INFORMÁTICA MÉDICA

TOMO II

COMPUTACIÓN

Capitulo 7. METODOLOGÍA DE LA INVESTIGACIÓN

La Ciencia es la progresiva aproximación del hombre al mundo real.

Max Planck

7.1 Introducción.

El estudio de la metodología de la investigación como el de cualquier otra ciencia exige en primer termino establecer el objeto de estudio de esta materia, el cual está vinculado con los métodos de investigación utilizados para la creación y desarrollo del conocimiento científico y las leyes generales a que están sometidos estos procesos en las diferentes ramas de la ciencia.

El camino del conocimiento como dijera Lenin, va de la contemplación viva (observación de los problemas que surgen y exigen solución) al pensamiento abstracto (creación de una teoría que explique el fenómeno o proceso en relación con el problema) y de este a la práctica (comprobar de hecho que la teoría explica el problema planteado).

Por ejemplo, si se quiere saber cuales son las causas de la hipertensión arterial, la respuesta hay que buscarla por medio de la aplicación estricta del método científico, que consiste como hoy día se conoce de una serie de pasos, de los cuales exponemos algunos de ellos, como son: establecer lo más preciso posible el problema a resolver y la(s) hipótesis a verificar, así como, los objetivos o tareas a llevar a cabo, y en relación con estos últimos aspectos, el tipo y diseño general de la investigación, que contempla elementos tales como, el universo de estudio, el diseño de la muestra, etc.

Estos y otros aspectos de la metodología de la investigación serán abordados a continuación.

7.2 La ciencia y el conocimiento científico. Revolución del conocimiento.

Toda operación cognoscitiva se dirige hacia un objeto con el que tiende a establecer una relación de la que surja una característica efectiva del objeto. Por tanto, las interpretaciones dadas en el curso de la historia de la filosofía se pueden considerar como interpretaciones de esta relación y como tal dirigirlas a dos alternativas fundamentales:

Tal relación es una identidad o semejanza, y la operación cognoscitiva es un procedimiento de identificación con el objeto o de su reproducción.

La relación cognoscitiva es una presentación del objeto, y la operación cognoscitiva un procedimiento de trascendencia.

Parte del conocimiento, es **ordinario**, es decir, no especializado, y otra parte es **científico**, o sea, se ha obtenido mediante el método de la ciencia y puede ser sometido nuevamente a prueba, enriquecerse y, llegado el caso superarse mediante el mismo método.

Por otra parte, la ciencia es un conocimiento de naturaleza especial, inventa y agrega conjeturas que van más allá del conocimiento común. Elabora sus propios cánones de validez y, en muchos casos, se encuentra muy lejos del conocimiento común; así, la sistematización coherente de enunciados fundados y contrastables se consigue mediante teorías, que son el núcleo de la ciencia y aspiran a ser racionales y objetivas. El sentido común no puede obtener más que una objetividad limitada

Estos planteamientos constituyen el centro de la información acerca del conocimiento y su formación y construcción epistemológica.

Tales conceptos e informaciones se encuentran en los libros de Metodología de la Investigación con enfoques más o menos amplios, por tal razón abordamos aquí, no sólo la formación del conocimiento, sino de la revolución de la información y del conocimiento. Según Luigi Valdés[1]:

...después de haber agotado el modelo de la Revolución Industrial la sociedad y las empresas buscaron una nueva fuente de generación de riqueza y la encontraron en el conocimiento...

Pero, ¿dónde se forma y organiza el conocimiento? Se parte de la relación sujeto – objeto; y ese sujeto que acumula, estructura y sistematiza la información, es el ser humano. Este conocimiento sistematizado en la información y en el propio personal portador constituye lo que hoy se llama el capital intelectual.

...el capital intelectual es todo el inventario de conocimientos generados por la empresa y expresados como tecnología; patentes, mejora de procesos, productos y servicios; la información: conocimiento de clientes, proveedores, competencia, entorno y oportunidades; y habilidades desarrolladas por el personal: solución de problemas en equipo, comunicación, manejo de conflictos, desarrollo de la inteligencia. Todo unido y orientado a crear valor agregado de forma continua para el cliente[2]

El capital intelectual son todos los bienes intangibles que se relacionan con el conocimiento, habilidades e información.

Dicho de esta manera, puede parecer un **sacrilegio** hablar de valor, consumo y clientes, en un sector humanitario como es el de la salud; sin embargo, una de las vías de perfeccionamiento de la atención, promoción, prevención y cuidado de la salud es la investigación científica, la producción de nuevos conocimientos; pero no se trata sólo de producirlos, sino de ponerlos en circulación, convertirlos en productos de valor agregado.

El énfasis del capital humano está justamente en el uso del conocimiento y la tecnología para sustituir cualquier factor de producción y perfeccionar integralmente la calidad de los servicios, cuando el objeto de perfeccionamiento sea ese.

¿En qué productos se expresa el conocimiento?: patentes, procesos, habilidades administrativas, tecnologías, información y experiencia. Todo este conocimiento es el capital intelectual, que es la expresión de la suma de lo que saben todos y que le confieren una ventaja competitiva en el mercado y en cualquier espacio de socialización. Su aprovechamiento dependerá, en gran medida, del paradigma económico en que se encuentre insertado: capitalista o socialista.

7.2.1 Información, inteligencia y conocimiento.

La estructuración del capital intelectual depende de estos tres elementos.

Los datos son simplemente un conjunto de números o de cifras, constituyen los bloques básicos que sustentan una economía (u organización y prestación) sustentada en la información y el conocimiento. Se presentan en tres formas: palabras, sonidos o imágenes, y su uso incluye las funciones de creación, manipulación, procesamiento, movimiento y/o almacenamiento.

Cuando los datos son ordenados se convierten en información básica.

Ellos son formas de expresar algo, mientras que la información es el arreglo de los datos en patrones que tiene una interpretación o significado. La información es el nivel inmediatamente superior a los datos.

---

[1] Valdés Luigi.Conocimiento es futuro. Hacia la sexta generación de los procesos de calidad..CONCAMIN. centro para la calidad total calidad..CONCAMIN. centro para la calidad total y la competitividad. 1996: 330y la competitividad. 1996: 330

[2] op. cit:331

La interpretación de la información se realiza gracias a la inteligencia de las personas y se convierte en conocimientos. Este conocimiento, es un acervo de información dinámica, se encuentra en el cerebro y se puede expresar como información. Él, además, es la aplicación y el uso productivo de la información.
La información es pasiva y estática, mientras que el conocimiento es activo y dinámico.
A medida que se modifica la información por medio de la inteligencia y conocimiento de la gente, se le está agregando valor.
Actualmente, en cualquier sistema empresarial o de servicios, uno de los elementos más importantes es la generación de valor agregado vía conocimiento para el usuario.
La generación de valor agregado por conocimiento son todas las ideas, sugerencias y cambios, propuestos por el personal de la empresa, orientados a mejorar los productos y los servicios que le ofrecen al cliente, y que, por consiguiente, aumentan su demanda y aceptación.
Los servicios de salud son provisores de servicios, y es importante cambiar la concepción de los especialistas en el uso de la información, desde la generación de conocimiento hasta la introducción de la innovación tecnológica en el perfeccionamiento del sistema.
Ocasionalmente, se desvaloriza la contribución del conjunto de sujetos que participan en los procesos. El capital intelectual no puede ser producido en masa, es singular y depende esencialmente de cuatro factores: Educación, experiencia, habilidades naturales y actitud.
La educación es la base donde descansa todo el capital intelectual. Abarca el desarrollo integral de todos los colaboradores y es el único medio que puede garantizar la generación de valor. Para avanzar en el proceso del conocimiento, la experiencia tiene que ser potenciada, se necesita entender la esencial del por qué de cada acontecimiento. Si una persona se queda con el nivel del cómo, cuándo y dónde, entonces, esa experiencia será difícilmente reproducible.
Para potenciar la experiencia es necesario estructurarla y sistematizarla como información, y, después, encontrar un medio para compartirla con toda la organización, así se puede multiplicar el conocimiento y aprender de la experiencia. Esto es poco frecuente, muy pocas veces la experiencia se estructura como información.
Valdés también señala que aunque no haya diferencia en la capacidad intelectual de la personas, sí puede haberla en ciertas habilidades naturales. Al crear un inventario de las habilidades naturales de las personas, de cada uno de los colaboradores, entonces, estas se pueden desarrollar y potenciar de manera particular a cada una, colocándola en el lugar pertinente.
El cuarto elemento es la actitud, que no es más que la diferencia entre tener que aprender y querer aprender.
El objetivo de este enfoque de la generación del conocimiento no está dirigido a cambiar la cultura del ser, sino a inducirle posibilidades.

7.2.2 Nuevas necesidades en el contexto de un mundo globalizado.

No basta hoy con tener un ejército de investigadores o especialistas para producir conocimientos, para asimilar tecnologías o perfeccionar procesos, desde la organización gerencial hasta la producción de equipos y medicamentos.
El marco más propicio para la generación de conocimientos, es el espacio académico, las universidades. Pero esa producción cognoscitiva no tendría validez, sino se crea una organización que permita su transmisión, su facilitación para la aplicación (sin desmedrar la importancia de los conocimientos básicos o puros). Hay que buscar la relación entre las universidades y los centros de producción y servicios, hospitales, farmacias y atención primaria de salud, lo que puede facilitar el paso natural de la producción del conocimiento a su

introducción en la práctica. Es necesario favorecer esa interfaces natural que caracteriza el sistema de salud cubano.

Es necesario, entonces, potenciar el capital intelectual a partir de una organización que contenga sistemas para promover la generación del conocimiento, estimular la experimentación y evaluar los resultados.

Hoy día, las instituciones de salud, en todos los niveles integran la formación de recursos humanos, la investigación, la innovación tecnológica como vía pera perfeccionar la organización y la atención a la salud.

Así, la estructura de las instituciones debe contribuir a potenciar la conducta de las personas, y promocionar el carácter creativo de cada sujeto en su puesto de trabajo.

La nueva cultura organizacional debe asumir la experimentación y la innovación como dos de sus principales valores. Permitir la crítica en el sentido de averiguar exactamente el por qué de las cosas y crear un genuino interés por la investigación.

Se debe fomentar humildad y curiosidad sobre el accionar de las instituciones. Hay que sensibilizar a los superiores, prepararlos, para oír propuestas. Debemos ser capaces de generar propuestas creativas, y el sustento de ellas es la generación de conocimientos.

Tenemos que valorar los resultados en función de los procesos, e indagar sobre cómo, por qué mejorar su estructuración, lo que generará resultados tangibles e intangibles más útiles en términos cognoscitivos y prácticos.

El conocimiento se mueve, de ahí que sea necesario aprovechar todas las fuentes de información, buscar su significado a partir de la integración sistémica de la misma, adicionarle valor y crear productos del conocimiento que consoliden las acciones y difundan las experiencias, desde donde se genera el conocimiento hasta su aplicación práctica, desde las esferas más teóricas de la ciencia hasta el terreno de la innovación tecnológica.

7.3. La investigación científica.

El proceso de investigación científica es la vía principal de obtención de nuevos conocimientos.

Se puede definir como un conjunto de acciones planificadas que se realizan con la finalidad de resolver, total o parcialmente, un problema científico determinado.

7.3.1 Clasificación.

Puede ser muy diversa acorde con el eje de clasificación que se asuma. A continuación expondremos dos ejes generales.

7.3.1.1 Según estado de conocimiento alrededor del problema.

Exploratoria. En este caso el conocimiento del problema es pobre y se necesita profundizar para delimitarlo correctamente. En este punto el investigador cuenta esencialmente con un área problema y el fin de este tipo de investigación lo constituye la delimitación de todas las partes que lo conforman.

Descriptiva. Se clasifica así, cuando se ha avanzado en el dominio del área problema y en la delimitación del problema mismo, pero el ámbito de conocimientos resulta limitado para buscar relaciones entre variables. La investigación descriptiva siempre se encuentra en la base de la explicativa.

Experimental. En este tipo el problema está bien identificado y definido, y permite avanzar en la búsqueda de relaciones causales. En este tipo de investigación resulta imprescindible la formulación de hipótesis que pretenden explicar las causas del problema o eventos que estén relacionados con las mismas.

Cuasi-experimental. Esta Clasificación está referida a aquellas investigaciones en que el investigador no puede manejar el factor de experimentación para evaluar el comportamiento de los resultados (experimentales), sino que tiene que organizar la observación de datos de manera que le permite verificar o

refutar las hipótesis (observacionales). Según Jiménez Paneque (1997) este tipo de investigación la ubica en los estudios explicativos, de tipo observacional.

7.3.1.2 Según el alcance de los resultados.

Fundamental. En este tipo de investigación el propósito está dirigido a la búsqueda de un nuevo conocimiento, pero no puede precisarse su relación con la práctica social.

Fundamental Orientada: Son aquellas investigaciones cuyo vínculo con la práctica social es indirecta y mediata. Sus resultados no tienen una aplicación inmediata, pero pueden conducir a resultados que sí la tienen.

Aplicada. El problema objeto de la investigación científica surge directamente de la práctica social, y genera resultados que son aplicables de manera inmediata.

De desarrollo. Es aquella investigación dirigida a completar, desarrollar y perfeccionar nuevos materiales, productos o procedimientos. Se incluyen aquí los estudios realizados para evaluar el resultado de la implantación de nuevos procedimientos o técnicas, como pudiera ser la evaluación de nuevos métodos de diagnóstico y tratamiento, evaluación de tecnologías sanitarias, etc.

7.3.2 Etapas de la investigación.

Las etapas de una investigación se refieren al conjunto de acciones que se deben realizar para que el desarrollo de la misma llegue a su destino final.

Estas son: Planificación, Organización, Ejecución, Evaluación, Redacción del informe final, Introducción en la práctica de los resultados, que pasamos a analizar a continuación.

7.3.2.1 Planificación.

Planificar es ordenar en secuencia lógica una serie de actividades que llevaran en conjunto, una vez que se hayan ejecutado todas, a la obtención de un objetivo o producto, en nuestro caso, los resultados de una investigación.

La planificación resulta la etapa más compleja de la organización de la investigación y termina con la elaboración del proyecto de investigación.

En este esquema organizativo deben quedar delimitados los aspectos siguientes:

El problema de investigación y los objetivos.

Las vías o métodos que se utilizarán para resolver dicho problema.

Los recursos necesarios y disponibles.

Los métodos y formas que se van a utilizar para recolectar la información.

La forma de procesamiento y análisis de la información.

En esta etapa se deben considerar todos los elementos necesarios que garanticen que el producto (o los productos) de la investigación, sea éste un nuevo conocimiento, un producto mejorado, una nueva tecnología o cualquier otro, se inserte en los procesos de información o en la práctica, de manera que el conocimiento no quede estático.

A la hora de planificar una investigación, además de los requerimientos propios, se requiere tener en cuenta tres niveles:

Nivel estratégico. El investigador estratégicamente debe conocer las políticas de desarrollo y los objetivos de la ciencia y la técnica en el país, los marcos y regulaciones de restricción y la posibilidad de asignación de recursos nacionales e internacionales. Tener en consideración estos elementos permite avalar, en el terreno cognoscitivo la posibilidad real de llevar a vías de hecho la investigación.

Por ejemplo, no basta con querer **conocer el impacto del tratamiento con antioxidantes en el paciente séptico**, sino que es necesario delimitar la relevancia y pertinencia del tema, acorde con las políticas de desarrollo de la ciencia y la técnica en el país. Si el tema no es de interés de los que soportan el financiamiento es poco probable que la investigación más importante, a juicio del o los investigadores, pueda llevarse a cabo.

Nivel táctico. Responde como complemento del nivel estratégico en la planificación y está relacionado con la programación que para la ciencia y la innovación tecnológica exista en el país y, en el caso particular que nos ocupa, en el sector salud.

En este nivel, es determinante tener en consideración el inventario de activos y recursos, dicho de otra manera, hay que conocer el potencial científico e informativo con que se cuenta para dar respuesta a las necesidades investigativas.

Es necesario prever las necesidades antes de comenzar la investigación; para saber con qué se cuenta, qué hace falta y planificar adecuadamente los recursos para los que se necesita financiamiento.

Nivel operacional. En este aspecto, se requiere delimitar adecuadamente el proceso de monitoreo y evaluación, del proyecto con el fin de garantizar su calidad y cumplimiento en el tiempo programado.

En resumen, al planificar se deben tener en cuenta las normativas establecidas por el sistema de ciencia e innovación tecnológica en cada país, las regulaciones políticas establecidas en términos de ciencia y tecnología, y las regulaciones internacionales que me permitirán en el momento oportuno buscar oportunidades para el financiamiento de un proyecto dado.

En nuestro país, actualmente, el Ministerio de Salud a delimitado los Programas Priorizados para el Sector, estos son:

Salud materna infantil.
Calidad de vida.
Enfermedades transmisibles.
Enfermedades crónicas no transmisible.
Accidentes.
Aptitudes físicas y mentales de los niños.
Investigación en sistemas y servicios de salud.
Medicamentos.
Medios diagnósticos.
Evaluación de tecnologías sanitarias.
Medicina natural, tradicional y termalismo.
Atención al adulto mayor.

Es conveniente que en la etapa de planificación se tenga en consideración las posibles fuentes de financiamiento, tanto de carácter nacional como internacional.

Las fundamentales fuentes de financiamiento nacionales son: el Ministerio de Ciencia Tecnología y Medio Ambiente (CITMA), el Ministerio de Salud Pública (MINSAP), el Gobierno (Consejo de Estado), otros ministerios, empresas y ramas de la economía nacional. Estas fuentes de financiamiento se expresan en convocatorias de financiamiento a programas, su correspondencia con la convocatoria como organismos es:

PROGRAMAS	ORGANISMOS
Nacionales	CITMA
Ramales	MINSAP

Territoriales	CITMA
Proyectos no asociados a programas	CITMA

7.3.2.2 Organización.

Es el proceso donde se produce el ajuste en la práctica de lo planificado y se toman las determinaciones finales. Esta etapa se corresponde con la recepción de recursos, entrenamiento de personal, estudio y caracterización del campo de estudio y de la preparación y validación de instrumentos.

7.3.2.3. Ejecución.

Esta etapa se corresponde con la puesta en marcha de la investigación, e incluye la recolección de la información, procesamiento y análisis.

7.3.2.4. Evaluación.

Al finalizar la investigación se relaciona lo real ejecutado contra lo planificado. Se evalúa el cumplimiento de los objetivos y de los resultados esperados respecto a aquellos que se han obtenido.

7.3.2.5. Redacción del informe final.

El informe final es el documento donde se recoge el resultado completo de la investigación en forma escrita.

7.3.2.6. Introducción de los resultados en la práctica.

Esta etapa se corresponde con la publicación o introducción en la práctica social del resultado obtenido. Debe haberse previsto en la etapa de planificación, la introducción del resultado, al igual que la determinación del usuario o **cliente** del producto de la investigación, que garantice la continuidad del proceso. En algunos casos, el resultado de una investigación y su introducción en la práctica social se realiza a partir de la elaboración de un proyecto de innovación tecnológica, al que nos referiremos en el tema correspondiente a tipos de proyectos.

7.4. El problema y los objetivos de investigación científica.

En todo proceso de investigación científica el elemento protagónico lo constituye la adecuada formulación del problema de investigación.

El problema de investigación científica puede entenderse como una discontinuidad o salto dentro del proceso del conocimiento, que surge debido a que el objeto de estudio, cuando es el caso queda comprendido dentro de aquella zona del saber en que se ignora la esencia del objeto, o esta es incompleta o lo que se conoce de ella es contradictorio.

Los pasos a seguir para delimitar el problema de investigación deben ser:

Identificación. Selección del objeto de estudio entre el conjunto de problemas (área problema) existentes en la realidad.

Delimitación. Identificación y señalamiento de todos los aspectos que forman parte **del problema seleccionado.**

Definición. Precisión dentro de todos los aspectos que involucran al problema, de aquellos que específicamente serán incluidos o analizados en la investigación, enmarcándolos en espacio y tiempo.

7.4.1. Características de la formulación del problema.

La formulación del problema se puede realizar en forma de pregunta o declaratoria. Siempre debe expresar con claridad la relación entre las variables a estudiar y la posibilidad de su validación empírica.

Debe cumplir los requisitos siguientes:

Ser objetiva y fundamentada, o sea, la formulación del problema debe estar inmersa en la construcción del marco conceptual de referencia que le da soporte.

Ser específica, es decir, en la medida en que sea más precisa, se visualizará con claridad qué se quiere conocer.

Ser contrastable empíricamente, el propio proceso de formulación del problema debe conducir a la valoración de la factibilidad de darle solución y a la utilidad o conveniencia de realizar el estudio.

Silva ha señalado que la estructura del pensamiento y del accionar de un investigador en la fase de formulación del problema discurre según los elementos siguientes:

1. **Expresar el problema nítidamente (mediante preguntas o hipótesis).**
2. **Fundamentar la necesidad de encararlo (beneficios esperados).**
3. **Exponer tanto el marco teórico en que se inscribe como los antecedentes en que reposa.**

Estos tres elementos resultan inseparables para una adecuada formulación del problema.

Silva refiere también las dificultades básicas que en ocasiones obstaculizan o impiden una formulación adecuada del problema de investigación, estos son:

No se informa detalladamente el origen del problema, ni se da un fundamento teórico, bibliográficamente respaldado.

Las preguntas no son explícitas, quedan subsumidas dentro de un planteamiento general usualmente borroso.

No deja bien delimitado el aporte real, la necesidad que cubre.

La formulación es ambigua.

El enunciado del problema incluye parte del método para resolverlo.

La concepción del problema carece de un enfoque crítico.

En resumen, considero que la formulación del problema constituye la base, la arrancada del proceso de investigación y resulta vital para garantizar su calidad, para conducirlo de manera coherente, organizada, y sistemática hasta el fin.

7.4.2. Objetivos de la investigación.

Estos se refieren a los aspectos que se desean estudiar, a los resultados intermedios y finales que se esperan obtener para dar respuesta al problema.

Si en el problema de investigación se define el ¿qué se quiere investigar? en los objetivos se define con precisión ¿a dónde se quiere llegar?.

Los fines de la formulación de objetivos están dirigidos a servir como una guía para el estudio, determinar sus límites y amplitud, orientar los resultados que se esperan alcanzar y visualizar las etapas del proceso que se pretende desarrollar.

Para lograr esto, en la formulación de los objetivos se deben considerar los aspectos siguientes:

La correspondencia con el problema de investigación

Enunciar el resultado unívoco, claro, preciso, factible y medible que se obtendrá una vez terminada la investigación

7.4.2.1 Clasificación.

Es frecuente en nuestro medio que los objetivos se dividan en generales y específicos, pero tal división no constituye un esquema inflexible; y se realizará cuando corresponda a la investigación que nos proponemos realizar.

Los **objetivos generales** nacen directamente del problema y constituye los propósitos de mayor alcance en el estudio.

Los **objetivos específicos** sintetizan la forma en que se alcanzaran los objetivos generales. Poseen un nivel de precisión mayor y se dice que constituyen **guías para la acción** porque a partir de la definición clara y precisa de los resultados que se pretenden alcanzar permiten delinear los métodos que se emplearán para conseguirlos.

Veamos, por ejemplo, cómo en una investigación de salud reproductiva, referida al estudio de la influencia de las relaciones de subordinación de género en la expresión del climaterio, el problema de investigación se estructura alrededor de las preguntas siguientes:

¿Las relaciones de género condicionan variaciones en el proceso salud-enfermedad de la mujeres climatéricas?

¿Cuáles son las características del síndrome climatérico que prevalecen en estas mujeres?

¿Las relaciones de género expresadas en el microsistema de la familia, la pareja y el trabajo doméstico e insertado, que rodean a la mujer, incrementan la aparición de síntomas climatéricos?

Los objetivos deben corresponderse con la pregunta o preguntas formuladas, en el ejemplo serían:

Comprobar que las relaciones de género modifican la expresión del síndrome climatérico.

Describir las características reproductivas y climatéricas de la población objeto de estudio.

Caracterizar las relaciones de género que prevalecen en el microsistema de la familia, la pareja y el trabajo insertado y doméstico, y evaluar su probable contribución en el incremento de la sintomatología climatérica.

Y la **hipótesis** sería: en el medio social y familiar en que vive la mujer de edad mediana prevalecen relaciones de género que favorecen el incremento de la sintomatología climatérica.

En resumen, los objetivos derivan del planteamiento del problema y se formulan sobre la base de las principales interrogantes que se desean contestar por medio del estudio. Los objetivos están enfocado hacia los resultados, a lo que se desea conocer, explorar, determinar y demostrar; o sea, son los objetivos los que orientan la formulación, las hipótesis, la definición de variables e indicadores del estudio y el plan de análisis de los datos.

7.4.3. El marco conceptual.

Está constituido por el cuerpo de teorías, conceptos, referentes y supuestos donde se inscribe el estudio que se pretende abordar. En este acápite se trata de mostrar la existencia de una construcción teórico conceptual donde el problema detectado puede inscribirse, y en este sentido será el instrumento de mayor utilidad para poder establecer las categorías de análisis y su relación, así como la definición de las preguntas claves, las hipótesis y las variables de estudio.

En el marco conceptual debe quedar claramente delimitada a inconsistencia en el proceso del conocimiento del que deriva el problema de investigación enunciado. El investigador en su fundamentación debe aportar ideas y evaluar el impacto a obtener por el conocimiento alcanzado, si la investigación llega a cumplir sus objetivos.

El desarrollo del marco conceptual exige una exhaustiva revisión bibliográfica a través de la cual se muestre el dominio de la literatura que versa sobre el tema, y se crean o construyen propuestas teóricas alternativas para explicar el fenómeno que se pretende estudiar.

7.4.4. Formulación de hipótesis.

Del marco conceptual de referencia, y teniendo como soporte las preguntas que definen el problema de investigación, se formulan las hipótesis o posibles respuestas a esas interrogantes, las que expresan las relaciones causales que se pretenden encontrar, verificar o rechazar.

Las investigaciones exploratorias y descriptivas, si bien no tienen un cuerpo de hipótesis estructuradas como relaciones causales de asociación, podrían tener hipótesis formuladas como supuestos o resultados que se esperan encontrar. En otros casos, tan solo se plantearía la necesidad de conocer el comportamiento del fenómeno, frente a la ausencia de conocimientos previos.

Necesariamente, la construcción de las hipótesis exigen al investigador identificar las variables del estudio y determinar sus niveles de independencia, dependencia o condicionalidad, según el modelo o marco teórico que sustenta el estudio.

El proceso de investigación constituye la interacción de un conjunto de elementos que permite, al final, contribuir al enriquecimiento de los cuerpos teóricos que nutren el conocimiento científico. La relación entre el qué y el a dónde resulta determinante para el inicio del proceso investigativo, en el que, como en una carrera de velocidad, lo determinante es la calidad de la arrancada, esta es la llave del éxito.

7.5. Métodos de la investigación.

Una vez definidos claramente el problema y los objetivos es necesario seleccionar los métodos, técnicas y procedimientos que se utilizarán para darle respuesta.

Una forma ordenada de definir los métodos será:

Tipo y diseño general del estudio.
Definición de universo de estudio y muestra.
Definiciones operacionales.
Procedimientos y técnicas para la recolección de información.
Métodos para el control de la calidad de los datos.
Procedimientos para garantizar los aspectos éticos.

7.5.1. Tipo y diseño general del estudio.

El diseño de la investigación se corresponde con la determinación de estrategias y procedimientos que han de seguirse para dar respuesta al problema, darle salida a los objetivos y comprobar las hipótesis; por tal razón, es necesario precisar con claridad el estado de conocimiento sobre el problema de investigación.

El estudio se puede clasificar en cuatro tipos, según se considere el:

1. **Tiempo de ocurrencia de los hechos y registro de la información**, en:

Retrospectivo: se indaga sobre hechos que ya han ocurrido

Prospectivo: se registra la información en la medida que van ocurriendo los hechos.

2. **Período y secuencia del estudio**¸ en:

Transversal: se hace un corte en el tiempo y se estudian las variables simultáneamente. El tiempo no es importante a como se dan los hechos.

Longitudinal: estudia una o más variables a lo largo de un período que varia según el problema de investigación y las características de las variables en estudio.

3. **Control que tiene el investigador sobre las variables**, en:

Caso control: se aplica en los estudios donde se desea conocer qué parte de la población que presenta determinado atributo o carácter estuvo expuesta a la causa o factor supuestamente asociado. Se parte del efecto (E) a la causa (C).

Estudio de cohorte: aplicable cuando interesa conocer qué parte de la población expuesta a la causa (C) presenta determinado resultado o efecto (E). En este tipo de estudio se excluye la población expuesta que presenta el efecto (E), y el grupo control lo constituyen los que no están expuestos a la variable condicionante, con el fin de comparar la expresión del efecto (E) en uno y otro grupo.

4. Análisis y alcance de los resultados, en:

Exploratorio: son estudios cuyo objeto fundamental es familiarizar al investigador con el problema a investigar.

Descriptivo: son aquellos estudios dirigidos a profundizar en el conocimiento del problema que se analiza; son utilizados con frecuencia para caracterizar un hecho o conjunto de hechos que caracterizan una población.

Analítico o explicativo: son aquellos dirigidos a responder por qué se produce determinado fenómeno y cuál es la causa o factor asociado al mismo. En este tipo de estudio se analizan relaciones causa-efecto.

Experimentales: son estudios que se caracterizan por la introducción y manipulación del factor causal para la determinación del efecto. Este tipo es muy utilizado en la clínica y en investigaciones biomédicas.

Cuasi experimental: se utiliza cuando el grupo control no se puede dejar sin la intervención. En tal caso se asimila un modelo que permite hacer una analogía con el tipo de estudio experimental.

De evaluación: están dirigidos a evaluar la eficiencia, eficacia y efectividad de algo, por ejemplo, acciones de salud, tecnologías, medicamentos, programas, etc.

7.5.1.1. Consideraciones generales para la selección del tipo de estudio.

El investigador debe seguir ciertos criterios que le permitan una selección precisa del objeto de estudio, entre ellas es recomendable tomar en consideración:

Las variables y su medición
El riesgo que implica para los sujetos en estudio
El tipo de relación que se busca entre las variables
El tiempo necesario para la observación del fenómeno
Los recursos disponibles para el estudio.

7.5.2. Definición de universo de estudio, muestra, unidades de análisis y de observación.

La definición inequívoca del universo en estudio, de las unidades muestrales, así como del alcance de la inferencia a ser realizada, son requisitos a tener en cuenta en toda investigación científica.

El investigador para conocer que unidades de la población va a usar como muestra deberá asesorarse con un especialista en esta rama del saber, conceptos tales como tamaño de muestra, representatividad de la muestra, tipo de muestreo empleado para la selección de la muestra, son básicos y sobre los cuales en los capítulos 8, 10 y 11 se puede profundizar en estos y otros aspectos de interés.

7.5.3. Elementos que se deben considerar en la definición de la muestra.

El investigador definirá la muestra de la población a estudiar acorde con los objetivos definidos para el estudio. Para seguir una secuencia lógica se procederá de la manera siguiente:

Definir la población, tamaño y elementos que la componen
Determinar la unidad muestral, la unidad de observación y sus características
Definir el tamaño de la muestra
Definir los procedimientos que deben seguirse en el proceso de selección de la muestra.

7.5.4. Definiciones operacionales.

La operacionalización de las variables es el proceso mediante el cual el investigador define las categorías y/o variables del estudio, tipos de valores (cuantitativos o cualitativos) que podrían asumir las mismas y los cálculos que se tendrían que realizar para obtener los valores de los indicadores, concebidos.

Todas las variables deben estar claramente definidas y operacionalizadas.

Para operacionalizar una variable, es necesario partir del cuerpo teórico que define el concepto de esta variable. Por ejemplo la variable nivel de escolaridad se conceptualiza como el último nivel de enseñanza aprobado por la persona encuestada; se especifica la forma de registro, para el caso que nos ocupa se registra como variable cualitativa ordinal con categoría: primaria incompleta, primaria completa, nivel medio (incluye Secundaria Básica, Pre universitario y técnico medio), nivel universitario.

Como se observa, este tipo de descripción evita posibles confusiones a la hora de definir e interpretar las variables. Este procedimiento favorece la formulación de los instrumentos para la recogida de la información.

Para facilitar el proceso de operacionalización de variables debe usted proceder de la manera siguiente, identifique y conceptualice las variables fundamentales contenidas en sus objetivos e hipótesis, después, evalúe si dentro de la variable principal no se contienen otros posibles aspectos contradictorios. Finalmente defina la forma de registro y cuando sea el caso, la codificación de los valores de las variables en los instrumentos de recolección de la información y especifique la escala de medición de cada una, y los indicadores a calcular (porcentaje, tasas, medio, desvío estándar u otros.

Observe el ejemplo siguiente:

Variable general:

Tener acceso a los servicios estomatológicos de una institución de salud determinada dentro de una región geográfica dada.

Variables asociadas (a la variable general):

1. **Accesibilidad geográfica**:

   (Tiempo en hrs./min. que tarda en llegar un sujeto desde un domicilio hasta el servicio.)

   (Tiempo en hrs./min. que demora en ser atendido por el servicio mencionado).

   Cuantos sujetos de una subregión prefijada al servicio de salud mencionado por unidad de tiempo (semana, mes, trimestre, semestre, año).

Accesibilidad económica:

   Cuanto gasta un sujeto en trasladarse desde su domicilio hasta llegar al servicio de salud mencionado.

   Cuanto invierte un sujeto al ser atendido por el servicio de salud de acuerdo con el problema de salud bucal que tenga.

   Cuanto gasta un sujeto de una subregión prefijada en acceder al servicio de salud mencionado (llegar hasta el lugar).

Accesibilidad cultural:

   Tienen conocimiento los sujetos del área sobre la atención en el centro de referencia.

7.5.5 Plan de análisis.

La decisión de los datos que se deben recolectar depende de los objetivos de la investigación, del material estudiado y del contexto en que se realizará dicha recopilación. El investigador debe limitarse a recoger la información que va a ser estudiada.Un elemento fundamental en el diseño de una investigación es la descripción del plan de análisis de los datos y la justificación del por qué se selecciona para tal fin. El análisis que se proponga debe ser coherente con los objetivos y las hipótesis del estudio.

Si se emplean técnicas estadísticas, se debe justificar convenientemente su uso. No es suficiente hacer mención de paquetes estadísticos (software), sino que es preciso dejar claro en qué se piensan emplear y qué resultados se espera obtener mediante su aplicación.

Para facilitar la planificación del plan de análisis de la información resulta recomendable hacerlo por objetivos, explicar con claridad cuál y por qué se selecciona el método escogido para el tratamiento de la información.

En resumen, la descripción de los métodos que se van a utilizar en el proceso de la investigación debe quedar expresada con claridad y precisión. Debe redactarse de tal manera, que un profesional con similar nivel del competencia al investigador que elabora el proyecto pueda llevar a cabo la investigación a partir de esta referencia.

Finalmente, para dar respuesta a los objetivos de la investigación se pueden utilizar diferentes métodos, cuantitativos y cualitativos, la selección dependerá del carácter, profundidad y delimitación del a dónde se quiere llegar con el conocimiento a obtener.

7.6 Proyecto de investigación.

Es el documento que contiene la exposición razonada de lo que se quiere estudiar o resolver, fundamenta la necesidad de su ejecución y expone como se realizará el proceso. El proyecto es un producto informativo de valor agregado.

Se pueden clasificar globalmente en:

1. Proyectos de investigación científica.

Aquellos cuyos objetivos están dirigidos a obtener nuevos conocimientos:

2. Proyectos de desarrollo y de innovación tecnológica.

Estos son los que sus objetivos están dirigidos a dar soluciones a problemas:

El mundo actual se desenvuelve entre lo que se ha dado en denominar una cultura de proyectos, entendida como el conjunto de conocimientos creados, aprendidos y transmitidos en relación con la gerencia, planificación, diseño y negociación de proyectos en la comunidad científica y tecnológica; así también se habla de una tecnología de proyectos en función del ordenamiento sistemático de conocimientos que es necesario realizar, referido al conjunto de métodos, ***know-how*** e instrumentos, así como a principios de gestión y organización diseñados para su empleo en la formulación y gerencia de proyectos

7.6.1 Funciones.

Entre las funciones fundamentales del proyecto de investigación están, la de **planificación**, la **administrativa** así como la **cognoscitiva**. La de planificación permite la evaluación y monitoreo de la investigación. Mientras que la administrativa permite la evaluación para aprobación institucional y financiera. Y por ultimo la cognoscitiva permite el desarrollo integral a partir del ejercicio intelectual que desarrolla el investigador para formular y abordar la investigación.

7.6.2 Partes integrantes del proyecto de investigación.

Las partes básicas componentes del proyecto de investigación se estructuran sobre una base lógica que permite su interrelación. Estas son:

- El **Título, Resumen y Datos de identificación**
- La **Introducción** que comprende: **Planteamiento del problema** (justificación científica), **Fundamento teórico** (antecedentes y marco conceptual), **Referencias bibliográficas** y **Objetivos de la investigación.**
- Los **Métodos** que abarca lo concerniente a: **Tipo y diseño general del estudio**, las **Definiciones operacionales**, la **Muestra, Unidad de análisis,** la **observación y los Criterios de inclusión y de exclusión**, los **Procedimientos y técnicas para la recolección de información**, los **Métodos para el control de la calidad de los datos**, y por último, los **Procedimientos para garantizar los aspectos éticos**.

- El **Plan de análisis** de los resultados que incluye los aspectos siguientes: Los **Métodos** y **Modelos de análisis de los datos**, los **Paquetes de análisis estadísticos**, el **Presupuesto**, y el **Cronograma,** y como parte final los **Anexos.**

A continuación describiremos algunos de los aspectos anteriormente expresados.

Título del proyecto. Debe ser claro y preciso, se debe corresponder con el problema científico y con el objetivo principal a estudiar.

Resumen. En él deben recogerse los elementos fundamentales que caracterizan el proyecto, el qué se quiere investigar, a dónde se quiere llegar, los métodos, los beneficios sociales y económicos del proyecto. Por lo general son cortos, se escribe en tiempo futuro, y deben contener entre 250 a 300 palabras como máximo.

Datos de identificación. En este acápite se especifica, el nombre del jefe del proyecto y de los otros investigadores, la experiencia profesional y los grados científicos y docente que ostenta. En algunos proyectos se solicita un pequeño ***currículo*** de los investigadores. Estas especificaciones permiten mostrar la competitividad curricular de los que realizarán la investigación.

De igual manera se debe identificar la o las instituciones que soportan la investigación. Así como los organismos o instituciones financiadoras.

Introducción. En ella se deja clara la identificación del problema, la justificación científica de su análisis y el marco conceptual, tal como se explicó en el epígrafe 7.3.3. Al final se debe reseñar todas las referencias bibliográficas.

Objetivos de la investigación. Deben quedar claramente expresados, ser específicos, corresponder a las salidas del producto de la investigación que se pretenden alcanzar, desde un nuevo conocimiento hasta un proceso de innovación tecnológica.

Métodos. Se explican detalladamente todos los métodos y técnicas a utilizar en el proceso de la investigación, cuidando la lógica de la exposición

Plan de análisis de los resultados. Este aspecto es fundamental en la preparación del documento del proyecto de investigación, se corresponde con el plan de análisis de los datos. Su planteamiento tiene que coincidir con el de los objetivos y de las hipótesis. Debe incluir, los métodos y modelos de análisis de los datos, y los paquetes de análisis estadísticos.

Presupuesto. La investigación es una inversión económica, por lo que exige aseguramientos y recursos que se dedicarán en la medida que se requieran para alcanzar los objetivos plasmados en el proyecto. Esto se materializa a través de acciones basadas en un plan lógico, el cual se debe corresponder con los costos estimados del presupuesto.

El total del presupuesto solicitado lo constituye la suma del importe de los gastos directos e indirectos.

DIRECTOS	INDIRECTOS
Gastos de personal	Consumo de agua
Materiales gastables	Mantenimiento
Equipamiento	Electricidad
Viajes	Teléfono
Otros gastos	(20 – 60% de los costos directos)

A continuación se detallan los gastos

a) Gastos de personal. Contempla categorías ocupacionales, salarios básicos, devengados y totales, así como el porcentaje de tiempo que cada investigador dedicará mensualmente al proyecto. El formato puede variar acorde a la estructura de presentación de los documento del proyecto.

NOMBRE	CAT.	SB	%V	SD	% SS	TOTAL $ MN	%TOTAL	TOTAL $ MN
	TITULO	610.00	55.45	665.45	79.85	745.29	10	810.81

SB salario básico
%V %
vacaciones

9.09% acumulado por vacaciones
SD salario devengado
%SS seguridad social – 12% del SD
% total% de tiempo a dedicar al proyecto cada mes
total cálculo por 11 meses de proyecto

b) **Material gastable**. Incluye todos los productos requeridos para llevar a cabo la investigación, se describen las unidades y formas que se presentan en el mercado, los precios por unidad o por forma de presentación, la cantidad solicitada y el importe total del producto.

PRODUCTO	UNIDAD	PRESENTACION	PRECIO $ MLC	CANTIDAD	TOTAL $ MLC
PAPEL	UNO	PAQUETE 500 HOJAS	$ 5.OO	10	$ 50.00

c) **Equipamiento**. Se enumeran los equipos necesarios para la ejecución de la investigación. El porcentaje de depreciación puede variar de un formato de proyecto a otro. El Ministerio de Ciencia Tecnología y Medio Ambiente de Cuba (CITMA) admite un indicador de 0,66 % por mes.

EQUIPO	CANTIDAD	VALOR INICIAL	DEPRECIACIÓN /MES	TOTAL $ MLC 11 MESES
PC	1	$ 700.00	$4.62	$ 50.82

Depreciación 0,66% por mes

d) Viajes: incluye los costos relacionados con la planificación de viajes a otras unidades, se incluirán los gatos de pasajes, alimentación y hospedaje, considerando el importe por día, número de personas y días.

e) Gastos indirectos: Son aquellos que se requieren para procesos generales de administración, construcción, mantenimiento, suministro de electricidad, agua, gas, etc. Habitualmente resulta complicado y tedioso el calcularlo con exactitud, entonces se estima entre un 25 y un 65% del monto solicitado para los gastos directos. La decisión para escoger el valor dentro de este rango depende del coeficiente aprobado por el centro responsable de la investigación, cuando es financiamiento interno o lo que determine el financista. En

etapas iniciales se recomienda un valor de 30% para subsidios de agencias gubernamentales y tal vez menos para las de origen privado.

f) **Otros gastos**. En este acápite se consignarán aquellos materiales no relacionados directamente con la ejecución de la investigación, pero sí necesarios para actividades colaterales como serán: reproducciones de materiales (fotocopias), combustible para viajes locales, etc. Forma parte de este punto la depreciación de equipos por mes, considerando el valor inicial.

Cronograma. Es el esquema viable y coherente del desarrollo en función del tiempo, de la movilidad de todos los requisitos del proyecto: físicos, materiales, humanos y de cualquier otro tipo en la medida en que sean necesario. Debe comprender los aspectos siguientes:

a) Forma detallada y cronológica de expresar las secuencias de actividades que corresponden a la fase de ejecución del proyecto.

b) Precisar en detalle las previsiones de la cronología estimada, a fin de coordinar mejor la adquisición de materiales y equipos, la prestación de servicios por terceros, y la realización directa de tareas que permiten poner en marcha el proyecto.

c) Tener en cuenta la secuencia obligada de las tareas a ejecutar y de los responsables de su ejecución.

Anexos: Incluye todos los materiales que puedan servir para aclarar el contenido del proyecto, por ejemplo, los cuestionarios o las guías de entrevistas a utilizar, mapas epidemiológicos, etc.

7.6.3 Tipos de proyectos.

Según el propósito a que se dirigen y la lógica de presentación, los proyectos pueden ser de diferentes tipos. A continuación le presentamos una tabla resumen de los mismos, al final del capítulo le anexamos la lógica de presentación de los diferentes proyectos.

TIPOS DE PROYECTOS		
TIPO	ORIENTACION	INTENCION
De creación científica	Producción de nuevos conocimientos	Cognoscitiva
De innovación tecnológica.	Obtención de productos tecnológicos.	Obtención de productos nuevos o mejorados. Introducir servicios nuevos o mejorados. Implantar procesos productivos o procederes médicos nuevos o mejorados Introducir y validar nuevas o mejoradas técnicas de gerencia y sistemas organizativos.
De evaluación	A un “saber” relacionado con atributos de eficacia, calidad, eficiencia o impacto	Evaluar sistemas, procesos, fármaco, e intervenciones.
De intervención	Ejecución de acción (es) sobre un objeto dado.	Cambio o transformación de algo.

Finalmente, es importante una cuidadosa elaboración del proyecto de investigación. El formato definitivo se ajustara según resulte pertinente. La organización lógica de presentación del proyecto, que abarca desde, la formulación del problema, los objetivos, la(s) hipótesis y los métodos, se pueden variar en la forma que se solicite o se necesario, pero de un modo u otro estos aspectos deberán estar contenidos en cualquier formato de proyecto de investigación.

7.7 Informe final.

El informe final de la investigación debe dejar claro el aporte científico y social, y fundamentar si los resultados obtenidos representan un avance científico cualitativo en el área de estudio, ya sea en un nuevo campo o en uno ya conocido. Este aporte puede ser a nivel teórico o experimental. Para fundamentarlo se debe definir con claridad, en el marco conceptual elaborado durante la etapa de planificación, el estado del arte en el que se encuentra el área específica de la investigación realizada, en tanto la intencionalidad debe velar por que no se repitan estudios ya realizados, dentro o fuera del país.

El aporte social, se demuestra, al justificar la correspondencia y pertinencia de los resultados obtenidos con las fundamentales líneas de dirección y estrategias de la política científica nacional y de la rama sectorial en que se inserte, en nuestro caso la rama de la salud.

En la exposición del informe de los resultados se debe mantener una coherencia global, de tal manera que unos planteamientos se deriven de otros en una cadena bien estructurada del discurso; con una estructuración lógica, y con el mantenimiento de una atención regular a las reglas gramaticales.

Las aseveraciones que deriven de los resultados obtenidos deben tener un fundamento real, ser expuestas de manera específica, concreta, objetivas; no se deben magnificar.

En los casos, en que como producto de la investigación derive una teoría científica, es necesario que quede fundamentada la discusión crítica con las fuentes teóricas; especialmente cuando se trata de la existencia de teorías, paradigmas, enfoques, corrientes y escuelas que sostengan puntos de vista divergentes. No es suficiente que se haya establecido una relación con una teoría determinada; toda investigación científica tiene que significar un paso adelante no sólo en la captación de datos, sino en la propia teoría. A partir de la confrontación entre teoría y datos, la primera saldrá enriquecida, ya que habrá tenido que adecuarse para poder explicar nuevos datos, a los que antes no se refería.

Debe quedar referido si la teoría utilizada explica los datos de la investigación. La investigación arroja finalmente, como resultado, determinada teoría explicativa. Más que evaluar los aspectos estrictamente formales, se trata de analizar en qué medida la teoría utilizada da cuenta de la nueva realidad, articulándola con otras realidades y con conocimientos previos acerca de éstas.

La investigación se torna válida cuando a través de ésta, la teoría adquiera mayor potencia heurística, esto es, tiene una mayor capacidad de explicación respecto de su estado previo, anterior a la investigación en cuestión.

Es necesario verificar si las hipótesis han sido comprobadas o refutadas.

En lo referente a los métodos utilizados, debe quedar claro si los procedimientos empleados se corresponden con los resultados que se han obtenido. Dichos procedimientos deben haber sido suficientemente estandarizados y establecidos con rigurosidad científica, y las técnicas utilizadas, sustentar la validez de los datos obtenidos.

La bibliografía utilizada debe ser actual y pertinente con el objeto de estudio, para la clínica un promedio de 5 años puede ser aceptable.

7.7.1 Partes del informe final de investigación.

La redacción de un informe final depende mucho de la calidad con que se haya elaborado el proyecto de investigación, tal es así que la adecuada redacción del mismo soporta alrededor de un 50% la elaboración de un buen informe final.

Por lo general, el informe de investigación se divide en subsecciones o acápites denominadas: **Título**, **Datos de identificación**, **Resumen**, **Introducción**, **Material y métodos**, **Resultados**, **Discusión**, **Conclusiones**, **Referencias bibliográficas** y **Anexos**.1.

Título: Debe ser corto, preciso y específico y corresponderse totalmente con el problema de investigación y el objetivo general del estudio. Debe permitir al lector identificar el tema fácilmente y a los especialistas en información viabilizar su catalogación.

Datos de identificación: Incluye los investigadores que han participado en la obtención de los resultados de la investigación, las instituciones científicas o académicas que permitieron su ejecución y los organismos, agencias o instituciones financiadoras del proyecto que sustenta el informe final que se redacta.

Resumen: Se exponen en una extensión no mayor a 250 palabras los objetivos y alcance del estudio, los procedimientos básicos, los métodos analíticos y observacionales, los principales hallazgos y las conclusiones.

Introducción: En el acápite debe quedar claramente identificado el problema de investigación, la justificación de por qué se realiza, y el estado de la teoría en que se inserta el tema.

Material y métodos: Es muy importante, que aquí queden expresados todos los procederes utilizados para dar respuesta al problema de investigación. Debe exponerse con claridad la validez de la muestra, por lo que se será explícito en cómo se produjo la selección, no basta con referir que fue un muestreo simple aleatorio, sino que es necesario dejar expresado con claridad cómo se escogió la misma.

Este acápite se debe exponer en una secuencia lógica, es decir, definir la población y grupo de estudio, el diseño seleccionado, la selección y asignación de sujetos a grupos de estudio, la intervención o tratamiento, las técnicas utilizadas para recolectar la información, los métodos de análisis y los de tratamiento de la información (análisis estadístico).

De los métodos ya establecidos se citará la referencia bibliográfica, se describirán brevemente aquellos que no son bien conocidos a los que se le realiza alguna innovación, y se explicarán con todo detalle los métodos nuevos o que estén sustancialmente modificados. En todos los casos se informará por qué se usan y se declararan las limitaciones si las tuviesen.

En el caso que se requiera del uso de drogas o fármacos, todas se identificarán con precisión, se dará el nombre genérico, las dosis y vías de administración.

Se especificarán los fundamentos éticos del estudio, tanto para investigaciones en humanos como con animales de laboratorio, en ambos casos se tendrán en cuenta las normativas de la Declaración de Helsinki de 1975 revisada en 1983 y la Guía del National Research Council.

Se describirán las pruebas estadísticas en detalle, de manera tal que un investigador de igual nivel de competencia pueda verificar los resultados notificados. Se debe evitar el uso exclusivo de pruebas de significación (valores de p); es recomendable utilizar intervalos de confianza. Si se utilizan métodos de asignación al azar, hay que explicar con nitidez la forma en que se realizó, de igual manera se detallará cuando se hayan empleado métodos de enmascaramiento. Es recomendable utilizar, como referencia de los métodos aplicados, libros de texto conocidos y evitar la cita de artículos, cuando esto sea posible.

Resultados: En este punto debe presentarse sólo la información pertinente a los objetivos del estudio, los hallazgos han de seguir una secuencia lógica y se mencionarán los relevantes, incluso aquellos contrarios a las hipótesis. Este informe será lo suficientemente detallado, de modo que permita justificar las conclusiones.
Se deben cuantificar los resultados obtenidos con medidas adecuadas de error o incertidumbre, notificar las reacciones al tratamiento si las hubiere, indicar el número de observaciones y el recorrido de los datos observados, así como la pérdida de participantes en el estudio y especificar las pruebas aplicadas para analizar los resultados.
Tenga en cuenta que el texto es la principal y la más eficiente forma de presentar los resultados; los cuadros (tablas) y los gráficos (ilustraciones) se utilizarán sólo cuando aporten claridad a la exposición de los resultados; para los mismos datos no utilice explicaciones en el texto, tablas y gráficos, sino que debe seleccionar el que sea más ilustrativo para el lector.
Debe tener cuidado en no repetir elementos expuestos anteriormente.
Discusión: Es la parte del informe donde el investigador aporta el nuevo conocimiento obtenido. En este momento se examinan e interpretan los resultados de la investigación y se insertan en el marco conceptual de referencia previamente construido, se discuten las coherencias y contradicciones y se evalúan y califican las implicaciones de los resultados con respecto a las hipótesis originales. Es el espacio en que se produce el **vuelo teórico** del investigador, de donde emergen los nuevos conocimientos y las hipótesis que se deberán verificar en nuevos estudios.
Conclusiones: Esta zona del informe final debe dejar explícita las respuestas a las preguntas de la investigación, planteadas en la introducción y que condujeron al diseño y realización de la misma.
No debe ser una exposición de resultados, por ejemplo, el 85% de las mujeres que tuvieron nacimientos con bajo peso fumaban, sino una generalización que pudiera ser, **hubo una alta correspondencia entre el habito de fumar en mujeres embarazadas y el bajo peso al nacer**.
Se debe evitar hacer conclusiones sin apoyo en los datos obtenidos y las discusiones superficiales, que en lugar de contribuir a enriquecer el estudio lo oscurecen y limitan.
Referencias bibliográficas: Las referencias permiten identificar las fuentes originales de ideas, conceptos, métodos, técnicas y resultados provenientes de estudios publicados anteriormente.
Las referencias pueden enumerarse de forma consecutiva, con números arábigos situados entre paréntesis, o como superíndices, en el orden que aparecen por primera vez en el texto.
También puede utilizarse el citar el autor por el primer apellido y colocar entre paréntesis el año de la publicación. En este caso las referencias se organizan por orden alfabético.
No deben utilizarse resúmenes en calidad de referencias, así como las observaciones no publicadas, la cita de citas. Las comunicaciones personales, aunque las escritas, nunca las orales, pueden ser insertadas entre paréntesis en el texto.
Se pueden incluir los trabajos aceptados pero que aún no han sido publicados y se añade la denominación “en prensa” entre paréntesis. Las referencias deben ser revisadas por los autores consultando los documentos originales.
Anexos. Se incluirá toda la documentación que complete la información obtenida en la investigación, y que por su carácter o configuración no encuadren apropiadamente dentro del cuerpo del artículo.

Capítulo 8. Estadística Descriptiva.

8.1 INTRODUCCIÓN.

El campo biomédico, al igual que otros campos del saber, aporta a la Estadística un sin número de valores diversos y cambiantes los que si no son debidamente recopilados, resumidos y presentados, no será posible un análisis e interpretación positiva de los mismos ni se podrán obtener conclusiones concretas de ellas.

La Estadística Descriptiva a través del método estadístico provee al investigador de técnicas y procedimientos que coadyuven a realizar esta tarea, tras la cual, mediante simples pasos y pocos resultados, será capaz de tener una valoración bastante cercana a la realidad de cómo es la información con que se cuenta y, si se ha preparado el camino, servirá de base para extrapolar estas conclusiones hacia un conjunto de datos más amplio.

Entre los objetivos de este capítulo se encuentran:

Explicar las etapas del método estadístico.

Explicar los conceptos fundamentales como son los de Estadística, Estadística Descriptiva, Estadística Inferencial, Población y Muestra.

Identificar las fuentes de información y los procedimientos de obtención de esta.

Diseñar e interpretar la tabla correspondiente a una distribución de frecuencias.

Calcular e interpretar las medidas de tendencia central, dispersión y posición relativa.

Calcular e interpretar las medidas para el análisis de frecuencias relativas en el campo de la salud.

Presentar mediante forma tabular y gráfica la información estadística.

8.2 ESTADÍSTICA. CLASIFICACIÓN. ESTADÍSTICA DESCRIPTIVA E INFERENCIAL.

La Estadística es una ciencia de amplia aplicación en todos los campos del saber humano cuyo nombre se deriva de su relación con la recolección de datos útiles para la administración de los estados.

El término estadísticas (en minúscula y plural) se refiere al conjunto de datos, mientras que el de Estadística (en mayúscula y singular) a la ciencia de la experimentación, encargada de las técnicas y procedimientos adecuados para la recolección, elaboración, análisis e interpretación de la información en estudio.

Una forma de clasificar la Estadística es en: Estadística Descriptiva y Estadística Inferencial.

- La Descriptiva. Está constituida por los métodos estadísticos destinados a la elaboración primaria de datos, o sea, que permiten la consolidación o resumen de la información y su posterior presentación.
- La Inferencial. Es la constituida por los métodos para el análisis y elaboración de los datos con vistas a ayudar a la interpretación de los resultados y lograr su objetivo esencial que es el de poder tomar decisiones, lo cual se hará con un grado de incertidumbre.

8.3 UNIVERSO Y MUESTRA.

Como universo (también conocido como población), se comprende en estadística, a un conjunto de elementos capaces de tener una o varias características en común bien definidas, y por muestra a cualquier subconjunto de la poblacion. definidas. P Así por ejemplo, el conjunto de madres de una provincia y el conjunto de madres de un municipio cualquiera de ellas, pueden constituir una poblacion y muestra respectivamente.

Estos términos son relativos, y a lo que en un momento se le denomina población puede ser una muestra en otro y viceversa. Así, si se analiza el ejemplo anterior, el conjunto de madres de una provincia será una muestra del conjunto de madres de toda Cuba, pero el conjunto de madres de un municipio puede representar una población y el conjunto de madres de un municipio menores de 20 años constituir una muestra.

Para algunos autores el universo constituye un concepto más general que población, y definen, entonces, tres términos: universo, población y muestra. Con lo expuesto basta para las necesidades de los temas a tratar, posteriormente estos conceptos seran retomados en el capitulo 10.

8.4 MÉTODO ESTÁTICO. ETÁPAS Y CARACTERÍSTICAS.

El método estadístico no es más que el método científico aplicado a una ciencia en particular, en este caso, la Estadística. Apropiado para resolver los problemas de las ciencias aplicadas, entre ellas las del campo biológico está dirigido a:

1. Obtener información
2. Organizar, resumir y presentarla en una forma adecuada.
3. Analizar e interpolar los resultados.

Se divide en 4 etapas:

1. Planificación de la investigación.
2. Recolección de la información.
3. Elaboración de los datos recogidos.
4. Análisis e interpretación.

Planificación de la investigación

Un paso natural en el proceso de llevar a cabo una investigación es su planificación, su organización, es decir, confeccionar un esquema organizativo tal que sea capaz de evaluar su factibilidad, que controle, facilite y evalúe su ejecución hasta alcanzar la meta propuesta.

Este esquema no puede ser rígido y debe permitir hacer cambios en los pasos a realizar en la consecución de la meta cuando se requiera.

En general la planificación constará de 4 subetapas:

1. **Planteamiento del problema**. Deberá partirse de la formulación del problema científico que vamos a investigar para lo que será necesario la definición de su naturaleza e importancia (**qué y por qué se va a estudiar**) y además la definición de los objetivos, o sea, las metas o fines que se pretenden alcanzar mediante la investigación.

   Los objetivos comúnmente se clasifican en generales, también planteados como finales o mediatos y específicos o inmediatos (**para qué y cómo** de la investigación).

2. **Búsqueda y evaluación de la información existente.**

   Después de plantear el problema y antes de efectuar la investigación se impone la revisión de toda la documentación posible sobre el tema y otros que puedan resultar afines o que bordean al mismo y proceden con la búsqueda y evaluación del material bibliográfico (debe recordarse que el problema investigativo aparece entre la frontera de lo conocido y desconocido científicamente, por lo que debe esclarecerse qué es lo que realmente se conoce). Este paso no solo se debe realizar en este momento, sino durante todo el proceso investigativo hasta su culminación.

3. **Formulación de hipótesis**. En ocasiones el propósito de la investigación es meramente descriptivo, pero en la mayoría de los casos lo que se busca es dar una respuesta que explique el fenómeno en estudio, y bajo todo esto subyace una hipótesis de trabajo la cual deberá ser corroborada o probada.

   Esta subetapa dependerá de distintos factores como son: las necesidades establecidas, los intereses actuales de los investigadores y los recursos materiales y humanos disponibles.

   Estas hipótesis de trabajo se traducirán en hipótesis estadísticas que son las que realmente serán contrastadas.

4. **Verificación de las hipótesis**. Esta etapa consistirá en la planificación de la prueba o contrastación de las hipótesis, que tendrá como resultado la aceptación o no de aquellas planteadas; esto conllevará al diseño de la investigación, o sea, a la planificación de la recogida, elaboración y análisis de la información.

   Aquí se estará planificando el desarrollo de la investigación, por lo que habrá que tomar en consideración a muchos y diferentes aspectos, desde definir cuál será la unidad de observación, qué características debe cumplir, qué cantidad de observaciones se van a tomar, cuál será el método de selección de las mismas, con qué recursos materiales y humanos se cuenta, hasta establecer qué condiciones deben ser impuestas o salvadas con vistas a que los elementos observados sean analizados de manera uniforme, en fin, a todo factor que pueda influir en el resultado final.

Recolección de la información

Después de realizar una cuidadosa planificación de la investigación o estudio se está en condiciones de comenzar a recolectar los datos, lo que constituye un paso a efectuar con mucho cuidado, escrúpulo y objetividad científica. Se deben evitarse, limitar o disminuir las posibles fuentes de error, según sea el caso (estos errores pueden estar asociados al observador, al método de observación o al elemento observado).

Es fundamental estar impuesto de que la validez de los resultados dependerá de la veracidad de la información tomada. La experiencia plantea que será particularmente importante atender a lo relacionado con:

- El universo, la muestra a tomar y los procedimientos utilizados para su selección.
- Los errores factibles en la recolección de la información y el modo de controlarlos.
- Los métodos y procedimientos utilizados en la recolección de los datos.
- El diseño de los formularios, documentos que contendrán la información recogida.

Elaboración de los datos recogidos

El procesamiento de la información atraviesa las subetapas siguientes:

1. Revisión y corrección de la información recolectada.
2. Clasificación y computación de los datos.
3. Presentación de la información.

No basta recolectar la información sino que esta debe ser despojada de los errores que pueden estar viciándola. Paso obligado será entonces la revisión de los datos acopiados, a fin de rectificar en los casos posibles y eliminarlos cuando no sea factible lo anterior.

En ocasiones no está en las manos del investigador la comparación con el dato original, pero sí es posible conocer si el dato con que se cuenta se encuentra dentro del rango establecido, o si uno faltante es deducible a través de alguna otra información adicional; de no poderse, generalmente habrá que desechar toda la información concerniente al individuo o elemento.

La masa de datos usualmente no dice mucho, sobre todo en la medida en que aumenta la cantidad de elementos; sin embargo, cuando es convenientemente clasificada y compilada, se obtiene un resumen que puede brindar o resaltar los detalles más significativos, pero desde luego, esto no sería importante si no se es capaz de presentar estos resultados de una forma clara, adecuada y entendible.

Análisis e interpretación

En esta fase, la información que fue sometida a un conjunto de métodos y procedimientos que permitieron desde recolectarla hasta elaborarla, sufrirá un proceso de análisis para definir e interpretar sus características más relevantes y representativas; en el caso que se hayan planteado algunas hipótesis, entoces será el momento de contrastarlas y llegar a conclusiones al respecto.

En la Estadística Descriptiva no se formulan hipótesis, por lo que el análisis e interpretación deberá limitarse a lo que muestra el grupo de datos estudiado (muestra) sin tratar de extender las conclusiones a un conjunto de datos más amplio (población).

A continuación se estudiarán algunas de estas etapas más profundamente.

8.4.1. Recolección de la información.

En esta etapa deben atenderse múltiples aspectos, que se desarrollarán a continuación:

8.4.1.1. Fuentes de recolección.

Como fue explicado anteriormente, todo estudio comienza con la planificación de lo que se va a realizar, en primer lugar de la información que se debe acopiar; para lograrlo, hay que conocer cuáles son los aspectos fundamentales que distinguen la información, sus fuentes y los modos de obtenerla.

Cuando se habla de la fuente de una información se están refiriendo al origen, al elemento que la produce, estas fuentes pueden ser **primarias** o **secundarias**, según si los datos se obtienen directamente del elemento que la origina, o de la información ya recolectada por otros, respectivamente.

En la mayoría de los casos los datos se obtienen de una fuente primaria, esto se puede lograr a través de dos procedimientos: la observación y el interrogatorio.

8.4.1.1.1. Observación

Es el procedimiento clásico de la investigación científica y es el más objetivo, pero no puede usarse siempre, ya que hay hechos que no están a la vista ni son deducibles, sino que solo son conocidos por los individuos bajo estudio, por ejemplo: cuánto tiempo hace que no acude al médico, si es alérgico a algún medicamento, si tiene frío, sed, o si está cansado, qué estudios realizará el próximo año, etc; tampoco es recomendable para analizar un conjunto grande de individuos, pues se encarece mucho la investigación al ser este un procedimiento que requiere de un personal altamente calificado.

8.4.1.1.2. Interrogatorio

Será el método a emplear cuando se necesita de las opiniones y conocimientos del individuo, desde luego, esto no es ideal porque su fidelidad dependerá de muchos factores como pueden ser el tipo y calidad de la pregunta (modo de plantearla es decir debe ser clara y concisa), buena memoria, del entrevistado y disposición de éste a ser veraz (desinhibido, serio) y hasta de su edad.

Si se realiza un estudio sobre alcoholismo, por ejemplo, es posible que la persona sienta cierto embarazo en afirmar que es o ha sido un alcohólico o que un pariente muy cercano lo es, por considerar la pregunta comprometedora o sentir vergüenza; si se necesita conocer qué enfermedades padeció en su niñez, una persona de edad avanzada pudiera no recordar algunas, o en otros casos no quedar claro si se están refiriendo a todo tipo de enfermedad o a ciertas en particular. También puede ocurrir que el individuo no tome con seriedad el interrogatorio y responda por responder.

Cada pregunta del **interrogatorio** debe ser bien analizada antes de realizarse. Estas han de responder a los objetivos de la investigación, no pueden ser ambiguas, ni que parezcan mal intencionadas, irrespetuosas o que sugieran la respuesta. Se debe contemplar si una pregunta puede ser generadora de otras y en qué orden deben efectuarse, en el citado estudio sobre alcoholismo, por ejemplo, será importante conocer si la persona bebe o no, qué bebe y cuánto bebe, este orden es importante; pues si no bebe, las otras preguntas sobran. El interrogatorio tampoco debe ser muy extenso ya que cansa al individuo y pierde el interés.

El interrogatorio puede ser **directo** y se realiza a través de **entrevistas** o **indirecto** mediante **cuestionarios**. Aunque ambos deben tomar en consideración todos los aspectos manejados anteriormente poseen características diferentes.

La **entrevista** tiene como ventaja que permite usar también la observación directa, puede ser más extensa y con preguntas más complicadas, sujeta a menos errores de omisión, pero las respuestas pueden estar condicionadas por suponer que no son anónimas (aunque no se registre el nombre, la persona piensa que sí), o por la impresión que cause el entrevistador sobre el entrevistado en cosas tan diversas como la forma de vestir, la inflexión de la voz y otras tantas que pueden variar las respuestas solicitadas. El resultado de la entrevista puede ser mejorado con la aplicación de técnicas cognitivas que ayuden a la memoria, como puede

ser ubicarlo en el tiempo a través de hechos importantes en la vida del sujeto. Generalmente el entrevistador es un individuo que ha sido entrenado para realizar este trabajo.

El **cuestionario** tendrá como ventajas su anonimato y la falta de "presión" para responder las preguntas, pero como desventaja estas tendrán que ser más sencillas, el documento no puede ser muy extenso, so pena que el interrogado se canse; muchas veces no se devuelve el cuestionario o se omiten respuestas, y por lo general, como esta información faltante no se puede recuperar, se debe eliminar la información completa sobre dicho individuo.

La entrevista tiene muchas variantes, ya que se puede realizar con la presencia de otra persona o que quien responda sea otra, o efectuarse por teléfono; por su parte el cuestionario puede ser entregado y recogido personalmente o por vía postal, etc. En general, este último introduce más error, pero ambos métodos deben probarse primero, o sea, hacerse un pilotaje.

8.4.1.1.3. Fuentes secundarias

Muchas veces es necesario acudir a una información ya recogida por otros, será importante, entonces, cerciorarse de su fidelidad (si los datos no son confiables es mejor desecharlos), y de la posibilidad de accesar a ella.

No tiene sentido volver a acopiar una información que ya existe y se tiene acceso a ella, significaría duplicar esfuerzo, gasto de tiempo, recursos materiales y humanos y, la consecuencia, encaren la investigación; pero si no cumple con los requisitos que se le exigen a cualquier información, habrá que buscar una fuente primaria.

8.4.1.2. Métodos de recolección según la frecuencia: encuesta, censo y registro.

La observación y el interrogatorio son los procedimientos de recolección de la información que de manera conjunta o usando uno sólo de ellos formarán parte de los métodos de dicho proceso. Estos métodos se diferencian entre sí no solo por la frecuencia de recogida, sino por el propósito de la investigación y la naturaleza de la información a acopiar.

De forma general puede plantearse que existen tres métodos fundamentales para realizar la recopilación de la información: encuesta, censo y registro y en ellos se utilizarán dos procedimientos de recogida: la observación y el interrogatorio (entrevista y cuestionario), esa información será volcada en el formulario.

Encuesta. Es el método que se realiza ocasionalmente siguiendo un propósito específico y un alcance restringido a un sector de la población. Por ejemplo, un médico de familia puede efectuar una encuesta para saber qué conocimientos tiene el sector de la población que atiende sobre el VIH-SIDA o sobre enfermedades de transmisión sexual.

El experimento también es un método de recolección ocasional pero a diferencia de la encuesta, cuya información se recoge a través del interrogatorio, este la toma, generalmente, mediante la observación, aunque en ocasiones utiliza el interrogatorio. En la encuesta los datos ya existenm, mientras que en el experimento no es así, y habría que "provocar" su aparición para poder recogerlos. Por ejemplo, para conocer si cierto medicamento tiene efectos positivos sobre personas hipertensas será necesario planificar un experimento.

En ocasiones se desea extrapolar de cierta forma los resultados obtenidos en las encuestas; en estos casos se distinguen dos tipos de encuestas, las correspondientes a muestras representativas y las correspondientes a grupos seleccionados, en el primero, a través de la muestra (debidamente seleccionada) se pretende inferir los resultados a toda la población; mientras que en el segundo, la muestra corresponde a grupos donde el fenómeno a estudiar es más frecuente y se utiliza para tener mayor conocimiento de dicho fenómeno.

Censo. Corresponde a una información que se recoge periódicamente, además este método completa y publica datos demográficos, económicos y sociales de un tiempo específico de todas las personas de un territorio determinado. Por ejemplo: se desea recopilar un conjunto de datos (edad, peso, tiempo de embarazo, embarazos anteriores, enfermedades que padece o padecidas, estado civil, dieta, condiciones de la vivienda, nivel de escolaridad y otros datos económico sociales) de las embarazadas del país en un momento dado, para lo que se planifica un censo nacional de las embarazadas.

Registro. Es aquel método que toma la información de manera continua (según se va produciendo) y sistemática sobre una cuestión determinada en la que puede referirse a lo que acontece en algo tan general como un país o algo tan particular como es una institución; por ejemplo: el registro de nacimientos de un municipio o la historia clínica de un paciente en su consultorio médico.

Resumiendo, estos métodos de recopilación de datos se van a diferenciar entre sí en cuatro aspectos fundamentalmente, que resumiremos en la tabla siguiente:

Aspectos	Encuesta	Censo	Registro
Frecuencia de recolección	Ocasional	Periódico	Continuo
Cobertura	Parcial	Universal	Universal
Temporalidad	Transversal	Transversal	Longitudinal
Propósito	Específico	General	Específico

Nota: parcial: región; transversal: un momento determinado; longitudinal: a lo largo del tiempo.

8.4.1.3. Formulario.

Toda información de interés para ser recopilada se registra en algún modelo para su posterior resumen o compilación; a este documento se le denomina formulario; ejemplos de él se tienen en las historias clínicas.

Generalmente, este instrumento recoge dos tipos de información: la correspondiente a los datos administrativos o de identificación y la referida a los datos propios del estudio que se está llevando a cabo.

De la información acopiada dependerán los resultados de la investigación o estudio que se realiza, por lo que el formulario debe ser elaborado con especial interés y cuidado, y será conveniente, considerar los aspectos siguientes:

- Que esté dividido al menos en dos secciones: la de los datos de identificación y la de los datos específicos del estudio.
- Que se tome en cuenta la secuencia en que se hacen las preguntas.

- Que las preguntas no sean reiterativas, ambiguas o que sugieran la respuesta. Deben ser claras, precisas y gramaticalmente correctas. No se debe indagar por aquellos datos que posteriormente no podrán ser utilizados porque no correspondan al interés del estudio, no sean confiables o no puedan completarse.
- No debe ser tan extenso que después pese o peligre llenarlo en su totalidad;
- Se debe decidir desde un inicio si el procesamiento será de forma automatizada para, de acuerdo al sistema a utilizar, realizar la codificación de las preguntas.
- El documento debe responder a los objetivos trazados.

Como puede apreciar, comúnmente el formulario se llena a través de preguntas y respuestas.

Alguno de los datos de identificación cambiarán de acuerdo con la investigación, ya que si en un caso pudieran ser nombres, dirección particular, centro de trabajo, sexo, edad, raza, lugar de nacimiento, nivel de escolaridad, etc., en otros se pudiera requerir más o menos datos o cambiar uno por otro (por ejemplo, nombres por el número de la historia clínica), a veces son anónimos y parte de esta información no es solicitada o elementos como el sexo y la edad, estre otros, corresponden a datos propios del estudio.

En ocasiones, el formulario debe estar acompañado, en el propio documento o aparte, de un instructivo que hará más comprensible el llenado del mismo.

Existen casos en que el formulario requiere ser anónimo para velar por la veracidad de las respuestas o por problemas éticos. Esto tiene como inconveniente que si se necesita rectificar algún dato será imposible, y probablemente haya que desechar a todos los provenientes del individuo. Ahora bien, a la hora de precisar la información lo que hace falta es algo (un nombre, un número, un alias, el número de la historia clínica o del expediente, etc.) que permita que se procese al individuo una sola vez y que cuando se revisen los datos, se tenga contra quién chequear y no haya error.

Aunque para el procesamiento de la información de forma automatizada las preguntas deben ser cerradas, esto no impide que algunas se dejen abiertas y su análisis posterior se haga aparte.

Ejemplo de formulario:

Supongamos se realiza un estudio acerca de la ingestión de café y su relación con algunas enfermedades.

La elección de las variables y la forma de la respuesta dependerá de los objetivos que persiga la investigación.

Los datos administrativos podrían ser:

Variante 1:

1) Nombre _____ 2) Consultorio _____

3) Sexo ___ 4) Edad ___ 5) Raza____ 6) Nivel de escolaridad _____**Variante 2:**

1) Nombre _____ 2) Consultorio_____

3) Sexo: Fem _____ 4) Edad 15-19 _____ 5) Raza B _____

Masc ____ 20-54 _____ N _____

55-64 _____ M _____

65 o más ____

6) Nivel de escolaridad: No _____ Nivel medio _____

(grados terminados) Primaria _____ Universitario _____

Secundaria _____

Entre los datos del estudio

Variante 1:

7) Toma café _____ 9) Cuantas tazas al día _____

8) Desde cuándo _____ 10) En qué momentos _____

Variante 2:

7) Toma café: 8) Desde cuándo:

Sí _____ <1año _____ > 5 años ____

No _____ 1-5 años ____ No corresponde ____

9) Cuántas tazas: 10) En qué momentos:

1 ____ Al desayunar ____

2 -3 ____ Detrás de las comidas ____

4 ó más ____ En cualquier momento ____

No corresponde____ No corresponde____

La respuesta será una cruz o marca.

De ambas variantes la segunda ofrece una forma de contestar que garantiza una respuesta más confiable, ya que "cierra" o enmarca la respuesta (un caso muy común será cuando se indaga por enfermedades que se padecen o se padecieron, en el cual se deben explicitar las principales de acuerdo con el estudio, y las

restantes agruparlas en otras). La respuesta **no corresponde** competería a los que no toman café (en la mayoría de los casos hay que tomar en cuenta este acápite) y la pregunta 10 podrá tener más de una respuesta, no así las anteriores.

Cuando se van a codificar las posibles respuestas de la pregunta se rellena con el código. Por ejemplo si se tiene Sexo (1. Fem, 2 Masc) ___, y se trata de una mujer se procedería asi Sexo (1. Fem, " Masc): 1

8.4.1.4. Errores en la recogida de la información.

En el campo biológico, al igual que en otros campos, existe una variación natural entre los valores de un individuo y otro acorde con la variable que se estudie, pero en el proceso de recogida de estos datos también se cometen errores que afectan en menor o mayor grado dichos valores y que pueden estar asociados al sujeto (quien realiza la recogida de la información), al objeto (la información) o al método de recogida empleado.

Con respecto al sujeto, estos errores dependerán entre otras cosas, de su preparación general y también específica sobre el trabajo que se va a realizar, así como de su capacidad; pero no deben olvidarse aspectos subjetivos tales como las condiciones de trabajo o las físicas y mentales del individuo. En el caso del objeto, estarán relacionados con las condiciones en que se encuentra y también el momento seleccionado para efectuar la recogida.

Por último, será no menos importante la fuente de error proveniente del método de recogida, pues incluso los errores mencionados anteriormente pueden depender de él. Aquí habrá que considerar si el método y procedimientos empleados son los más apropiados de acuerdo con la información que se va a recopilar.

Los instrumentos o medios que se utilicen para recoger o medir la información introducirán errores que se deberán detectar, controlar, medir y disminuir o eliminar, según sea el caso.

En muchas ocasiones estos errores son sistemáticos, como el que puede introducir un instrumento de pesaje o uno de medida de longitud, y se pueden establecer correcciones, pero a veces pasan inadvertidos; en otras, los errores pueden provenir de respuestas no fiables de un interrogatorio.

8.4.1.5. Papel del médico en la recolección de la información.

El médico participa activamente en la recolección de los datos, que se inicia desde que el enfermo entra en contacto con él.

Se hace necesario una descripción correcta y sistematizada de la enfermedad actual para una buena obtención, almacenamiento y procesamiento de los datos. En medida dependerá de:

- La información que se desprende del interrogatorio o anamnesis, el cual debe estar bien dirigido y orientado.
- Los datos del examen físico, que debe seguir una pauta exploradora.
- La correcta anotación de los diversos acontecimientos diagnósticos y terapéuticos en la historia clínica.
- Una evaluación en general antes de su egreso, cuando sea el caso.

De la limpieza, escrupulosidad, objetividad y celo del médico en la recogida de la información dependerá el futuro del éxito de los estudios, investigaciones y los resultados que brindan las distintas instancias del sistema de salud del país.

8.4.2. Procesamiento de la información.

Todo estudio o investigación lleva como respuesta un determinado volumen de información del que no se obtendrá nada en claro, si después de revisada y corregida dicha información esta no se procesa o elabora.

El procesamiento de la información transita por tres etapas, estas son:

1. Organización
2. Resumen
3. Presentación

8.4.2.1. Organización de la información. Características.

Mientras mayor sea el conjunto de datos con que se cuenta, menos factible será conocerlo y percatarse de cuáles son sus características comunes o cuán diferente es un dato de otro, es por eso que se hace necesario organizar la información. Para lograrlo es imprescindible conocer primero ciertos conceptos.

Variable

Es una característica que puede tomar valores diferentes. Las variables pueden ser numéricas o no.

Escala cualitativa

Es la que permite clasificar los elementos solo de acuerdo con los atributos comunes que exhiben cada uno de ellos. Por ejemplo, el sexo, que toma los valores femenino y masculino permite clasificar los individuos de una población en dos clases: la de las femeninas y la de los masculinos, asimismo pudieron analizarse la categoría ocupacional, el nivel de escolaridad y la raza como variables cualitativas y los valores que asumen son datos cualitativos.

Esta escala se subdivide en dos tipos:

1. **Cualitativa nominal.** Cuando la escala cualitativa no presenta categorías ordenadas, es decir, no es posible establecer diferencias de rango entre ellas. El sexo, la raza, el estado civil son variables cuyos posibles valores no establecen un orden entre sí.
2. **Cualitativa ordinal.** Cuando existen categorías ordenadas que permiten establecer comparaciones entre ellas. Ejemplo de ella se observa cuando se habla del grado de quemaduras que pueden tener un grupo de pacientes: quemaduras de 1er., de 2do. o de 3er. grado; otro ejemplo, sería el nivel de escolaridad: primario, ..., universitario; etc.

Escala cuantitativa

Aquí la escala no se distingue por poseer cierto atributo o no, sino por su cantidad. Será posible determinar entonces en cuánto se diferencia un elemento de otro. Valores como el nivel de hemoglobina, el nivel de colesterol y la temperatura son cuantitativas. Una escala cuantitativa puede clasificarse en dos tipos:

1. Una escala se considera discreta cuando solo admite un número finito de valores

numéricos o infinito numerable. Como ejemplo puede apuntarse la cantidad de veces que una persona asiste al médico en 1 año o el número de embarazos que ha tenido una mujer en su vida fértil.

2. Continua. Se distingue porque entre dos valores dados siempre es posible encontrar valores intermedios. El peso es una variable cuantitativa continua porque, por ejemplo, entre 62 y 63 Kg, existen los valores 62.2 ó 62.5 pero entre 62 y 62.2 estarán valores tales como 62.15 ó 62.07. Se plantea que esta escala surge por medición y la cuantitativa discreta por conteo.

8.4.2.1.1. Distribuciones de frecuencia.

Una forma de organizar los datos es agruparlos en diferentes clases, según una escala única, es decir, construir una distribución de frecuencia que será cualitativa o cuantitativa discreta o cuantitativa continua en dependencia de la escala empleada. Particularmente será importante cuando el conjunto de datos es grande.

Confeccionar la distribución de frecuencias consistirá en definir las clases en que serán agrupados los elementos, clasificarlos y, por último, calcular cuántos pertenecen a cada clase.

Distribución de frecuencias para datos cualitativos

Cuando la variable bajo estudio es cualitativa, las clases corresponderán a las categorías que esta presente.

Suponga que un médico de familia quiere conocer la composición por raza, de los pacientes con hipertensión arterial que él atiende, para ello utilizaría una distribución de frecuencias.

Primero tendrá que considerar cuáles son las distintas clases; pudieran ser: blanca, negra, mestiza y amarilla. En esto hay que observar un detalle muy importante y es que una buena escala debe ser exhaustiva y constar de clases mutuamente excluyentes, o sea, debe permitir la clasificación de cualquiera de los elementos, pero además estos deben ser clasificados en una clase y solamente en esa.

En el caso del ejemplo, si no existe ningún individuo de la raza amarilla, se puede obviar esta categoría; pero de existir, sería un error omitirla, ya que el individuo quedaría fuera de la clasificación (la escala no sería exhaustiva).

Después de definidas las clases o categoría se procede a la ubicación de cada elemento en la clase que le corresponda y, por último, al conteo, para conocer el número de ocurrencias en cada una de ellas, valor conocido como frecuencia absoluta de la clase. En el ejemplo citado sería como sigue:

Clases	absoluta	Relativa
Blanca	58	0.547
Negra	17	0.160
Mestiza	29	0.274
Amarilla	2	0.019
Total	106	1.000

Esta sería la distribución de frecuencias, las clases vendrían dadas por las diferentes categorías que aporta la raza, los valores de frecuencia absoluta significarían que 58 de los pacientes hipertensos son de la raza blanca, 17 de la raza negra y así sucesivamente.

La frecuencia relativa de la clase se define como el cociente de la frecuencia absoluta de la clase y el total:

$F_r = f_a$ /Total

Su resultado es el peso que tiene la clase respecto al total. Se diría aquí que la raza blanca tiene un peso de 0.55 (58/106) sobre el total de pacientes hipertensos en ese consultorio médico.

En ocasiones se completa esta distribución de frecuencias añadiendo una columna con los porcentajes.

Vale destacar en los totales que la suma de las frecuencias absolutas tiene que coincidir con el total de elementos bajo estudio, la suma de las frecuencias relativas debe ser igual a la unidad (1) y la suma de los porcentajes al cien por ciento (100 %).

Distribución de frecuencias para datos cuantitativos continuos

En este caso, las clases no pueden establecerse tan fácilmente como para la escala cualitativa.

Suponga que se cuenta con las estaturas de 44 embarazadas provenientes de un estudio de embarazo ectópico:

167	174	148	168	153	155	149	185	177	162
173	161	169	154	163	164	178	164	168	159
194	177	184	181	156	159	184	176	167	164
168	159	161	165	159	167	171	166	174	162
169	163	159	165						

Para dar una referencia de cómo es el comportamiento de la estatura es muy difícil (como ya se expresó, mientras mayor número de datos se tenga más difícil será), se hace necesario organizar los datos y tampoco dirá mucho si lo que se hace es relacionar todos los datos, pero si se establecen varias agrupaciones o conjuntos de valores donde se pueden mover los datos, la situación se torna diferente; en el ejemplo pudiera ser (140, 149); (150, 159) y así, los conjuntos de valores que se definan; lo que interesaría saber es cuántos de los datos se encuentran en la primera agrupación, cuántos en la segunda, etc.

Al conjunto de valores o agrupaciones se les conoce como clases o intervalos de clase.

Será importante, entonces, definir cuántos y cuáles serán los intervalos de clase en las que se distribuirán los datos.

Existen diferentes criterios en cuanto a la cantidad de clases o intervalos que debe tener una distribución de frecuencia de datos cuantitativos: entre 5 y 15, entre 6 y 15, entre 8 y 15, entre 10 y 20. Lo importante será no seleccionar tantos que la distribución aporte poco más que si se relacionaran los datos (mucha diferencia interelementos) o tan pocos intervalos que apenas se noten las diferencias entre ellos (poca diferencia interelementos), por supuesto, cuando los datos son muy parecidos, necesitará tener la mayor cantidad razonable de intervalos y si son muy diferentes la menor, para en el primer caso resaltar las diferencias entre los elementos, mientras que en el segundo destacará las similitudes. Si son pocos datos no deberá definir muchos y por el contrario si hay gran cantidad no deberá definir pocos intervalos.

Es decir, seleccionar el número de intervalos dependerá de la cantidad y del comportamiento de los datos, por lo que será importante conocer el recorrido o rango de los mismos, este se define como:

Rango = límite superior de los datos – límite inferior de los datos

(valor máximo) (valor mínimo)

Después de elegir la cantidad de intervalos de clase, debe definirse la amplitud de los mismos y en qué valores comienzan y terminan, o sea, aquellos que representan el valor mínimo y el máximo de cada clase denominados límite inferior y límite superior respectivamente.

La amplitud de las clases se determina por la fórmula siguiente: $$a = \frac{\text{Rango}}{\text{Número de clases}}$$

En el ejemplo dado, el recorrido es igual a Rango = 194 – 148 = 46cm y si se utilizarán 8 clases, la amplitud se calculará como:

$$a = 46/8 = 5.75\,cm$$

El resultado se aproxima al entero inmediato superior, en este caso a 6. Entonces, la amplitud de cada intervalo será 6 cm y los datos se distribuirán en 8 clases aproximadamente, si el primer intervalo comienza muy cercano al límite inferior de los datos.

La distribución de frecuencias podría ser:

Intervalos	**Frecuencia absoluta**
154-160	
160-166	
166-172	
172-178	
178-184	
184-190	
190-196	

Pero el problema estribará en que específicamente los valores 154; 166; 178; y 184 ¿dónde se clasificarían? ¿en los intervalos (148,154); (160,166); (172,178) y (178,184) respectivamente o en los intervalos (154,160); (166,172); (178,184) y (184,190)?; es decir, se impone establecer un convenio que permita determinar a qué clase pertenece un elemento dado y que cumpla que todo elemento del conjunto pertenezca una clase, pero además solo a esa.

Que el valor en duda pertenezca al intervalo de arriba o al de abajo es solo cuestión del criterio que se defina y para hacerlo existen diferentes formas. Será mucho más fácil indicarlo con el ejemplo que se está desarrollando.

En el caso que se desee que el valor pertenezca al primer intervalo o de arriba, ellos deberán ser definidos como alguna de las variantes siguientes, entre otras:

(1)	(2)	(3)
148 - 154	148 - 154	148 - 154

>154 - 160 154^{+} - 160 154.001 - 160

>160 - 166 160^{+} - 166 160.001 - 166

. . .

. . .

. . .

Aquí 154.160 y 166 siempre clasificaran en el primer intervalo de la escala en que aparezcan dichos valores.

Si se desea que el valor en duda pertenezca al segundo intervalo en que aparece, algunas variantes pueden ser las siguientes:

(1) (2) (3)

148 – <154 148 – 154^{-} 148 – 153.999...

154 – <160 154 – 160^{-} 154,001 – 159.999...

160 – <166 160 – 166^{-} 160,001 – 155.999...

En este caso los valores 154, 160 y 166 pertenecerán al segundo intervalo en que aparecen dichos valores.

Otra forma de resolver la problemática anterior sería si los intervalos se definieran como:

148 – 153

154 – 159

160 – 165

.

.

.

Esta distribución se construye comenzando por 148 y cada nuevo límite inferior será el resultado de sumarle 6 al anterior).

Los números de la izquierda son los límites inferiores de las clases y los que aparecen a la derecha son los límites superiores. En este último caso se contempla, además, el uso de los llamados límites reales, que serían entonces los verdaderos extremos de las clases, aquí en el ejemplo se tendría:

147.5 – 153.5

153.5 – 159.5

159.5 – 165.5

Los límites reales nunca pueden coincidir con los datos, de manera tal que los límites anteriores se definieron así porque no existía entre los valores ninguno con decimales; si así hubiera sido los límites reales se hubieran definido hasta la centésima y así sucesivamente. Se calculan como la semisuma del límite superior de una clase y el límite inferior de la siguiente:

$(153 + 154)/2 = 153.5$

$(159 + 160)/2 = 159.5$

En los primeros intervalos, la amplitud de la clase está dada por la diferencia entre sus límites:

154 – 148 = 6

160 – 154 = 6

y en este último por las diferencias entre los límites reales

153.5 – 147.5 = 6

159.5 – 153.5 = 6

o por las diferencias de los dos límites reales inferiores (o superiores) sucesivos.

$M_c = \frac{148 + 154}{} = 151$ o $M_c = \frac{147.5 + 153.5}{} = 150.5$

Volviendo a la distribución de frecuencias:

Intervalo	F.A.	F.R.	F.A.A.	F.R.A.	M_C	P_C
148 - <154	3	0.068	3	0.068	151	6.8
154 - <160	8	0.182	11	0.250	157	18.2
160 - <166	11	0.250	22	0.500	163	25.0
166 - <172	10	0.227	32	0.727	169	22.7
172 - <178	6	0.136	38	0.863	175	13.6
178 - <184	2	0.045	40	0.908	181	4.5
184 - <190	3	0.068	43	0.976	187	6.8
190 - <196	1	0.023	44	0.999	193	2.3
TOTAL	**44**	**0.999**				**99.9**
		≈ 1.000				≈ 100.0

F.A.: Frecuencia Absoluta; F.R.: Frecuencia Relativa; F.A.A.: Frecuencia Absoluta Acumulada; F.R.A.: Frecuencia Relativa Acumulada; M_c:Marca de Clase o Punto Medio y P_c: Porcentaje.

Las frecuencias absolutas acumuladas y las relativas acumuladas se calculan sumando todas las frecuencias absolutas y relativas hasta la clase señalada (incluyéndola) respectivamente. Por ejemplo, la frecuencia absoluta acumulada de la clase (160,166) se calcula sumando 3 + 8 + 11 = 22 que son las de las clases anteriores hasta ella, y se interpretaría como que 22 embarazadas del estudio tienen una estatura por debajo de (o hasta) 166 cm, o que 22 embarazadas tiene una estatura entre 148 y 166 cm.

La frecuencia relativa acumulada de esa clase se calcularía sumando 0.068 + 0.182 + 0.250 = 0.500 y nos diría que el peso de los datos hasta ahí (de 148 a 166 cm) es la mitad del total, o sea, que la mitad de los datos se concentra ahí.

Las frecuencias acumuladas son factibles de calcular cuando la variable es cuantitativa o incluso cuando es cualitativa ordinal, ya que en ambos casos existe un orden establecido, no así en las cualitativas nominales donde no tendría sentido.

En algunos casos, el intervalo inicial y/o el final puede ser abierto, esto ocurre con frecuencia cuando los primeros intervalos y/o los últimos apenas tienen valores o no tienen, por ejemplo si la distribución anterior fuese:

Intervalo	**F.A.**
148 – <154	1
154 – <160	0
160 – <166	0
166 – <172	21
172 – <178	12
178 – <184	8
184 – <190	0
190 – <196	2
TOTAL	44

La distribución anterior pudiera transformarse así:

Intervalo	**F.A.**
< 166	1
166 – <172	21
172 – <178	12
178 – <184	8
184 y más	2
TOTAL	44

O de otro modo:

Intervalo	**F.A.**
<166	1
166 – <172	21
172 – <178	12
178 – <184	8
≥184	2
TOTAL	44

Al agrupar los datos de esta manera los valores originales se pierden, aunque esto será menos importante, mientras mayor sea la cantidad de datos.

Por supuesto, la distribución será diferente si se utiliza otro valor de amplitud de intervalo.

Distribución de frecuencias para datos cuantitativos discretos.

En el caso de que la variable pueda tomar pocos valores diferentes, cada valor puede constituir una clase, por ejemplo, número de visitas al médico en un mes por paciente, número de hijos de una familia, cantidad de embarazos en una mujer, etc.; pero cuando no es el caso, como puede ser la cantidad de pulsaciones por minuto o el total de glóbulos rojos por paciente es necesario definir los intervalos aun cuando se sabe que solo serán factibles algunos valores dentro del mismo.

8.4.3. Resumen de los datos.

Para destacar las características de los datos será necesario condensarlos, o sea, resumirlos, hecho que puede lograrse por medio de diferentes medidas descriptivas, que dependerán de si los datos son cuantitativos o cualitativos.

8.4.3.1. Medidas descriptivas para datos cuantitativos.

Existen diferentes medidas resumen para variables cuantitativas como pueden ser las de localización o posición y las de dispersión.

8.4.3.1.1. Medidas de tendencia central: media o promedio aritmético, mediana, moda. Propiedades.

Son medidas de localización o posición central, es decir, definen el medio o el centro del conjunto de datos, que es un valor típico o representativo del mismo.

Exponentes de este tipo de medidas son la media aritmética, la mediana, la moda, la media geométrica, la media armónica, la media cronológica y la media ponderada. Se estudiarán las tres primeras.

Media arítmetica

Se conoce comúnmente como promedio o promedio aritmético. En este caso el centro del conjunto de datos se define como la suma de todos los valores dividido entre el total de datos.

$$\text{Media} = \frac{\sum \text{todos los datos}}{\text{Número total de datos}}$$

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$$

donde x_i representa a los diferentes valores del conjunto.

En el ejemplo de las embarazadas

$$\bar{x} = \frac{\sum_{j=1}^{44} x_j}{44} = \frac{167 + 174 + \cdots + 165}{44} = \frac{7341}{44} = 166.8$$

La estatura media de esas embarazadas es de 166.8 cm.

Propiedades de la media aritmética

1. Siempre existe.
2. Es única; existe una y solo una media aritmética.
3. Si a cada elemento de un conjunto de datos se le suma una constante, la media aritmética de nuevo conjunto será igual a la media aritmética del primer conjunto más la constante.

   $$x_2 = \overline{x_1 + c} = x_1 + c$$

   Ejemplificando, si se tienen los datos 12; 8; 11; 7 para el primer conjunto, entonces $x_1 = 9.5$.

   Segundo conjunto, sumándole 3: 15; 11; 14; 10

   $$x_2 = x_1 + 3 = 12.5$$

   Tercer conjunto, sumándole – 5: 7; 3; 6; 2

   $$x_3 = x_1 - 5 = 4.5$$

4. En un conjunto de datos, la sumatoria de todos los valores menos la media aritmética es igual a 0 y se expresa como:

   $$\sum_{j=1}^{n}(x_j - \overline{X}) = 0, \text{ donde si } X_j = x_j - \overline{X}, \text{ entonces } \sum_{j=1}^{n} X_j = 0$$

   $$\sum_{j=1}^{n}(x_j - \overline{X}) = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \overline{X},$$

   $$\text{pero } \overline{X} = \frac{\sum_{j=1}^{n} x_j}{n} \text{ por lo que}$$

   $$\sum_{j=1}^{n} x_j = n \cdot \overline{X} \text{ y como } \overline{X} \text{ es una constante, entonces}$$

   $$\sum_{j=1}^{n} \overline{X} = n \cdot \overline{X}, \text{ sustituyendo de arriba, se tiene que}$$

   $$n \cdot \overline{X} - n \cdot \overline{X} = 0$$

   Es decir, que si a cada elemento del conjunto le restamos la media aritmética y después los sumamos el resultado es 0.

   Por ejemplo, (12-9.5) + (8-9.5) + (11-9.5) + (7-9.5) = 0

5. Si se tienen conjuntos dados C_1, C_2 , ..., C_n , cada uno con m datos y $\text{con medias } \overline{X}_1, \overline{X}_2, \overline{X}_3, \cdots, \overline{X}_n$ respectivamente, la media aritmética general o media aritmética de todos los datos es igual a la media aritmética de las medias aritméticas de cada conjunto

   Esto es C_1 con datos $x_{11}, x_{21}, \ldots x_{m1}$

   C_2 con datos $x_{12}, x_{22}, \ldots x_{m2}$

   :

C_n con datos $x_{1n}, x_{2n}, ...x_{mn}$

Suponga que en el ejemplo de las estaturas de las embarazadas, estas estuvieran agrupadas en 4 conjuntos de 11 cada uno, donde las primeras 11 pertenecen al conjunto C_1 y así sucesivamente.

C_1: 167, 174, 148, 168, 153, 155, 149; 185, 177, 162, 173

C_2: 161, 169, 154, 163, 164, 178, 164; 168, 159, 194, 177

C_3: 184, 181, 156, 159, 184, 176, 167; 164, 168, 159, 161

C_4: 165, 159, 167, 171, 166, 174, 162, 169, 163, 159, 165

Entonces se tiene que:

$$\overline{X}_1 = 164.64,\ \overline{X}_2 = 168.27,\ \overline{X}_3 = 169.00,\ \overline{X}_4 = 165.45$$

$$\overline{X} = \frac{164.64 + 168.27 + 169.00 + 165.45}{4} = \frac{667.36}{4} = 166.8$$

6. Si cada elemento de un conjunto de datos se multiplica por una constante, la media aritmética del nuevo conjunto será igual a la media aritmética del primer conjunto multiplicada por la constante.

$$X_2 = \overline{k \cdot X_1} = k \cdot X_1$$

$$X_2 = \overline{k \cdot X_1} = \frac{\sum k \cdot x_1}{n} = \frac{k \cdot \sum x_1}{n} = k \cdot X_1$$

Aplicando lo planteado anteriormente al conjunto de datos: 12 ; 8 ; 11 ; 7, se tiene que: $\overline{X}_1 = 9.5$

Al multiplicar por 2 los datos iniciales se obtienen los valores siguientes: 24 ; 16 ; 22 ; 14. Entonces: $\overline{X}_2 = 19$

7. Es afectada por valores extremos. Esto se le señala como una desventaja.

   Si en el conjunto de datos en vez de 11 hubiera un 49, la media aritmética aumentaría a 19, valor bastante diferente a 12; 8 y 7 cuya media aritmética sería 9 y también bastante diferente a 49.

Mediana. Propiedades.

Se define como el valor que divide a un conjunto de datos ordenados a la mitad.

Para calcular la mediana de un conjunto de datos primeramente será necesario ordenar de menor a mayor o viceversa y después encontrar la posición central.

Si el número de datos es impar la mediana corresponde al valor que ocupa la posición (n+1)/2; si el número de datos es par, existirán dos valores centrales por lo que la mediana se define como el promedio de ambas (o semisuma), estos valores centrales ocupan las posiciones n/2 y (n/2) + 1.

En los datos de estatura del estudio sobre mujeres embarazadas para encontrar la mediana primeramente se ordenarán los valores:

148 149 153 154 155 156 159 159 159 159

159 161 161 162 162 163 163 164 164 164

165 165 166 167 167 167 168 168 168 169

169 171 173 174 174 176 177 177 178 181

184 184 185 194

n es par ya que es igual a 44; luego, los valores centrales serán los que ocupan las posiciones

n/2 = 44/2 = 22 y (n/2) + 1 = 22 + 1 = 23

que corresponden a los valores 165 y 166 por lo que la mediana será igual a:

$$\text{Mediana} = \frac{165 + 166}{2} = 165.5$$

El 50 % de las mujeres embarazadas miden menos de 165.5 cm y el 50 % de ellas tienen una estatura superior a ese valor.

Si en vez de 44 hubieran 43 valores (suponga que el valor 194 no existe) el valor central ocupa la posición: $\frac{n+1}{2} = \frac{43+1}{2} = 22$ que corresponde a 165.

se interpretaría entonces que 50 % de las mujeres embarazadas que se estudian tienen una estatura inferior a 165 cm y el otro 50 % una estatura superior a este valor.

Como puede verse en los ejemplos cuando n es impar la mediana es un valor que pertenece al conjunto de datos y cuando es par puede pertenecer o no.

Las principales propiedades de la mediana son:

1. Siempre existe
2. Es única
3. No se afecta por valores extremos. Por ejemplo, considere que entre las mujeres embarazadas en vez de un caso con 194 cm haya uno de 211 cm, entonces la mediana seguiría siendo 165.5 cm

Moda. Características.

Es el valor que más se repite dentro de un conjunto de datos, es decir, el de mayor frecuencia.
En el ejemplo visto 159 cm es la moda ya que es el valor que más se repite, 5 veces.
Entre las principales características de la moda se encuentran

1. No siempre existe, ya que si ninguno de los valores del conjunto de datos se repite no hay moda
2. No siempre es única. Por ejemplo, en la serie de datos siguientes:
   8, 15, 21, 22, 22, 25, 29, 29, 33
   los números 22 y 29 se repiten 2 veces por lo que ambos serían la moda, en este caso se dice que es bimodal; si hay tres modas, trimodal y en general, plurimodal.
3. En ocasiones puede usarse para datos cualitativos. Por ejemplo, en un consultorio determinado en el mes de febrero la enfermedad de moda pudo haber sido la respiratoria aguda ya que fue la más frecuente.

De estas tres medidas la media aritmética es la más ampliamente utilizada.
Cuando los datos están organizados en forma de distribuciones de frecuencia también se les puede calcular la media, la mediana y la moda; se dice entonces que se calculan estas medidas para datos agrupados.

Media aritmética para datos agrupados.

Sin recurrir a los datos originales se calcula esta media aritmética considerando que el punto medio o marca de clase es el elemento que mejor representa o sustituye al valor original dentro de cada clase. La fórmula es la siguiente:

$$\bar{x} = \frac{\sum_{j=1}^{k} f_j \cdot M_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} f_j \cdot M_j}{n}$$

Donde:

$f_{i:}$: frecuencia absoluta de la clase i.

M_i: marca de clase o punto medio de la clase i.

k: total de clases.

n: total de observaciones.

En el ejemplo de estaturas de las embarazadas, utilizando los límites reales:

$$\bar{X} = \frac{3 \cdot 150.5 + 8 \cdot 156.5 + \cdots + 3 \cdot 186.5 + 1 \cdot 192.5}{3+8+11+10+6+2+3+1}$$

$$= \frac{451.5 + 1252 + 1787.5 + 1685 + 1047 + 361 + 559.5 + 192.5}{44}$$

$$= \frac{7336}{44} = 166.7, \text{ valor diferente pero cercano al promedio de los datos simples que era de 166.8 cm}$$

Mediana para datos agrupados.

En este caso, primeramente debe encontrarse la clase mediana, es decir, la clase donde se encuentra la mediana y luego calcular la misma.

La fórmula será:

$$\text{Mediana} = L_{ME} + \frac{((n/2) - (\Sigma f)_{ME})}{f_{ME}} \cdot a$$

Donde:

L_{ME}: límite real inferior de la clase mediana.

n: número total de las observaciones.

$(\Sigma f)_{ME}$: suma de las frecuencias absolutas hasta la clase anterior a la clase mediana.

f_{ME} : frecuencia absoluta de la clase mediana.

a: amplitud de la clase mediana.

En el ejemplo de estaturas de las embarazadas, primero hay que encontrar la clase mediana, como hay 44 observaciones, esta clase será aquella donde se alcance 22 observaciones (para ello puede usarse la frecuencia acumulada), esta será (160, 165). Entonces, como:

$L_{ME} = 159.5$ $(\Sigma f)_{ME} = 11$ $a = 6$

$n = 44$ $f_{ME} = 11$

$$\text{Mediana} = 159.5 + \frac{(44/2) - 11}{11} \cdot 6$$
$$= 159.5 + 6$$
$$= 165.5 \text{ cm}$$

Este valor divide el conjunto de datos en dos partes tales que el 50 % de estas mujeres embarazadas miden menos de 165.5 cm y el resto por encima, en este caso el valor coincidió con el calculado para datos simples.

Moda para datos agrupados

Para calcular la moda también es necesario encontrar previamente la clase modal, o sea, aquella que contiene la moda, la fórmula en este caso será:

$$\text{Moda} = L_{MO} + \frac{(f_{MO} - f_A)}{(2f_{MO} - f_P - f_A)} \cdot a$$

Donde:

L_{MO}: límite real inferior de la clase modal.

f_{MO}: frecuencia absoluta de la clase modal.

f_A: frecuencia absoluta de la clase anterior a la clase modal.

f_P: frecuencia absoluta de la clase posterior a la clase modal.

a: amplitud de la clase modal.

En el ejemplo que se está desarrollando, la clase modal es (160, 165).

$L_{MO} = 159.5$ $f_A = 8$ $a = 6$

$f_{MO} = 11$ $f_p = 10$

$$\text{Luego se tiene, Moda} = 159.5 + \frac{(11 - 8)}{(2 \cdot 11 - 10 - 8)} \cdot 6 = 159.5 + \frac{3}{22 - 18} \cdot 6$$
$$\text{Moda} = 159.5 + 18/4 = 164 \text{ cm}$$

Como se observa, este valor esta algo alejado al encontrado para los datos simples (159 cm).

Cuando la distribución de los datos es simétrica, la media aritmética, la mediana y la moda

coinciden; si la distribución es asimétrica, donde la mayoría de los datos se encuentran hacia la izquierda (asimétrica a la derecha), la moda es menor que la mediana, la que a su vez es menor que la media aritmética; si la distribución es asimétrica, donde la mayoría de los datos se encuentra hacia la derecha (asimétrica hacia la izquierda), la moda es mayor que la mediana y esta es mayor que la media aritmética.

Entre otras medidas de tendencia central se encuentra la media arimética ponderada, que es una media aritmética que toma en cuenta que cada elemento de la serie tiene un peso diferente dentro del conjunto de datos, en la fórmula cada elemento estará afectado por ese peso, ponderación o factor.

$$\overline{x} = \frac{\sum_{i=1}^{n} \omega_i \cdot x_i}{n}$$

Donde:

ω_i: peso de la observación i.

x_i:observación i.

n: total de observaciones.

Esta fórmula es similar a la media aritmética para datos agrupados, si se consideran las x como las marcas de clase y las frecuencias absolutas de las clases, las ponderaciones. También puede usarse dicha fórmula cuando existen muchos datos repetidos y las ω_i representarían las frecuencias con que se repite cada valor x_i.

8.4.3.1.2 Medidas de dispersión: rango o recorrido, varianza y la desviación estándar.

Las medidas de tendencia central no bastan para caracterizar un grupo de datos; por ejemplo, si se tiene la siguiente serie de datos correspondientes al nivel de glucosa en sangre de un grupo de pacientes: 3.8; 7.6; 4.9; 6.3; 3.4, la media aritmética es 5.2 pero con la serie:

5.2; 5.0; 5.3; 5.4; 5.1

también la media es 5.2 y son dos conjuntos de datos totalmente diferentes. Estos últimos datos son muy parecidos entre sí, por ejemplo entre el valor máximo y el mínimo solo hay una diferencia de 0.4 unidades, mientras que en el primer conjunto los datos son bastante diferentes entre sí y alejados del valor medio, si se compara con el segundo conjunto.

Esto indica que para describir las características de los datos cuantitativos se hace necesario considerar otras medidas como son las de variación o dispersión, que miden cuánto se alejan o dispersan los datos respecto de un valor dado o entre sí. Las más comunes son: rango, varianza, desviación estándar, desviación media,

desviación cuartilar, coeficiente de variación y otros. Mientras más se parezcan los valores, o sea, mientras más cercanos sean ellos entre sí menor, será el valor de la medida de dispersión.

Rango o recorrido

Como vimos cuando se construyeron las distribuciones de frecuencia, estas se definió como la diferencia entre el valor mayor y el menor.

$$\text{Rango} = V_{\text{máximo}} - V_{\text{mínimo}}$$

En la primera serie de datos el rango es de 0.4 unidades (5.4 –5.0) y en la segunda 4.2 unidades (7.6 – 3.4).

Aunque es una medida fácil de calcular y casi obligada para tener una primera idea de la variación de la información puede resultar engañosa.

El primer conjunto de datos visto de una manera ordenada sería: 3.4; 3.8 ; 4.9; 6.3; 7.6

y el rango 4.2 unidades, pero si el conjunto fuera 3.4 ; 3.7; 3.4 ;3.5 ;7.6 el rango seguiría siendo el mismo, sin embargo, los cuatro primeros datos son bastante parecidos y 7.6 es un valor atípico.

Varianza.

Una medida de dispersión intuitiva sería aquella que aportara el promedio de la desviación de cada dato con respecto a la media aritmética pero el hecho es que se tiene que, $\sum(x_j - \bar{x}) = 0$,

(ya que esto se vió en las propiedades de la media aritmética), una forma de resolver esta situación se obtiene tomado el valor modular de cada desviación, o sea,

$$\sum_{j=1}^{n} |x_j - \bar{x}|$$

con esto todas las desviaciones serán positivas, esta medida dividida entre n es conocida como desviación media. Otra forma de solución es elevando al cuadrado estas desviaciones:

$$S^2 = \frac{\sum(x_j - \bar{x})^2}{n}$$

y se obtiene la varianza, promedio de las desviaciones cuadráticas de cada valor respecto a su media aritmética.

Cuando se trabaja solo con un conjunto de datos o muestra y no la población entera se acostumbra multiplicar este valor por el factor n/(n-1) cuyo resultado es un valor que representa una mejor aproximación de la varianza de toda la población, entonces,

$$s^2 = \frac{\sum(x_j - \bar{X})^2}{n-1}$$

Si en el conjunto de datos: 3.4; 3.8 ; 4.9; 6.3; 7.6 se va a calcular la varianza

$$s^2 = \frac{(3.4-5.2)^2 + (3.8-5.2)^2 + (4.9-5.2)^2 + (6.3-5.2)^2 + (7.6-5.2)^2}{5-1}$$

$$= \frac{(1.8)^2 + (1.4)^2 + (0.3)^2 + (1.1)^2 + (2.4)^2}{4}$$

$$= \frac{12.26}{4}$$

$s^2 = 3.06$

La varianza para datos agrupados se definiría como

$$S^2 = \frac{\sum_{j=1}^{k} f_j(M_j - \bar{X})^2}{n-1}$$

Donde:

f_i: frecuencias absolutas de cada clase.

M_i: marcas de clase.

$\bar{x}$: la media aritmética del conjunto de datos.

n: número total de observaciones, $n = \sum_{j=1}^{k} f_j$

k: cantidad de intervalos de clase.

Esta fórmula sirve cuando en una serie de datos simples se repiten muchos de ellos, así, f_i representaría la cantidad de veces que se repite la observación x y n la cantidad de observaciones diferentes.

Desviación estándar.

Se define como la raíz cuadrada de la varianza, por lo que tendrá la misma unidad de medida que la media aritmética. Su fórmula es:

$$s = \sqrt{S^2}$$

Utilizando el mismo ejemplo:

$$s = \sqrt{3.06} = 1.75$$

Propiedades de la varianza y la desviación estándar

Siempre son positivas, ya que se define la varianza como una suma de valores al cuadrado.

Si a cada elemento del conjunto de datos se le suma una constante, la varianza y la desviación estándar del nuevo conjunto de datos no cambia. En el ejemplo desarrollado, $s^2 = 3.06$ y $s = 1.75$; si a la serie de datos se le suma 0.5

3.9 ; 4.3; 5.4; 6.8; 8.1; s^2 sigue siendo 3.07 y s igual a 1.75.

$s^2(x + c) = s^2(x)$ y $s(x + c) = s(x)$, donde $s^2(x)$ es la varianza de la variable x.

Si cada elemento del conjunto de datos es multiplicado por una constante (c), la varianza y la desviación estándar del nuevo conjunto de datos serán iguales a c^2s^2 y cs respectivamente, y se tiene que :

$s^2(cx) = c^2s^2(x)$ y $s(cx) = cs(x)$

Suponga que cada elemento de la serie debe ser multiplicado por 1,2; los valores serían:

4.68; 5.16; 6.48; 8.16; 9.72

$s^2 = 4.41 = 1.44 \cdot 3.06$

$s = 2.10 = 1.2 \cdot 1.75$

Mediana para datos agrupados.

En este caso, primeramente debe encontrarse la clase mediana, es decir, la clase donde se encuentra la mediana y luego calcular la misma.

La fórmula será:

$$\text{Mediana} = L_{ME} + \frac{((n/2) - (\Sigma f)_{ME})}{f_{ME}} \cdot a$$

Donde:

L_{ME}: límite real inferior de la clase mediana.

n: número total de las observaciones.

$(\Sigma f)_{ME}$: suma de las frecuencias absolutas hasta la clase anterior a la clase

mediana.

f_{ME} : frecuencia absoluta de la clase mediana.

a: amplitud de la clase mediana.

En el ejemplo de estaturas de las embarazadas, primero hay que encontrar la clase mediana, como hay 44 observaciones, esta clase será aquella donde se alcance 22 observaciones (para ello puede usarse la frecuencia acumulada), esta será (160, 165). Entonces, como:

$L_{ME} = 159.5$ $(\Sigma f)_{ME} = 11$ $a = 6$

$n = 44$ $f_{ME} = 11$

$$\text{Mediana} = 159.5 + \frac{(44/2) - 11}{11} \cdot 6$$
$$= 159.5 + 6$$
$$= 165.5 \text{ cm}$$

Este valor divide el conjunto de datos en dos partes tales que el 50 % de estas mujeres embarazadas miden menos de 165.5 cm y el resto por encima, en este caso el valor coincidió con el calculado para datos simples.

Moda para datos agrupados

Para calcular la moda también es necesario encontrar previamente la clase modal, o sea, aquella que contiene la moda, la fórmula en este caso será:

$$\text{Moda} = L_{Mo} + \frac{(f_{Mo} - f_A)}{(2f_{Mo} - f_P - f_A)} \cdot a$$

Donde:

L_{MO}: límite real inferior de la clase modal.

f_{MO}: frecuencia absoluta de la clase modal.

f_A: frecuencia absoluta de la clase anterior a la clase modal.

f_P: frecuencia absoluta de la clase posterior a la clase modal.

a: amplitud de la clase modal.

En el ejemplo que se está desarrollando, la clase modal es (160, 165).

$L_{MO} = 159.5$ $f_A = 8$ $a = 6$

$f_{MO} = 11$ $f_p = 10$

$$\text{Luego se tiene, Moda} = 159.5 + \frac{(11-8)}{(2\cdot 11\ \ 10\ \ 8)}\cdot 6 = 159.5 + \frac{3}{22\ \ 18}\cdot 6$$

$$\text{Moda} = 159.5 + 18/4 = 164 \text{ cm}$$

Como se observa, este valor esta algo alejado al encontrado para los datos simples (159 cm).

Cuando la distribución de los datos es simétrica, la media aritmética, la mediana y la moda coinciden; si la distribución es asimétrica, donde la mayoría de los datos se encuentran hacia la izquierda (asimétrica a la derecha), la moda es menor que la mediana, la que a su vez es menor que la media aritmética; si la distribución es asimétrica, donde la mayoría de los datos se encuentra hacia la derecha (asimétrica hacia la izquierda), la moda es mayor que la mediana y esta es mayor que la media aritmética.

Entre otras medidas de tendencia central se encuentra la media arimética ponderada, que es una media aritmética que toma en cuenta que cada elemento de la serie tiene un peso diferente dentro del conjunto de datos, en la fórmula cada elemento estará afectado por ese peso, ponderación o factor.

$$\overline{X} = \frac{\sum_{j=1}^{n} \omega_j \cdot x_j}{n}$$

Donde:

ω_i: peso de la observación i.

x_i:observación i.

n: total de observaciones.

Esta fórmula es similar a la media aritmética para datos agrupados, si se consideran las x como las marcas de clase y las frecuencias absolutas de las clases, las ponderaciones. También puede usarse dicha fórmula cuando existen muchos datos repetidos y las ω_i representarían las frecuencias con que se repite cada valor x_i.

8.4.3.1.2 Medidas de dispersión: rango o recorrido, varianza y la desviación estándar.

Las medidas de tendencia central no bastan para caracterizar un grupo de datos; por ejemplo, si se tiene la siguiente serie de datos correspondientes al nivel de glucosa en sangre de un grupo de pacientes: 3.8; 7.6; 4.9; 6.3; 3.4, la media aritmética es 5.2 pero con la serie:

5.2; 5.0; 5.3; 5.4; 5.1

también la media es 5.2 y son dos conjuntos de datos totalmente diferentes. Estos últimos datos son muy parecidos entre sí, por ejemplo entre el valor máximo y el mínimo solo hay una diferencia de 0.4 unidades, mientras que en el primer conjunto los datos son bastante diferentes entre sí y alejados del valor medio, si se compara con el segundo conjunto.

Esto indica que para describir las características de los datos cuantitativos se hace necesario considerar otras medidas como son las de variación o dispersión, que miden cuánto se alejan o dispersan los datos respecto de un valor dado o entre sí. Las más comunes son: rango, varianza, desviación estándar, desviación media, desviación cuartilar, coeficiente de variación y otros. Mientras más se parezcan los valores, o sea, mientras más cercanos sean ellos entre sí menor, será el valor de la medida de dispersión.

Rango o recorrido

Como vimos cuando se construyeron las distribuciones de frecuencia, estas se definió como la diferencia entre el valor mayor y el menor.

$$\text{Rango} = V_{\text{máximo}} - V_{\text{mínimo}}$$

En la primera serie de datos el rango es de 0.4 unidades (5.4 –5.0) y en la segunda 4.2 unidades (7.6 – 3.4).

Aunque es una medida fácil de calcular y casi obligada para tener una primera idea de la variación de la información puede resultar engañosa.

El primer conjunto de datos visto de una manera ordenada sería: 3.4; 3.8 ; 4.9; 6.3; 7.6

y el rango 4.2 unidades, pero si el conjunto fuera 3.4 ; 3.7; 3.4 ;3.5 ;7.6 el rango seguiría siendo el

mismo, sin embargo, los cuatro primeros datos son bastante parecidos y 7.6 es un valor atípico.

Varianza.

Una medida de dispersión intuitiva sería aquella que aportara el promedio de la desviación de cada dato con respecto a la media aritmética pero el hecho es que se tiene que, $\sum(x_j - \bar{X}) = 0$,

(ya que esto se vió en las propiedades de la media aritmética), una forma de resolver esta situación se obtiene tomado el valor modular de cada desviación, o sea,

$$\sum_{j=1}^{n} |x_j - \bar{x}|$$

con esto todas las desviaciones serán positivas, esta medida dividida entre n es conocida como desviación media. Otra forma de solución es elevando al cuadrado estas desviaciones:

$$S^2 = \frac{\sum(x_j - \bar{X})^2}{n}$$

y se obtiene la varianza, promedio de las desviaciones cuadráticas de cada valor respecto a su media aritmética.

Cuando se trabaja solo con un conjunto de datos o muestra y no la población entera se acostumbra multiplicar este valor por el factor n/(n-1) cuyo resultado es un valor que representa una mejor aproximación de la varianza de toda la población, entonces,

$$s^2 = \frac{\sum(x_j - \bar{X})^2}{n-1}$$

Si en el conjunto de datos: 3.4; 3.8 ; 4.9; 6.3; 7.6 se va a calcular la varianza

$$s^2 = \frac{(3.4-5.2)^2 + (3.8-5.2)^2 + (4.9-5.2)^2 + (6.3-5.2)^2 + (7.6-5.2)^2}{5-1}$$

$$= \frac{(-1.8)^2 + (-1.4)^2 + (-0.3)^2 + (1.1)^2 + (2.4)^2}{4}$$

$$= \frac{12.26}{4}$$

$s^2 = 3.06$

La varianza para datos agrupados se definiría como

$$S^2 = \frac{\sum_{j=1}^{k} f_j (M_j - \bar{X})^2}{n-1}$$

Donde:

f_i: frecuencias absolutas de cada clase.

M_i: marcas de clase.

$\bar{x}$: la media aritmética del conjunto de datos.

n: número total de observaciones, $n = \sum_{j=1}^{k} f_j$

k: cantidad de intervalos de clase.

Esta fórmula sirve cuando en una serie de datos simples se repiten muchos de ellos, así, f_i representaría la cantidad de veces que se repite la observación x y n la cantidad de observaciones diferentes.

Desviación estándar.

Se define como la raíz cuadrada de la varianza, por lo que tendrá la

misma unidad de medida que la media aritmética. Su fórmula es:

$$s = \sqrt{S^2}$$

Utilizando el mismo ejemplo:

$$s = \sqrt{3.06} = 1.75$$

Propiedades de la varianza y la desviación estándar

Siempre son positivas, ya que se define la varianza como una suma de valores al cuadrado.

Si a cada elemento del conjunto de datos se le suma una constante, la varianza y la desviación estándar del nuevo conjunto de datos no cambia. En el ejemplo desarrollado, s^2 = 3.06 y s = 1.75; si a la serie de datos se le suma 0.5

3.9 ; 4.3; 5.4; 6.8; 8.1; s^2 sigue siendo 3.07 y s igual a 1.75.

$s^2(x + c) = s^2(x)$ y $s(x + c) = s(x)$, donde $s^2(x)$ es la varianza de la variable x.

Si cada elemento del conjunto de datos es multiplicado por una constante (c), la varianza y la desviación estándar del nuevo conjunto de datos serán iguales a c^2s^2 y cs respectivamente, y se tiene que :

$s^2(cx) = c^2s^2(x)$ y $s(cx) = cs(x)$

Suponga que cada elemento de la serie debe ser multiplicado por 1,2; los valores serían:

4.68; 5.16; 6.48; 8.16; 9.72

$s^2 = 4.41 = 1.44 \cdot 3.06$

$s = 2.10 = 1.2 \cdot 1.75$

8.4.3.1.3. Medidas de dispersión relativa: coeficiente de variación, variables estandarizadas y puntuaciones estándar.

Cuando se desea comparar la variación existente entre dos conjuntos de datos puede suceder que ambos grupos no estén medidos en la misma unidad o que uno de los mismos exhiba valores mucho más elevados que el otro, por lo que el resultado de la medida de dispersión debe ser mayor en este primer conjunto, sin que por ello sus valores sean menos parecidos entre sí que los del segundo conjunto; se impone entonces buscar una medida que permita ver la dispersión de una forma relativa y no absoluta.

De manera general la definición será:

$$\text{Medida de dispersion relativa} = \frac{\text{Medida de dispersion absoluta}}{\text{Medida de tendencia central}}$$

Coeficiente de variación.

De las medidas de dispersión relativa la más común resulta ser el coeficiente de variación, en este caso se toma como medida de dispersión absoluta a la desviación estándar y como medida de tendencia central a la media aritmética, así la definición es: $CV_i = \left(\frac{S_i}{\bar{X}_i}\right) \cdot 100$.

lo que representa un porcentaje y permitirá comparar cualquier conjunto de datos.

Por ejemplo, se cuenta con los resultados de una investigación sobre el nivel de colesterol en sangre en los habitantes de un municipio de un país, los cuales arrojan una media aritmética de 5.3 unidades y una desviación estándar de 2.05 unidades, en otros resultados de un municipio

contiguo un estudio similar arrojó un valor de 4.8 unidades para la media aritmética y de 1.9 unidades para la desviación estándar.

Los coeficientes de variación serán:

Grupo 1

$$CV_1 = (S_1/\bar{X}_1)\cdot 100$$

$$= [2.05/5.3]\cdot 100$$

$$= 38.7\ \%$$

Grupo 2

$$CV_1 = (S_2/\bar{X}_2)\cdot 100$$

$$= [1.9/4.8]\cdot 100$$

$$= 39.6\ \%$$

En el segundo conjunto se obeserva una variación algo mayor que en el primero.

Para valores de la media aritmética cercanos a 0 el coeficiente de variación deja de ser útil.

Variables estandarizadas y puntuaciones estándar.

Otra forma de medir la dispersión sin depender de las unidades de medida es lo que se conoce como la estandarización de la variable.

El valor de una variable estandarizada se obtiene a partir de la fórmula siguiente:

$$z = \frac{x - \bar{x}}{s}$$

donde esta f´romula o expresión se aplica a cada dato individualmente, la cual mide la desviación de la media en unidades de la desviación estándar, esta variable no depende de las dimensiones usadas al presentar el numerador y el denominador la misma unidad de medida.

Cuando las desviaciones de la media vienen expresadas en unidades de la desviación estándar se plantea que estas se encuentran en unidades o puntuaciones estándar y se emplean para comparar distribuciones.

8.4.3.2. Medidas de posición relativa: cuartiles, deciles y percentiles.

Los cuantiles son medidas de posición relativa los cuales dividen al conjunto ordenado de datos en varias partes iguales.

Los cuartiles, deciles y percentiles son cuantiles que dividen el conjunto ordenado de datos en cuatro (Q_i), diez (D_i) y cien (P_i) partes iguales respectivamente.

El cuartil 3 (Q_3) será el valor que divide al conjunto ordenado de datos de manera tal que 75 % de los valores del mismo está por debajo del cuartil y 25 % restante por encima; el decil 6 (D_6) será el valor que divide al conjunto ordenado de datos de forma tal que 60 % de los mismos es menor que él y 40 % restante es mayor;

el percentil 97 es el valor que divide al conjunto ordenado de datos de manera tal que 97 % de los datos se encuentra por debajo de él y el otro 3 % se encuentra por encima.

Como puede verse el cuartil 2, el decil 5 y el percentil 50 coinciden con la mediana.

Forma de cálculo.

Independientemente del cuantil que se desee calcular se utiliza el mismo procedimiento:

1. Ordenamiento del conjunto de datos.
2. Cálculo de la posición que ocupa el cuantil.

   Determinación del porcentaje del total de observaciones de acuerdo con el cuantil planteado:

   a) Si el número resultante es entero se toma en cuenta esta posición y la siguiente.

   b) Si por el contrario, el resultado es un número fraccionario la posición será el próximo valor entero.

3. Determinar el cuantil

Para el 2a) se promedian los valores de las dos posiciones encontradas.

Para el 2b) el cuantil será el valor que corresponde a la posición hallada.

En el ejemplo de las estaturas de las embarazadas, si se deseara calcular Q_3, D_6 y P_{97}, se deben seguir los pasos enunciados anteriormente:

Se ordenan los datos (ya se ordenaron cuando se calculó la mediana).

- Q_3

Q_3® 75 %

75/100 . 44 = 33 y como el valor es entero

Entonces se toman las posiciones 33 y 34 y se promedian los valores correspondientes.

$$Q_3 = \frac{173 + 174}{2}$$

Posición Valor

33 173

34 174

$Q_3 = 173.5$

El 75 % de las embarazadas tiene una estatura inferior a 173.5 cm.

- D_6

D_6 ® 60 %

(60/100)· 44 = 26.4, como el valor es fraccionario se debe tomar la posición 27, entonces D_6 = 168.

El 60 % de las embarazadass tiene una estatura menor de 168 cm y el resto superior:

- P_{97}

P_{97} ® 97 %

97/100 . 44 = 42.7, de nuevo el valor es fraccionario por lo que se aproxima a 43 y se busca el valor correspondiente a esta posición.

$P_{97} = 185$

El 97 % de las estaturas de estas embarazadas corresponde a un valor inferior a 185 cm y solo 3 % a un valor superior a este.

Utilizando los cuartiles se pueden definir el intervalo intercuartilar, que es aquel que se encuentra entre el cuartil 1 y el 3 y la desviación cuartilar o cuartílica o recorrido semi intercuartílico que es la media de la diferencia entre estos dos cuartiles ($DQ = (Q_3 - Q_1)/2$), tanto el recorrido intercuartílico ($Q_3 - Q_1$) como la desviación intercuartilar son medidas de dispersión pero este último es de uso más frecuente.

También se define el recorrido del percentil 10 – 90, esto será $P_{90} - P_{10}$ (o $D_9 - D_1$), además puede emplearse el recorrido del semi percentílico 10 – 90 que será igual a $(P_{90} - P_{10})/2$, aunque no es habitualmente utilizado.

Al emplear los cuartiles 1 y 3 puede definirse una medida de tendencia central:

$Q_M = (Q_1 + Q_3)/2$

También podrá definirse una medida de dispersión relativa tomando Q y la desviación intercuartílica:

$CV = [(Q_3 - Q_1)/2]/[(Q_1 + Q_3)/2]$

$= (Q_3 - Q_1)/(Q_1 + Q_3)$

La cual pudiera expresarse en forma de porcentaje si se multiplica por 100.

8.4.3.3. Medidas para resumir datos cualitativos: razón, índice, proporción, porcentaje y tasa.

Cuando los datos son cualitativos solo se pueden utilizar para referenciar o comparar las diferentes categorías a las frecuencias, ya que cada elemento lo que aporta es su pertenencia o no a cada categoría.

Aunque pudiera usarse la frecuencia absoluta, esta puede resultar engañosa. Por ejemplo, si se dice que la cantidad de mujeres hipertensas en el consultoriuo médico A es de 30 y en otro B es de 20, a primera vista se diría que la hipertensión arterial representa un problema de salud más preocupante para el consultorio A que para el B, lo que no sería cierto si en el consultorio A existiera un total de 150 mujeres y en el B un total de 80. En el primer caso, 1 de cada 5 mujeres es hipertensa, mientras que en el segundo lo es 1 de cada 4.

Por lo visto anteriormente, es preferible usar como medidas descriptivas a diferentes frecuencias relativas, las que se conocen también como indicadores. Cuando los indicadores son aplicados al campo de la salud reciben el nombre de indicadores de la salud.

Los indicadores pueden ser razones, índices, proporciones, porcentajes y tasas.

Razón.

Es una fracción del tipo a/b, donde a y b se refiren a hechos diferentes.

Si en el consultorio A hay 25 hombres hipertensos puede calcularse la razón de mujeres hipertensas con respecto a los hombres hipertensos en dicho consultorio, por medio de, $R = 30/25 = 6/5$.

Es decir, por cada 6 hombres hipertensos hay 5 mujeres hipertensas en ese consultorio.

Indice.

Es la razón multiplicada por 100: $I = (a/b) \cdot 100$

Así el ejemplo anterior, se tiene que: $I = \frac{20}{30} \cdot 100 = \frac{6}{5} \cdot 100$

Interpretandose como que, por cada 600 hombres hipertensos hay 500 mujeres hipertensas.

Proporción.

Es toda fracción del tipo p = a/n, donde a es el numero de elementos que tienen una (o varias)

características en comun respecto de un total de n elementos considerados.

Ejemplo:

En el consultorio A se vió que de las 150 mujeres, 30 son hipertensas por lo que la proporción de mujeres hipertensas es: P = 30/150 = 1/5

Como ya se vió, de cada 5 mujeres que corresponden a ese consultorio, 1 de ellas es hipertensa.

Porcentaje.

Es la proporción multiplicada por 100: $P = \frac{a}{n} \cdot 100\% \; P = \frac{30}{150} \cdot 100\% = 20\%$

En el ejemplo anterior:

El 20 % de las mujeres pertenecientes a ese consultorio son hipertensas o de cada 100 mujeres pertenecientes a ese consultorio 20 son hipertensas.

Tasa.

Se define como la relación por cociente entre el número de veces en que sucede un determinado fenómeno y la población expuesta al riesgo de ocurrencia de ese fenómeno:

$$T = \frac{a}{a+b} \cdot 10^m$$

Donde: a es la frecuencia con que ha ocurrido un determinado fenómeno durante un período determinado, entonces, m: es un entero positivo que se

determina en cada caso particular de tasa a calcular.

a + b será igual al total de personas expuestas al riesgo durante el mismo lapso de tiempo.

Como el numerador está contenido en el denominador, se multiplica este cociente por una potencia de 10 que haga este valor mayor de la unidad para que sea más fácil su interpretación.

La tasa es una proporción multiplicada por una potencia conveniente de 10 y expresa la probabilidad de ocurrencia de un fenómeno en una población específica y en un período determinado.

Las tasas más comunes o importantes en el campo de la salud son las de mortalidad, morbilidad, natalidad y letalidad. Por ejemplo: La tasa cruda de natalidad se define como:

$$TN = \frac{\text{Total de nacidos vivos}}{\text{Población total}} \cdot 1000$$

Si este resultado fuera 15.2 (TN = 15.2) se interpretaría que por cada 1000 habitantes nacen vivos 15.2 niños y del mismo modo se procedería como cualquier otro.

8.4.3.4. Medidas para conocer la forma de la distribución de los datos: Los momentos.

Para datos cuantitativos se han visto diferentes medidas que ayudan a poder realizar una caracterización o descripción de la distribución del conjunto de datos con que se cuenta, medidas que permiten conocer el centro de la distribución, otras que sirven para ver cuán alejados están estos datos entre sí y medidas para saber la distribución de estos datos de manera porcentual, o sea, conocer los intervalos (sus cotas) donde se encuentra reunido un porcentaje determinado de los valores; pero contribuiría a tener una idea más completa de la información, si se contara con medidas que permitieran conocer la forma de la distribución de los datos.

Con vistas a definir estas medidas se hace necesario identificarse con lo que se conoce en Estadística como momentos.

Momentos.

El r-ésimo momento con respecto a un origen c es definido de la forma siguiente:

$$M_R = \frac{\sum_{i=1}^{n}(x_i - c)^R}{n}$$

para diferentes valores de c y R aparecerán distintas medidas estadísticas, por ejemplo, si c = 0 y R = 1 (primer momento con respecto a cero) podrá reconocer a la media aritmética; si c = x y R = 0 (momento cero con respecto a la media) el resultado es igual a cero (revisar propiedades de la media aritmética) por el contrario si c = x y R = 2 se obtendrá la varianza poblacional.

Los momentos más comunes son los que se calculan con respecto a cero y a la media aritmética, así:

$$M_{aR} = \frac{M_R}{S^R} = \frac{M_R}{\sqrt{M_2^R}} = \frac{M_R}{\sqrt{M_2^R}}$$

debe recordarse que:

$S = \sqrt{M_2}$, es la desviación estándar poblacional.

Será fácil comprobar que Ma_1 es igual a 0 y Ma_2 es igual a $M_{a1} = \frac{M_1}{\sqrt{M_2}} = \frac{0}{\sqrt{M_2}} = 0$ 1 ya que $M_1 = 0$ y $M_2 = s^2$

$$M_{\alpha_2} - M_2 / \sqrt{M_2^2} - S^2 / \sqrt{(S^2)^2} - 1$$

$$M_R = \frac{\sum_{i=1}^{m} f_i (x_i - x)^R}{n}$$

Cuando los momentos se calculan para datos agrupados usualmente se le realizan correcciones (correcciones de Sheppard) a los momentos 2 y 4, así para los momentos con respecto a la media.

Donde: m: número de clases.

n: es el número total de observaciones

Así, M_2 corregido = $M_2 - (1/12)a^2$

M_4 corregido = $M_4 - (1/2)a^2 M_2 + (7/240)a^4$

Donde a es la amplitud de la clase

8.4.3.5 La asimetría y la curtosis, y sus diversas formas de expresión numérica.

La asimetría.

La forma de la distribución de datos puede caracterizarse de dos maneras: por su desviación de la simetría, lo que se conoce como asimetría, o por su grado de agudeza, lo que recibe el nombre de curtosis (o kurtosis).

Cuando la distribución de los datos tiene una cola más larga a la derecha del máximo central que a la izquierda (existen menos datos en el lado que corresponde a los valores más altos), la distribución se define como asimétrica a la derecha o que tiene asimetría positiva. Si, por el contrario, tiene una cola más larga hacia la izquierda se define como asimétrica a la izquierda, o que tiene asimetría negativa.

Una medida de la asimetría viene dada por el coeficiente momento de asimetría (momento respecto a la media aritmética). (Aquí s se refiere al valor poblacional, es decir, se divide entre n y no entre n – 1).

$$Ma_3 = \frac{M_3}{s^3} = \frac{M_3}{(\sqrt{M_2})^3} = Ma_3 = \frac{M_3}{\sqrt{M_2^3}}$$

Otra medida de asimetría es la que se define como Ma_3^2.

En el ejemplo de las esrtructuras de las embarazadas:

$$Ma_3 = \frac{M_3}{(\sqrt{M_2})^3} = \frac{476.8716}{\sqrt{(93.7701)^3}} = 0.5088$$

Si Ma_3 es positiva, la distribución será asimétrica positiva o asimétrica a la derecha, y si es Ma_3 negativa, entonces la distribución será asimétrica a la izquierda.

Existen otras medidas de asimetría más sencillas como son los coeficientes de asimetría primero y segundo de Pearson definidos respectivamente como: $CA_1 = (\bar{X} - \text{moda})/s$ y $CA_2 = 3(\bar{X} - \text{mediana})/s$.

También se definen otros coeficientes de asimetría, tales como:

El coeficiente cuartílico de asimetría, $CCA = (Q_3 - 2\ Q_2 + Q_1)/(Q_3 - Q_1)$, y el coeficiente de asimetría percentílico, $AP = (P_{90} - 2\ P_{50} + P_{10})/(P_{90} - P_{10})$.

Aplicado al ejemplo de las estaturas de las embarazadas sería:

$CA_1 = (166.8 - 159)/9.79 = 0.8$

$CA_2 = 3\ (166.8 - 165.5)/9.79 = 0.4$

CCA = (173.5 – 2 . 165.5 + 160)/(173.5 – 160) = 2.5 /13. 5 = 0.19

CAP = (181– 2 . 165.5 + 155)/(181 – 155) = 5/26 = 0.19

Por supuesto, mientras mayor sea el valor absoluto del coeficiente, mayor será el grado de asimetría.

Curtosis.

Como se mencionó anteriormente la curtosis (o kurtosis) mide el grado de esbeltez o agudeza de la distribución.

Si la distribución presenta un pico relativamente elevado, se dice que es leptocúrtica; si más bien es achatada, se dice que es platicúrtica y si es un término medio, se le denomina mesocúrtica. Un

ejemplo de esta última es la distribución normal, de la cual se comentará posteriormente y la que muchas veces sirve como patrón de comparación.

Una medida de la curtosis es el coeficiente momento de curtosis (momento con respecto a la media aritmética).

En la curva normal este valor es 3, por lo que también en ocasiones la curtosis se define como $Ma_4 - 3$; entonces, si este valor es positivo, la distribución es leptocúrtica y platicúrtica, si es negativo.

$$Ma_4 = M_4/M_2^2$$

$$Ma_4 = \frac{28748.6240}{(95.7701)^2}$$

$$= 3.1344 ,$$

En el ejemplo:

pudiera considerarse esta distribución como mesocúrtica al ser este valor bastante cercano a 3.

Utilizando los cuartiles y los percentiles se define también otra medida de la curtosis, el coeficiente de curtosis percentílico.

$K = Q/(P_{90} - P_{10})$, donde $Q = (Q_3 - Q_1)/2$

Este coeficiente en la distribución normal tiene un valor de 0.263.

Siguiendo con el mismo ejemplo:

$K = (173.5 - 160)/[2\cdot(181 - 155)\ ] = (13.5)/(2\cdot 26) = 0.260$; cercano al valor de la normal.

8.4.3.6. Sensibilidad y especificidad.

En la mayoría de las ocasiones en que se investiga o utiliza una nueva prueba para el diagnóstico de una enfermedad se desea medir su eficacia. Para esto se han llegado a definir diferentes términos, entre ellos la sensibilidad y la especificidad de la prueba en estudio.

Suponga que cierta prueba A es usada para el diagnóstico de la enfermedad B, para lo que se analizó un conjunto de 5 000 personas, las cuales presentaron el resultado siguiente:

De las 4 000 que no presentaron la enfermedad, 3 600 fueron bien clasificadas y de las 1 000 que sí la presentaban, 950 fueron bien clasificadas. Si se distribuyen estos resultados en una tabla de doble entrada:

	Resultados		
Personas	**Positivos**	**Negativos**	**Total**
Enfermos	950	50	1 000
Sanos	400	3 600	4 000
Total	1 350	3 650	5 000

La sensibilidad se define como la proporción de enfermos que son bien clasificados, es decir, que resultan positivos (también como el cociente de **verdaderos positivos** y la suma de **verdaderos positivos** más **falsos negativos**). La sensibilidad medirá la certeza de la prueba para detectar la enfermedad.

Aquí:

Sensibilidad = 9500/1000

= 0.95

La especificidad se define como la proporción de sanos bien clasificados, es decir que resultan negativos (también se puede definir como el cociente de dividir los **verdaderos negativos** y la suma de los **verdaderos negativos** y **falsos positivos**). La especificidad medirá la capacidad de la prueba para detectar la ausencia de patología.

Aquí:

Especificidad =3 600/4 000

= 0.90

Pero para saber interpretar correctamente los resultados de la prueba se hace necesario conocer también los llamados valores predictivos que son, la proporción de positivos que tienen la enfermedad.

950/1 350 = 0.704

y la proporción de negativos que no tienen la enfermedad

3 600/3 650 = 0.986

En el caso de que se estuviera en presencia de una enfermedad con una frecuencia baja pudiera suceder que la sensibilidad y la especificidad fueran altas y, sin embargo, no sucediera así con los valores predictivos, por lo que siempre es conveniente, para una mejor interpretación, calcular también estos valores.

Usando el concepto de probabilidad (el que se estudiará en el capitulo 9) , la sensibilidad y la especificidad pudieran definirse como:

Sensibilidad = P(Prueba positiva/enfermo)

Especificidad = P(Prueba negativa/sano)

ya que son probabilidades condicionales.

8.4.3.7 Riesgo relativo y razón de productos cruzados.

Existen otras proporciones a las ya vistas que son usadas con el objetivo de medir el grado de asociación entre un factor de riesgo y una enfermedad presente en ciertas poblaciones. Uno de estos casos es la que se denomina riesgo relativo, para los estudios prospectivos (estudios de cohortes), que son los que se inician con la identificación de individuos con y sin el factor que se va a investigar. Estos factores se determinan sin saber cuáles individuos padecen o padecerán la enfermedad.

Se define riesgo como la probabilidad de desarrollar una enfermedad en un período determinado, esto será el resultado de dividir el número de sujetos que llegan a desarrollar la enfermedad entre el total de sujetos que podían desarrollar la enfermedad o estaban expuestos a ella al inicio del período.

El riesgo relativo tendrá como fórmula: $RR = p_1/p_2$

interpretándose como una razon entre, la probabilidad (frecuencia) de desarrollar una enfermedad en presencia del factor de riesgo (p_1) respecto a la probabilidad (frecuencia) de desarrollar la enfermedad

en ausencia del factor de riesgo (p_2).

Un riesgo relativo de valor 1 significa que la presencia del factor de riesgo no aumenta el peligro de padecer la enfermedad y un riesgo relativo de valor x (>1) significaría que, como término medio, los sujetos que presentan este factor tienen un peligro de padecer la enfermedad x veces mayor que los que no lo presentan.

Si los datos estuvieran colocados en una tabla de doble entrada

	Factor de riesgo	**Sin factor de riesgo**
Enfermos	a	c
No enfermos	b	D

$p_1 = a/(a + b)$ y $p_2 = c/(c + d)$

Suponga que se realiza un estudio sobre el desarrollo de una cierta enfermedad y su posible relación con el hábito de fumar. Los resultados expresados en una tabla de doble entrada son los que siguen:

	Hábito de fumar	
Estado de salud	Fumador	No Fumador
Enfermos	150	90
No enfermos	1 850	2 910
Total	2 000	3 000

En el ejemplo:

RR = (150/2 000)/(90/3 000) = 0.075/0.030 = 2.5

Esto significa que el riesgo de contraer la enfermedad es 2.5 veces más elevado para los que fuman con respecto a los que no lo hacen.

Cuando los estudios que se realizan son retrospectivos (o de caso–control), es decir, estudios que se inician con la identificación de los individuos que tienen la enfermedad –casos– y los que no la tienen –controles–. Aquí los casos y los controles se identificaron sin saber si estuvieron expuestos o no a los factores que serán investigados. En estos casos no es posible calcular el riesgo relativo pero si lo que se conoce como la razón de los productos cruzados (odds ratio), la que también es posible calcular en los estudios prospectivos.

El **odds** o ventaja del éxito de un suceso se define como el cociente p/q donde, p es la proporción en que ocurre un suceso y q = 1–p (proporción en que no ocurre).

Cuando se consideran dos sucesos y las proporciones de ocurrencia de ambos son p_1 y p_2 se define la razón de productos cruzados (odds ratio) como el cociente:

$OR = (p_1 / q_1) / (p_2 / q_2) = (p_1 q_2 / p_2 q_1)$

también se puede definir como:

OR = ad/bc

Así en el ejemplo,

OR = (150 x 2910)/(90 x 1850) = 2.62

El odds del primer grupo es 2.62 veces mayor que el segundo grupo.

Cuando el odds es mayor que 1, indica una mayor frecuencia de que ocurra el suceso en el primer grupo que en el segundo; y cuando es menor que 1, indicará lo contrario.

En enfermedades poco frecuentes, la razón de productos cruzados se aproxima al riesgo relativo; de hecho, la primera puede ser estimada sin conocer realmente las incidencias de la enfermedad en las poblaciones.

8.4.3.8. Asociación entre dos variables.

Cuando se trabaja con más de una variable, en muchas ocasiones interesa la posible relación entre ellas, por lo que se busca construir medidas que puedan describir esta relación. A continuación se analizarán casas particulares.

8.4.3.8.1. Variables cuantitativas y los coeficientes de correlación.

Suponiendo que se cuenta con dos variables cuantitativas la relación entre ambas pudiera estar descrita por una línea recta, de manera tal que si una de las variables dependiera de la otra (la primera sería designada por **y** o dependiente y la segunda por **x** o independiente) se pudiera calcular esta primera variable en función de la segunda a través de una línea recta (recta de regresión lineal); y una forma de medir el grado de relación lineal entre dichas variables es mediante el cociente :

r^2 = Variación explicada / Variación total

partiendo de que la variación de la variable **y** puede ser explicada mediante la recta de regresión lineal, r^2 mediría el grado en que esta recta "cubre" la variación total, si el cociente es 0, la recta no explica la

variación de **y**; si es 1, la variabilidad de **y** es totalmente explicada por la recta, de manera general será un valor positivo entre 0 y 1 y este cociente es llamado coeficiente de determinación lineal, es usual encontrarlo multiplicado por 100; su raíz cuadrada, medida más conocida y utilizada, se le denomina como coeficiente de correlación lineal, simbolizado por r y el que será:

- Positivo, si al aumentar los valores de **x** también aumentan los de **y** y viceversa.
- Negativo, si al aumentar los valores de **x** disminuyen los de **y** y viceversa.

Trabajando con la fórmula correspondiente a r^2 se puede llegar a:

$$r^2 = \frac{\left[\sum (x_i - \bar{x})(y_i)\right]^2}{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}$$

Este cociente puede ser interpretado como el cociente de la covarianza al cuadrado de las variables **x** e **y** entre el producto de las varianzas de **x** y de **y**. La covarianza es una medida de la variación conjunta de las variables **x** e **y**; en el caso particular de que x e y son la misma variable, la covarianza se convierte en varianza.

Por ejemplo, se quiere conocer si la estatura de los padres está relacionada con la de los hijos (un problema clásico) para lo que en cierta

población se estudian los casos de 16 padres y sus primogénitos varones. Los valores en centímetros son:

Padre (X)	**Hijo (Y)**	**Padre (X)**	**Hijo (Y)**
1.70	1.68	1.95	1.84
1.84	1.90	1.87	1.89
1.63	1.70	1.69	1.64
1.74	1.74	1.83	1.79
1.87	1.93	1.94	1.98
1.85	1.80	1.71	1.75
1.79	1.83	1.77	1.78
1.97	2.04	1.81	1.83

Se tendrá que:

$$r^2 = \frac{(0.1432)^2}{(0.1480)(0.1786)} = 0.7758$$

También $r^2 =$ 77.58 %, es decir casi el 78 % de la variación de la variable **y** estaría explicada por las variaciones de la variable **x**.

Ya que r = 0.88, valor positivo, cuando aumentan los valores de la variable **x**, aumentan los de la variable **y** y viceversa y cuando disminuyen los valores de la variable **x**, disminuyen los de la variable **y** y viceversa.

Correlación por rangos

Existen ocasiones en que no resulta conveniente usar los verdaderos valores de las variables o estas están medidas en una escala que no permite usar el coeficiente de correlación lineal; por ello se define el coeficiente de correlación por rangos de Spearman, para estos casos, si hay interés de encontrar una medida de la relación entre dos variables.

Aquí los valores de cada variable son sustituidos por el rango o lugar que ocupan dentro de la serie de datos, por ejemplo, si se cuenta con los valores 9; 3; 5; 7; 12; 5, estos serán sustituidos por 5; 1; 2.5; 4; 6; 2.5,

(como el 5 se repite y ocupa los lugares 2 y 3, estos valores se promedian)

La fórmula del coeficiente de correlación por rangos es:

$$r = 1 - \frac{6 \cdot \sum_{j=1}^{n} D_j^2}{n \cdot (n^2 - 1)}$$

Donde:

D_j: diferencias de rangos de x_i e y_i.

n: es el número de pares de valores de x e y.

De existir relación entre las variables, los D_i deben ser 0 o muy cercanos a él y por lo tanto r muy cercano a 1.

Existe interés en conocer si los resultados que se obtienen para medir el colesterol en sangre a través de un método A está relacionado con los que se obtienen mediante otro método indirecto, el B, para ello se utilizó una muestra de 15 personas a las que se les determinó el nivel de colesterol por ambos métodos.

Los valores originales ya fueron sustituidos por la posición (rango) que ocupan dentro de cada serie de datos.

A	B	D_1	A	B	D_1	A	B	D_1
4	4	0	12	13	-1	13	12	1
6	5	1	8	7	1	14	14	0
1	2	-1	8	8	0	10	10	0
2.5	1	1.5	15	15	0			
11	10	1	5	6	-1			
8	10	-2	2.5	3	-0.5			

Sustituyendo en la fórmula de spearman se tiene que:

$$r = 1 - 6 \cdot \left[\sum_{j=1}^{15} D_j^2 \Big/ \left[15 \cdot (15^2 - 1) \right] \right]$$
$$= 1 - 6 \cdot \left[(0^2 + 1^2 + \cdots + 0^2 + 0^2) / (15 \cdot 224) \right] = 1 - 6 \cdot (13.5/3360)$$
$$= 1 - 0.024 = 0.976.$$

Como el valor del coeficiente de correlación de sperarman es cercano a 1, esto significa que ambas formas de medición estan relacionadas.

8.4.3.8.2. Variables cualitativas y las medidas de asociación dependientes de la variable χ^2.

Cuando se analizan datos de dos variables cualitativas (por supuesto que se pueden considerar mas de 2 variables, 3, 4, etc) es habitual, para el análisis de su interrelación, por ejemplo, contruir una **tabla de contingencia (TC)** que no es más que una **estructura rectangular (o arreglo compuesto) de f filas y c columnas**, y por tanto con **f·c casillas o celdas**, (donde se clasifica información proveniente de un conjunto, que posee (en general) n datos), significando **f y c el total de valores posibles (usualmente categorias o clases de una escala)** de cada una de las 2 variables, respecto a las cuales se clasifican los datos observados.

Posteriormente en el capitulo 12, dedicado al estudio de los métodos no parametricos, se retomara este tema, complementandose el analisis de las mismas mediante el estudio de las pruebas de hipotesis asociadas a estas tablas.

Consideremos ahora para ilustrar lo expresado, y ademas introducir otros aspectos relacionados con las TC, la situacion siguiente:

Se conoce que el hábito de fumar está relacionado con los trastornos respitatorios. Para corroborarlo en una cierta población se toma una muestra de 180 sujetos los cuales son clasificados en si: no fuman, fuman de modo leve, moderado o severo; y si tienen trastornos respiratorios o no.

Estos datos pueden disponerse en una TC del modo siguiente:

No	20	100	120
Leve	15	15	30
Moderado	12	8	20
Severo	8	2	10
Total	55	125	180

Como se observa, es una TC de 8 casillas, compuesta de 4(= f) filas y 2(= c) columnas, con un gran total de180(= n) datos clasificados.

Los valores como 12 y 100 se les denomina, **frecuencia (absoluta) o valor observado,** se simbolizan por $\mathbf{O_{ij}}$ (aunque también por $\mathbf{n_{ij}}$) , por ejemplo,12 es $O_{31}(n_{31})$ y 100 es $O_{12}(n_{12})$, significando el 12, que existen 12 personas entre las 180 que padecen de trastornos respiratorios con habito de fumar moderado.

Aquellos como, **120, 30, 20, 10 y 55, 125** se les conoce como **totales o frecuencias marginales por filas y columnas** respectivamente, estos totales por filas y columnas se acostumbran a simbolizar por $T_{i\cdot}$ y $T_{\cdot j}$ (aunque también mediante la notación $n_{i\cdot}$ y $n_{\cdot j}$) respectivamente. Así $T_{1\cdot}=120, T_{2\cdot}=30, \cdots, T_{\cdot 1}=55, T_{\cdot 2}=125$. Para estos totales debe cumplirse en general que:

$$T_{i\cdot} = O_{i1} + O_{i2} + O_{i3} + \cdots + O_{ic} = \sum_{J=1}^{c} O_{ij}\text{, para i desde 1 hasta f}$$

$$T_{\cdot j} = O_{1j} + O_{2j} + O_{3j} + \cdots + O_{fj} = \sum_{i=1}^{f} O_{ij}\text{, para j desde 1 hasta c.}$$

$$T_{1\cdot} = 120 = 20(O_{11}) + 100(O_{12}) = \sum_{J=1}^{2} O_{1j}\text{;}$$

$$T_{\cdot 2} = 125 = 100(O_{12}) + 15(O_{22}) + 8(O_{32}) + 2(O_{42}) = \sum_{i=1}^{4} O_{i2}$$

De este modo,

El total general de la tabla es n. Se ha de cumplir siempre que:

$$n = \sum_{i=1}^{f}\sum_{j=1}^{c} O_{ij} = \sum_{j=1}^{c} T_{\cdot j} = \sum_{i=1}^{f} T_{i\cdot}$$

O sea que, el total general de TC, debe tanto ser igual, a la suma de todos los valores observados (O_{ij}), asi como ser la suma de todos los totales marginales por columna o por filas.

Si las 2 variables no están asociadas estos valores O_{ij} tienen un comportamiento que puede ser analizado mediante el valor **$(T_{i\,.} \cdot T_{.\,j})/n$**, es decir, el producto de los totales marginales correspondientes a la fila y la columna que se interceptan en el valor observado O_{ij}, dividido por n. Es usual en este cuerpo teorico denotar al valor $(T_{i\,.} \cdot T_{.\,j})/n$ por medio de **e_{ij}**, siendo conocidos por el nombre de **frecuencia o valor esperado**.

Teóricamente bajo el supuesto hipotetico hecho al inicio del parrafo anterior las frecuencias observadas y esperadas deben ser valores muy proximos entre si y en ocasiones hasta coincidir. Por ello en la práctica, desviaciones grandes en uno u otro sentido suelen tomarse como un indice de que algo falla en relacion con el supuesto hecho.

En el ejemplo de la posible relación entre el hábito de fumar y los trastornos respiratorios, al calcular las frecuencias esperadas, sus valores serán:

$e_{11} = 36.7$ $e_{12} = 83.3$ $e_{31} = 6.1$ $e_{32} = 13.9$

$e_{21} = 9.2$ $e_{22} = 20.8$ $e_{41} = 3.1$ $e_{42} = 6.9$

(Recordemos que $e_{11} = \frac{T_{1\cdot} \cdot T_{\cdot 1}}{n} = \frac{120.55}{180} = 36.7$ y que de modo analogo se calculan las demás.)

Una medida de la asociación entre las dos variables debe estar basada en la diferencia existente entre estas frecuencias esperadas y las realmente observadas, mientras mayor sea esta medida mayor evidencia habrá de que existe esa asociación.

Entre las medidas de asociación suele considerarse a la figura estadística de carácter numérico siguiente:

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\left[(O_{ij} - e_{ij})^2 / e_{ij}\right],$$

como una de las de más frecuente uso, denominada como **valor chi-cuadrado.**

Donde: i : varia desde1hasta f filas

j : varia desde1hasta c columnas o_{ij}: frecuencias observada correspondiente a la casilla de indices

i,j.e_{ij}: frecuencias esperadas correspondiente a identica casila.

Esta fórmula es equivalente a: $\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\left[o_{ij}^2/e_{ij}\right] - n$, donde n es la frecuencia total o total de datos observados.

Esta medida de asociación χ^2 será siempre un valor positivo; con valor mínimo 0, cuando no existe asociación entre ambas variables.

Mas adelante en el capitulo 9, dedicado al estudio de las nociones básicas de la teoría de las probabilidades y la inferencia estadistica, se vera que, este valor también es conocido como **estadigrafo chi-cuadrado.**

Utilizando los datos del ejemplo desarrollado, el valor chi (o ji) cuadrado es:

$\chi^2 = 10.9 + 24.5 + 23.6 + 20.6 + 120.0 + 10.8 + 4.6 + 0.6 - 180 = 35.6$

Como se puede apreciar, es un valor al parecer (a grosso modo) **bastante alejado de 0**, por tanto, un indice de que ambas variables de la tabla, **habito de fumar y trastornos respiratorios** pueden estar en algún grado asociadas. Se dijo **a grosso modo**, ya que a este nivel todavia no se puede establecer un criterio que permita justificar que en realidad el valor χ^2 observado (a veces denotado por χ^2_{obs}) esta **bastante alejado de 0**, es decir, decidir si lo destacado en negrita es cierto o no.

En el caso en que abunden valores observados (Oij) en la tabla con valor menor que 5 se propone utilizar la llamada corrección de Yates.

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\left[\left(|o_{ij} - e_{ij}| - 0.5\right)^2/e_{ij}\right]$$

cuya diferencia con el valor anterior consiste en que, al valor absoluto de la diferencia entre la frecuencia observada y esperada general se le resta 0.5, quedando los restantes procederes iguales.

Como se dijo al inicio de este epigrafe las TC se pueden usar en relación con una sola variable, como, cuando lo que se quiere conocer es, si la variable bajo estudio se comporta semejante a una distribución conocida o teórica. Luego, de acuerdo con las clases o categorías de la escala que se establezca usar, se calculan según la distribución teórica, las frecuencias esperadas tomando en cuenta el tamaño de muestra utilizado (en este aspecto radica la diferencia básica con relación con el esquema anterior).

Posteriormente estas se comparan con las frecuencias observadas, calculandose el valor chi-cuadrado por medio de la expresión siguiente:

$$\chi^2 = \sum_{i=1}^{m}\left[(o_i - e_i)^2/e_i\right]$$

donde m es el número de clases o categorías de la escala en uso.

Otra medida de asociación, es el coeficiente de contingencia que se define como: $C = [c^2 / (c^2 + n)]^{1/2}$.

C es un valor menor que 1 y mientras mayor sea, mayor será el grado de asociación.

En el ejemplo planteado:

$C = [35.6 / (35.6 + 180)]^{1/2} = 0.406$

Cuando el número de filas y columnas coinciden, se define el coeficiente de correlación de atributos como:

$r = \{ c^2 / [n (k - 1)] \}^{1/2}$

donde k es el número común de filas o columnas de la tabla.

Este coeficiente toma siempre un valor entre 0 y 1; para el caso en que k = 2 se le llama coeficiente de correlación tetracórica.

8.4.4. Presentación de la información.

No basta resumir la información, es necesario presentarla y hacerlo de tal modo que sea claramente entendible y sin necesidad de textos acompañantes aclaratorios para su interpretación, la presentación de datos puede llevarse a cabo a través de:

- Tablas o cuadros estadísticos.
- Gráficos.

8.4.4.1. Tablas estadísticas

Como el nombre lo indica, la información resumida se presenta en forma tabular. El cuadro estadístico no debe estar recargado ya que, lejos de aclarar, confunde. Cuando se necesita dar mucha información es preferible construir varias tablas.

Básicamente ambos métodos de presentación constan de las mismas partes: título, cuadro o gráfico contentivo de los datos y las notas explicativas o calce.

Título.

Debe ser completo y conciso. Un título completo significa que a través de él debe conocerse el contenido de la tabla, para ello será importante que el mismo responda a las preguntas qué, cómo, cuándo y dónde; es decir:

- Qué se estudia, de qué trata la tabla, cuál es su contenido.
- Cómo, cuáles son las características que se toman en cuenta para clasificar lo que se estudia.
- Cuándo, a qué fecha se refiere el estudio.
- Dónde, de qué lugar se toman los datos.

Que el título sea conciso significa que se debe ser lo más breve posible, sin que ello conlleve a perder o sacrificar la característica de la claridad.

Cuadro.

Estará conformado por un grupo de filas y columnas en cuyas intersecciones, llamadas celdas o casillas vendrán dispuestos los datos.

En la primera fila se colocan los encabezamientos de las columnas que deben ser breves y explícitos. Generalmente, la última fila se reserva para los totales.

La primera columna se llama columna matriz y contiene los nombres de las diferentes clases establecidas para la variable en estudio; a las demás columnas se les señala como columnas auxiliares.

La tabla puede clasificar más de una variable. En el caso de dos, las categorías de una irían por las filas y las de la otra irían por las columnas. A medida que se incorporen 3 o más variables se complicaría la tabla, lo cual pudiera no ser conveniente.

Notas explicativas o calce.

Aquí, generalmente, se inscribe la fuente de la información reflejada y cualquier aclaración que se desee hacer acerca del contenido de la tabla.

De modo convencional, las notas explicativas pueden colocarse en la parte superior cuando afectan todo el contenido y en la inferior, si solo se van a referir a determinadas casillas, lo que debe ser indicado por un número o una letra entre paréntesis.

El calce solo se emplea a veces; pero siempre que los trabajos no sean originales la fuente debe incluirse, para indicar de dónde se extrajo la información y para que si otra persona lo desea pueda recurrir al trabajo original.

Tabla 1

1 Relación de entidades que prestan servicios estomatológicos. cuba,1998.

	2	3
	Entidad	**Número**
	Clinicas Estomatológicas	166
	Policlínicos	275
	Centros escolares	182
5		
	Centros de trabajo	107
	Hospitales	127
	Instituciones sociales	20

	Otras instituciones	437
	Total	1 314

6

7 Fuente: La Salud Pública en Cuba. Hechos y Cifras. Dirección Nacional de Estadística, MINSAP, Cuba,1999.

Donde los números anteriores significan:

1-Título

2-Columna matriz

3-Columna auxiliar

4-Fila de encabezamiento

5-El cuadro propiamente

6-Fila de totales

7-Notas explicativas

En este caso, el título es completo, pues a través de él sabemos **qué** y **cómo** se estudia (los servicios estomatológicos mediante las entidades que lo ofertan), así como **dónde** y **cuándo** (en 1998 en Cuba). Además es conciso, porque de forma breve indica lo anterior. Si el título fuera "Relación de las diferentes entidades que de una forma u otra brindan algún tipo de servicio estomatológico en Cuba durante 1998", estaría completo pero muy extenso; si faltara Cuba o 1998, no se sabría ni de dónde son los datos ni a qué fecha se refieren. Si dijera "Relación de Entidades.CUBA, 1999", no se sabría qué se estudia o si se planteara "Servicios estomatológicos. Cuba, 1998", no se indicaría que lo que interesa aquí de estos servicios son las entidades que los ofertan porque faltaría el cómo se estudian los servicios estomatológicos.

En el cuadro se reflejan los encabezamientos de las columnas entidad y número, la columna matriz está compuesta por las clases o categorías consideradas para las entidades (clínicas estomatológicas, policlínicos, etc.). Esta tabla solo tiene una columna auxiliar que contiene cantidades correspondientes a las categorías que se establecieron en las entidades, pudiera haberse adicionado otra columna auxiliar, si se hubieran considerado los porcentejes que aportaba cada clase. La última fila contiene el total de columna.

En las **notas explicativas** colocamos la **fuente** ya que los datos no son originales y es necesario indicar de dónde fueron extraídos los mismos.

Otro ejemplo:

Tabla 2

Mortalidad por enfermedades infecciosas y parasitarias según Edad.

Cuba, 1998 (Tasa por 100 000 habitantes)

Grupos etáreos	Tasa
0- 4	35.3
5-14	1.6
15-39	5.0
40-64	28.6
65 o más	449.5
Total (bruta)	54.5
Total (ajustada)	45.3

Fuente: La Salud Pública en Cuba Hechos y Cifras, Dirección Nacional de Estadística, MINSAP, Cuba, 1999.

Cómo leer la tabla.

La tabla se lee de arriba hacia abajo y su lectura es fácil si se tiene en cuenta qué significa cada parte de la misma.

Bastará saber a qué encabezamiento de columna y no fila pertenece el dato; por ej. el 107, indicará que en 107 (número, frecuencia, cantidad; esto corresponde al encabezamiento de columna) centros de trabajo (clasificación de las entidades, esto es referido a la fila) se brindan servicios estomatológicos en Cuba en 1998 (esto lo aporta el título).

La lectura y comprensión de la tabla también exige que el lector extraiga información relevante de toda la que se aporta. En el caso que se analiza, podría ser que el tipo de entidad que más ofrece esta clase de servicio es el policlínico y después los centros escolares (no se toma en cuenta la categoría **otras instituciones**, que estará formada por varias sin saber cuáles).

También son válidas las recomendaciones establecidas por Wallis y Roberts en "Statistics: A New Approach", 1956 que relacionamos a continuación:

1. Leer cuidadosamente el título y las notas explicativas.

2. Averiguar las unidades de medida utilizadas.

3. Fijarse en el promedio o porcentaje general del grupo.

4. Relacionar el promedio general del grupo con cada una de las variables que se estudia.

5. Relacionar entre sí los promedios o porcentajes de las variables que se estudian.

6. Buscar irregularidades en los datos.

8.4.4.2. Gráfico.

El gráfico constituye otra forma de presentar una información ya resumida, tanto él como la tabla permitirán dar de forma rápida y clara, de un golpe de vista, una idea del comportamiento de los datos.

El gráfico resulta menos preciso que la tabla (fundamentalmente cuando se trata de variables cuantitativas), pero más fácil de entender. Se utiliza sobre todo para destacar la tendencia que sigue un fenómeno o una variable, o para resaltar una particularidad de los datos o alguna relación entre variables.

Como se planteó anteriormente el gráfico consta de 3 partes:

1. El título. Los requerimientos coinciden con los de la tabla.
2. El gráfico. Dependerá del tipo de variable con que se esté trabajando.Existe una gran diversidad de gráficos, los más comunes son, para las variables cualitativas y cuantitativas discretas los diferentes gráficos de barras y el de sector o pastel y para las variables cuantitativas continuas, el histograma y el polígono de frecuencia.
3. Las notas explicativas. Es válido lo que se planteó para las tablas. Aquí se incluirá, en los casos que precise la leyenda, la cual aparece con mayor frecuencia debajo o a la derecha del gráfico.

8.4.4.2.1 Tipos de gráficos.

Existen diversos tipos de gráficos, se estudiarán los más usados.

Gráficos de barras

Se utilizan para variables cualitativas y cuantitativas discreta y se representan a través de un sistema de coordenadas.

El más sencillo es el gráfico de barras simples, en este caso se estudia una sola variable, cuyas categorías se colocan en el eje de las abscisas y las frecuencias de cualquier tipo, en el eje de las ordenadas.

Este tipo de gráfico se usa también para representar series cronológicas de pocos datos.

Los valores de las frecuencias se disponen mediante barras o rectángulos separados, que deben deben ser de igual amplitud y la distancia entre ellos no mayor que el ancho de la barra ni menor que la mitad del mismo.

Ejemplo 1:

Dado el siguiente resumen estadístico:

Tabla 3

Casos notificados de hepatitis viral y varicela en las provincias centrales, Cuba, 1998.

Provincias	**Cantidadhepatitis viral**	**Cantidadvaricela**	**Total deCasos**
Villa Clara	1 814	1 133	2 947
Cienfuegos	456	526	982
Sancti Spíritus	1 368	1 309	2 677

Ciego de Avila	218	831	1 049
Camagüey	777	1 023	1 800
Total	4 633	4 822	9 455

Fuente: La Salud Pública en Cuba. Hechos y Cifras Dirección Nacional de Estadística, MINSAP, Cuba, 1999.

La representación de los datos correspondientes a los casos notificados de varicela en las provincias centrales se aprecia en la figura 1.1

Figura 8.1

Debe tomarse en cuenta que se logra una mejor apariencia y comprensión de un suceso médico dado mediante gráficos de barras, si las mismas se colocan de mayor a menor o viceversa, siempre que no se viole algún orden establecido, como por ejemplo, si la variable en estudio está referida a la época del año (primavera, verano, otoño e invierno), a los días de la semana, etc.; en el ejemplo expuesto si no es necesario que las provincias aparezcan en ese orden, sería mejor representarlo del modo siguiente:

Ejemplo 2:

Figura 8.2

También será importante que la escala tomada para la frecuencia (eje "y" de la figura) comience en cero y no se "parta o interrumpa" ya que de no ser así el gráfico puede dar una idea equivocada del verdadero comportamiento de la variable.

Ejemplo 3: supóngase que se tiene la siguiente distribución de frecuencias.

Tabla 4

Prevalencia de lepra. Provincias Pinar del Río, La Habana y Matanzas, 1998.

Provincia	**Cantidad**
Pinar del Río	40
La Habana	19
Matanzas	21

Fuente : La Salud Pública en Cuba. Hechos y Cifras. Dirección Nacional de Estadísticas, MINSAP, Cuba, 1999.

Una posible representación de esta información que se nos pudiera presentar es:

Figura 8.3

En este caso la visión que ofrece la figura 8.3, en un primer golpe de vista es como si la prevalencia de la lepra en Matanzas fuera 3 veces la de La Habana, cuando en realidad son prácticamente similares (21 contra 19 respectivamente); así mismo Pinar del Río parece tener una prevalencia de varias veces la de las otras

provincias, cuando realmente es de aproximadamente el doble, debe observarse que lo dicho es consecuencia de que la escala del eje y, comienza en 18 y no en 0, como ya se dijo.

Gráfico de barras múltiples

La situación problémica más frecuente donde se usa este tipo de gráfico es aquella donde interviene el uso de dos variables, pudiendo ser estas cualitativas, cuantitativas discretas o combinaciones de estas.

En este caso en un sistema de ejes coordenados, sobre el horizontal se colocan las categorías representativas de una de las variables, y sobre cada una de estas categorías se levantan tantas barras como categorías tenga la otra.

Ejemplo 4:

Tomando en cuenta los datos de la tabla 3 (ver ejemplo 1), se puede ilustrar el uso de este tipo de gráfico (barras múltiples) mediante la figura 8.4 siguiente:

Figura 8.4

Este gráfico se puede interpretar de varias formas, una de ellas, es comparar los casos de hepatitis viral o de varicela por provincia, también la relación o comportamiento de ambas enfermedades en cada provincia y entre estas.

Gráfico de barras proporcionales.

También se conoce como gráfico de barras compuestas. Al igual que el anterior se utiliza para representar más de una variable, aunque puede ser usado cuando hay una sola.

En este caso, en vez de usar varias barras por cada categoría de una de las variables, se construye una sola barra y respecto de esta se representan las categorías de la otra variable que interviene en el estudio de un fenómeno dado, en forma proporcional a la frecuencia observada en al tabla estadística, generalmente en forma porcentual.

Ejemplo 5: considérese para esta ilustración de nuevo los datos de la tabal 3. El gráfico de barras proporcionales correspondiente a este resumen de datos es:

Figura 8.5

Aquí se trabaja con valores porcentuales, por ejemplo, en la provincia de Villa Clara hay un total de 2 947 casos notificados de ambas enfermedades, 1 814 corresponden a la hepatitis viral, equivalente al 61.6 %, y 1 133 a varicela que representan el 38.4 %. Los valores porcentuales calculados, permitirán realizar una comparación de las enfermedades dentro de cada provincia y también si esta podría ser o no la misma relación entre las demás provincias.

Los datos anteriores pueden ilustrarse de otro modo como se ejemplifica a continuación.

Figura 8.6

De esta forma, la comparación en primera instancia se hace dentro de cada enfermedad.

Seleccionar entre este gráfico y el anterior depende de los intereses del investigador, así como seleccionar entre uno de este tipo y el de barras múltiples.

Gráfico de sector o pastel.

Este es un gráfico que se usa para la representación de una variable cualitativa o cuantitativa discreta, brinda la misma información que el de barras simples. Ambos se utilizan con fines comparativos.

Esta presentación emplean un círculo dividido en sectores, en el que el tamaño de cada sector se corresponde con el aporte de cada categoría de la variable.

El empleo de este tipo de gráfico se sustenta sobre el cálculo de la magnitud del ángulo correspondiente al sector que representan cada categoría. Para ello debe recordarse que todo circulo abarca un ángulo de 360^{o} y que el total de las frecuencias absolutas de cualquier resumen de datos constituye el 100 %, bastará, entonces, multiplicar el porcentaje correspondiente a cada característica por 3.6 para conocer cuántos grados tiene el sector que representará la misma.

La explicación de esto es que si:

360º ------- 100 %

$$x = \frac{360 \cdot 1}{100} = 3.6^{o}$$

x --------- 1 %

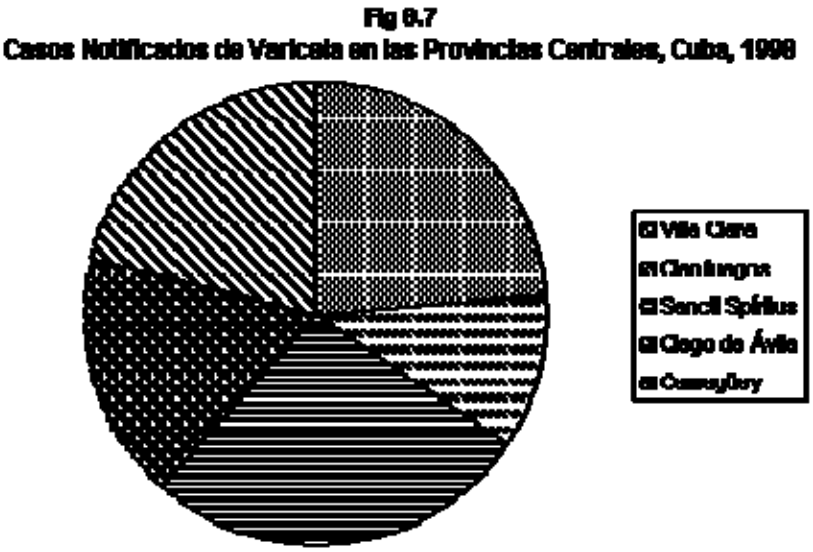
Cada 1 % representa 3.6º. En el ejemplo para los casos notificados de varicela se tendría.

Ejemplo 6: si se utilizaran los casos notificados de varicela de la tabal 3 y se transforman según se explicó anteriormente, se obtendrá la siguiente tabla.

Tabla 5

Provincia	Casos reportados de varicela	%	multiplicado por 3.6 (en º)
Villa Clara	1133	23.50	84.60
Cienfuegos	526	10.91	39.28
Sancti Spíritus	1309	27.15	97.44
Ciego de Avila	831	17.23	62.03
Camaguey	1023	21.21	76.36
Total	4822	100.00	~360º

A partir de estos datos ya es posible obtener la representación siguiente:



Los valores (porcentuales o de frecuencia) no tienen que estar incluidos pero hacen al gráfico más preciso y asi no habrá que recurrir a una tabla para conocer el valor exacto (esto es válido para los otros gráficos) defecto que en muchas ocasiones se les señala.

Gráfico aritmético simple.

Para representar a una variable cualitativa o cuantitativa discreta pudiera utilizarse un gráfico de puntos, mantener la forma verbal que consistiría en colocar en el eje de las abscisas las diferentes características de la variable mientras el de las ordenadas estaría destinado para las frecuencias y la representación correspondería simplemente a un punto. Sería por tanto absurdo unir estos puntos buscando una tendencia, ya que los valores intermedios en el eje de las abscisas no existen y unir los puntos daría la idea de una continuidad irreal. Pero cuando esta variable se refiere al tiempo, es decir, se quiere representar una serie cronológica, esta continuidad sí existe, y aunque la serie puede ser representada por un gráfico de barras, más comúnmente se hace por un gráfico de puntos que se unen entre sí por segmentos y se le llama **gráfico aritmético simple**, dado que en el eje de las ordenadas (frecuencias) se toma la escala aritmética. Si para la representación fuera necesario usar la escala logarítmica, se llamaría **gráfico semilogarítmico**.

Ejemplo 7: dado el siguiente resumen estadístico, el cual describe el comportamiento del total de nacidos vivos durante el período 1988-1998.

Tabla 6

Total de nacidos vivos por años, en 1988-1998, Cuba.

Año	Nacimientos	Año	Nacimientos
1988	187 911	1994	147265
1990	186 658	1996	140276
1992	157 349	1998	151 080

Fuente: Annual Health Statistics Report 1999, MINSAP, Cuba,1999.

A partir de estos datos se elabora el gráfico aritmético simple (Figura 8.8) como se muestra a continuación.

Figura 8.8

Se hace notar que en este tipo de gráfico es admisible el uso de interrupciones en los ejes, en este caso el eje horizontal o del tiempo. Pero también puede ser en ambos, como cuando solo se desea reflejar la tendencia del fenómeno.

Esto puede ilustrarse del modo siguiente.

Figura 8.9

También se puede construir el gráfico colocando el valor del tiempo en el intermedio de los intervalos como se muestra en la figura (Figura 8.10).

Figura 8.10

Se supone que la tendencia del fenómeno dentro de un intervalo de tiempo es la que aparece en el gráfico, situación que pudiera no ser real de conocerse valores intermedios, por ejemplo, en el año 1989 el total de nacidos vivos fue de 184 891, en 1995 de 147 170 y en 1997 de 152 681, esto significa que en el primer tramo de 1988 al 90 el total de nacidos baja y después sube, en el tramo de 1994 al 96 al principio el total se mantiene prácticamente estacionario antes de disminuir y en el correspondiente al 1996-98, del año 96 al 97 aumenta pero del 97 al 98 disminuye.

Histograma.

Cuando la variable que se está analizando es cuantitativa continua se necesita un gráfico que refleje esta característica, el más frecuentemente usado es el histograma (ver figura1.11) que consiste en un gráfico de barras unidas, es decir, sin separación estre ellas, donde el eje horizontal se indican de acuerdo con la variable y sus datos a representar, los intervalos o clases; cada uno de los cuales se corresponderá con una y solo una barra, reflejando la altura de cada barra un valor proporcional a la frecuencia alcanzada por la variable en dicha clase o intervalo.

Ejemplo 8: como ilustración de lo expresado, considérese el cuadro estadístico siguiente:

Tabla 7

Tasas de fertilidad acorde con la edad materna, 1980, Cuba

Edad de la madre	Tasa	Edad de la madre	Tasa
15–19	86.3	35–39	16.2
20–24	116.8	40–44	4.6
25–29	70.9	45–49	1.8

Fuente: Annual Health Statistics Report 1999, MINSAP, Cuba, 1999. Tasa por 1000 mujeres de esa edad.

Esta distribución admite la representación siguiente mediante un histograma.

Figura 8.11

Los rectángulos o barras tendrán una amplitud dependiente de la de los intervalos (luego a intervalos de diferentes amplitudes corresponderán barras que reflejen esta característica en el gráfico) y como siempre la altura de la barra será proporcional al valor de la frecuencia usada, en el ejemplo una tasa. Si las amplitudes son diferentes, la altura de los rectángulos se obtendrá dividiendo la frecuencia entre la amplitud de la clase. Este gráfico también puede ser usado cuando la variable es cuantitativa discreta y el recorrido es muy amplio, por ejemplo, con la frecuencia cardíaca.

Polígono de frecuencias.

También es muy utilizado para las variables cuantitativas continuas el polígono de frecuencias. Este gráfico es muy similar al aritmético simple, pero los valores que se colocan en el eje de las abscisas corresponden a los puntos medios o marcas de clase. Aunque hay autores que no lo plantean así, el polígono (a diferencia del gráfico aritmético simple) corta al eje de las abscisas para dar idea de área.

Para lograr esto hay que construir un intervalo imaginario anterior a la primera clase y uno posterior a la última, ambos con frecuencia 0

Ejemplo 9:

Para ilustrar este tipo de gráfico, considérense los datos de la última tabla. La distribución que allí aparece puede representarse así: Figura 8.12.

En este ejemplo se considerarán los intervalos (10, 15) y (50, 55) inexistentes con frecuencia 0.

Figura 8.12

El histograma es más utilizado que el polígono; este último se usa sobre todo cuando se quiere comparar dos o más series de datos en el mismo gráfico.

Gráfico de frecuencias acumuladas u ojiva

Es un gráfico para variables cuantitativas, similar al aritmético simple o al polígono de frecuencias, en el que se utilizan los puntos de un plano y los segmentos que los unen entre sí; la frecuencia que se usa es la frecuencia acumulada, por lo que permitirá responder a la pregunta de ¿cuántos casos hay por debajo de determinado valor?

Como ya se mencionó, en el eje de las ordenadas se colocan las frecuencias acumuladas y en el de las abscisas los intervalos de clase, pero el punto se coloca en la intersección del valor de la frecuencia acumulada y el límite superior del intervalo. También puede trabajarse con los límites reales y empezar con frecuencia 0, que correspondería al primer límite real inferior.

Ejemplo 10:

Suponga que en un consultorio médico se tienen los siguientes datos correspondientes a la distribución por edades de los hipertensos:

Tabla 8

Número de hipertenso por grupos etareos en un consultorio médico

		Frecuencia
Edades	**Hipertensos**	**Acumulada**
15–24	1	1
25–34	2	3
35–44	8	11
45–54	12	23
55–64	16	39
65–74	7	46
75–84	4	50

85–94	2	52
Total	52	

Así los datos referentes a la última columna se representan como:

Figura 8.13

Si se deseara conocer cuántos hipertensos en dicho consultorio son menores de 50 años, bastará levantar una línea en ese punto paralela al eje de las ordenadas y en el lugar de la intersección con la curva trazar otra paralela al eje de las abscisas. En el valor donde dicho segmento de recta corte al eje de las ordenadas indicará de forma aproximada el resultado 17.

Este gráfico es usual cuando se desea resumir series cronológicas, es decir, cuando lo que más interesa es el valor acumulado a través del tiempo.

En ocasiones se prefiere trabajar con el porcentaje en vez del valor absoluto, en otras, utilizar ambos a la vez; en el último caso se puede trazar hacia la derecha otro eje de ordenadas que contenga los porcentajes y entonces trabajar con cualquiera de los dos como se refleja en la figura siguiente:

Figura 8.14

Gráfico de tronco y hojas

Este guarda similitud con un histograma que tenga invertido los ejes (los valores en las ordenadas y las frecuencias en las abscisas), pero las barras van a estar representadas por los propios valores.

Ejemplo 11:

Suponga que se cuenta con las estaturas de 29 adolescentes y se quiere representarlas gráficamente. Los datos correspondientes son:

157,168,149,162,158,163,148,149,154,167,157,158,153,164,172,163,156,

169,154,150,154,161,170,149,154,153,158,164,168.

El gráfico de tronco y hojas quedaría representado como se indica a continuación (Figura 1.15):

Figura 8.15

14 8 9 9 9

15 0 3 3 4 4 4 4

15 6 7 7 8 8 8

16 1 2 3 3 4 4

16 7 8 8 9

17 0 2

Debe observarse que para la realización del gráfico anterior se consideraron los intervalos (145,149); (150,154); (155,159); (160,164); (165,169) y (170,174). Los primeros dígitos (14, 15, 16 y 17) representan el tronco y el último las hojas. Si el recorrido de la serie es grande, puede darse la situación en que los primeros dígitos tengan que repetirse dos veces utilizando 2 filas para colocar la primera vez, los valores más bajos y la segunda los valores más altos. En el caso contrario como en el ejemplo solo se utilizaría una fila. Si se tuvieran decimales, el tronco podrían ser los enteros y las hojas estarían reservadas para los decimales. Los valores se ordenan para tener una visión más clara de la distribución.

Este gráfico aventaja al histograma en que es muy fácil de construir y no se pierden los datos originales.

Gráfico de caja y bigote

Es propio para los datos cuantitativos y contribuye a dar una clara imagen del comportamiento de estos, trabaja con los valores mínimo y máximo y los cuartiles.

Ejemplo 12:

Usando la serie de datos del ejemplo anterior se puede construir el gráfico siguiente:

Figura 8.16

Como se observa, entre el primer y tercer cuartil se construye una caja (rectángulo) y a la altura del valor mediana se traza una línea horizontal que llega los bordes. Note además que los bordes inferior y superior de la caja coinciden con los valores del primer y tercer cuartil. Los llamdos bigotes son las líneas verticales que saliendo de los bordes superior e inferior se extienden hasta los valores mínimo y máximo del conjunto de datos ordenados.

Se puede apreciar que, la distribución de los datos es bastante uniforme ya que las separaciones entre los diferentes cuartiles es bastante parecida, aunque entre el primer y segundo cuartil hay una separación algo menor, lo que indica una mayor concentración.

Diagrama de dispersión

Se utiliza cuando se quiere estudiar o descubrir si existe relación o asociación entre dos variables cuantitativas.

Se trata sencillamente de un gráfico de puntos sobre un sistema de ejes coordenados donde una de las variables, la que se considera independiente, se coloca en el eje de las abscisas y la otra, la dependiente, en el eje de las ordenadas ver figura 8.17.

Ejemplo 13:

Suponga que se dispone de los valores de estatura y peso de 29 adolescentes, tal como se indica en el cuadro que se muestra a continuación.

Estatura (cm)	**Peso (kg)**	**Estatura (cm)**	**Peso (Kg)**	**Estatura (cm)**	**Peso (kg)**
148	50	154	60	163	65
149	48	156	61	164	67
149	53	157	57	164	68
149	54	157	62	167	69
150	61	158	60	168	67
153	57	158	64	168	69
153	52	158	63	169	70
154	55	161	64	170	68
154	58	162	66	172	69
154	51	163	65		

El diagrama de dispersión correspondiente a estos datos se representan en la figura siguiente:

Figura 8.17

Recomendaciones

En la mayoría de los gráficos se utiliza un sistema de coordenadas para su trazado, empleándose el eje de las abscisas (eje x) para los valores o categorías de la variable en estudio y el eje de las ordenadas (eje y) para las frecuencias (absolutas, relativas y acumuladas, etc.), aunque en

ocasiones esto se invierte en general por un problema de espacio.

La longitud de ambas escalas debe ser la misma o algo mayor la del eje de las abscisas, será conveniente probar con más de una escala para buscar cuál da una impresión más correcta y no distorsionada de lo que se estudia.

El eje de las ordenadas debe comenzar en cero a menos que los valores sean muy altos y con poca variación; siendo así el caso, no importaría interrumpir la escala, pues esto no afectaría la apreciación correcta del fenómeno. En el eje de las abscisas no debe constituir un problema la interrupción de la escala, ya que solo lo que se está produciendo es un traslado o corrimiento.

En los ejes de coordenadas deben aparecer los nombres de las variables y, si es necesario, las unidades de medida utilizadas.

En el caso de interesar no solo la tendencia del fenómeno sino los valores en sí, los datos originales deben aparecer reflejados en el gráfico; si no, será imprescindible que lo acompañe una tabla donde aparezcan registrados dichos valores. Por supuesto, esto no siempre será posible.

8.6 EJERCICIOS RESUELTOS.

1. A un grupo de 52 pacientes de un policlínico se le ordena un chequeo del que se recopilan los valores de ciertas variables como el sexo, la raza, la edad, el nivel de colesterol, el nivel de hemoglobina y otros. Tomando en cuenta que las valoraciones de hemoglobina y sexo fueron las siguientes:

12.8 F	14.5 F	13.2 F	13.4 M
12.6 M	13.6 F	12.8 F	11.4 M
10.9 M	9.9 M	11.4 M	13.4 F
12.7 M	14.5 F	12.3 F	12.3 F
11.8 M	13.0 M	14.5 M	11.8 F
11.9 F	13.4 M	11.0 M	11.0 F
11.1 F	12.3 F	12.3 F	12.6 F
11.2 F	11.2 M	13.3 M	10.6 F
11.3 M	13.2 M	11.4 M	13.3 F
11.4 M	12.3 F	13.3 F	13.2 F

14.0 M	12.8 M	13.2 F	12.7 F
14.2 M	13.8 M	10.1 F	12.3 F
13.8 F	12.4 M	13.6 M	12.7 F

Para la hemoglobina:

a) Construya una distribución de frecuencias con 7 clases, en la que se muestre, además de la frecuencia absoluta de cada clase, la relativa, la absoluta acumulada, la relativa acumulada y la marca de clase de cada intervalo.

b) Calcule la media aritmética, la mediana y la moda.

c) Calcule el rango, la varianza, la desviación estándar y el coeficiente de variación.

d) Calcule el coeficiente momento de asimetría y el de curtosis.

e) Calcule el cuartil 1, el decil 8 y el percentil 15.

Para el sexo:

f) Construya una distribución de frecuencias.

g) Calcule la razón de mujeres respecto a hombres.

h) Calcule el porcentaje de hombres que tienen la hemoglobina por debajo de 132 mmoles/l.

Respuesta:

a) Distribución de frecuencias

Rango = 145 – 99 = 46

al dividir este valor entre 7, que es aproximadamente el número de clases con que se quiere contar, dá más de 6, por lo que la amplitud de las clases se aproxima a 7 unidades.

La distribución de frecuencias correspondiente a la situación planteada se muestra a continuación en la tabla siguiente:

Intervalo Frecuencia Frecuencia FrecuenciaAbsoluta FrecuenciaRelativa Marca de

de clases	Absoluta	Relativa	Acumulada	Acumulada	clase
	FA	FR	FAA	FRA	Mc
99–106-	2	0.0385	2	0.0385	102.5
106–113-	7	0.1346	9	0.1731	109.5
113–120-	8	0.1538	17	0.3269	116.5
120–127-	9	0.1731	26	0.5	123.5
127–134-	14	0.2692	40	0.7692	130.5
134–141-	8	0.1538	48	0.923	137.5
141–148-	4	0.0769	52	0.9999	143.5
Total	**52**	**0.9999 @ 1**			

b) La media aritmética será la suma de todos los datos dividido entre 52

$$\bar{x} = \frac{12.8 + 14.5 + \ldots\ldots\ldots + 13.6 + 12.7}{52} = \frac{644}{52} \cong 12.4$$

La mediana será la media de los valores que ocupan, después de ordenar los datos, los lugares 26 y 27, ya que n es par.

Mediana = (126 + 127)/2 = 12.65 mmoles/l

La moda como valor más frecuente, es 123, el que se repite 6 veces.

c) El rango se define como la diferencia entre el valor mayor y el más pequeño, ya se había visto al construir la distribución de frecuencias que este valor es 46 mmoles/l.

Fórmula de la varianza:

$$s^2 = \frac{(12.8 - 12.5)^2 + \ldots\ldots\ldots + (12.7 - 12.5)^2}{52 - 1}$$

= 1.3002

La desviación estándar:

s = 1,14; que es la raíz cuadrada de la varianza

El coeficiente de variación, que es la desviación estándar entre la media aritmética por 100: (1,14/12,5) . 100 es:

CV=9,1 %

d) El momento de asimetría y la curtosis según la fórmula:

$$Ma_3 = \frac{M_3}{S^3}, donde\ M_3 = \frac{\sum(X_i - \overline{X})}{n}$$

Se tiene que: $M_3 = 0.3544\ y\ S = 1.2752,$ luego sustituyendo

$$Ma_3 = -0.3544/\left|\sqrt{(1.2752)^3}\right|$$

$$= -0.2461$$

La expresión de la curtosis es: $Ma_4 = M_4/M_2^2$

Se tiene que: $M_4 = 3.8556/(1.2752)^4 = 2.3719$

Donde $M_4 = \frac{\sum(Xi - \overline{X})^4}{n}\ y\ M_2 = S^2$

e) Para obtener el cuartil 1 (percentil 25) habrá que calcular el 25 % de 52, que es 13, un número entero, por lo que el cuartil 1 será el promedio de los valores que ocupan las posiciones 13 y 14, estos valores son 114 ambos, entonces:

C_1 = 114 mmoles/l

El decil 8 (o percentil 80) será el valor que ocupa la posición 42, ya que 80 % de 52 es 41.6; al ser decimal se debe aproximar a la posición siguiente:

D_8 = 134 U

Para encontrar el percentil 15, hay que calcular el 15 % de 52 que es 7.8, por lo que el percentil 15 será el valor que ocupa la posición 8, esto es, 112 U.

f) La distribución de frecuencias para el sexo se presenta a continuación.

Sexo	**Frecuencia Absoluta**	**Frecuencia Relativa**
Femenino	28	0.54
Masculino	24	0.46

Total	52	1.00

La razón de mujeres respecto a hombres se calculan:

g) Razón $_{F/M}$ = 28/24 = 7/6, es decir, en el grupo de pacientes por cada 7 mujeres habían 6 hombres.

h) De los 24 hombres, 15 tienen la hemoglobina por debajo de 13.2 mmol/l, entonces el porcentaje pedido será:

P_c = (15/24) x 100

= 62.5 %; el 62.5 % de los hombres estudiados tienen la hemoglobina por debajo de 13.2 U.

2. Según es planteado en el libro "La Salud Pública en Cuba. Hechos y Cifras", de la Dirección Nacional de Estadística del MINSAP, publicado en el año 1999, el país contaba en 1998 para la formación del personal de salud con 21 facultades de Medicina, 4 de Estomatología, 4 institutos superiores de Ciencias Médicas, 57 institutos politécnicos de la salud y 2 centros nacionales de perfeccionamiento.

   a) Construya una tabla estadística y un gráfico adecuado para reflejar dicha información.

   Una tabla que presenta la información disponible será la siguiente:

   Entidades para la formación del personal de salud de Cuba, 1998.

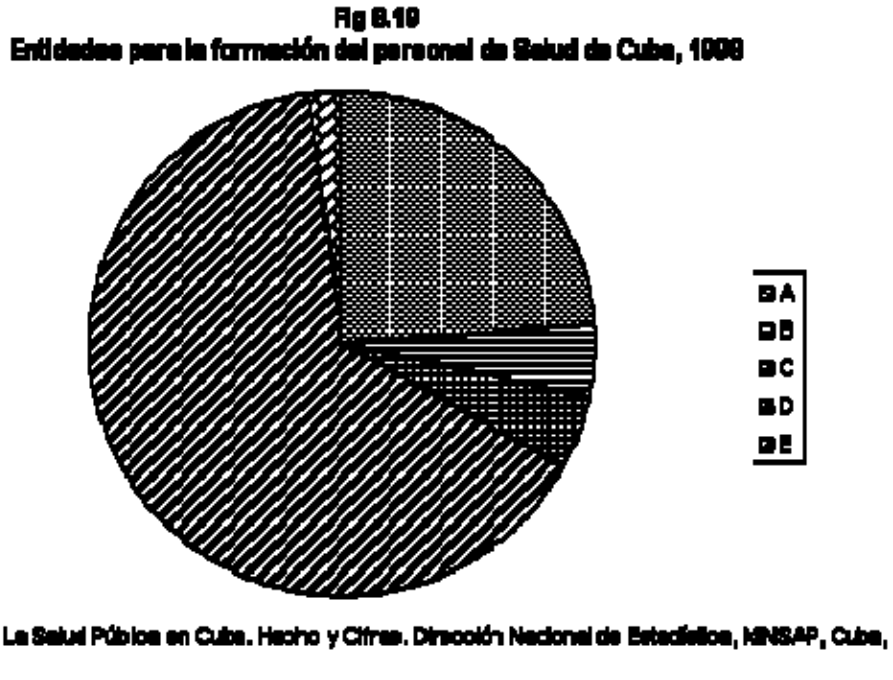
Entidad	**Cantidad**	**Porcentaje**
Facultades de Medicina	21	23.9
Facultades de Estomatología	4	4.5
Institutos superiores de Ciencias Médicas	4	4.5
Institutos politécnicos de la salud	57	64.8
Centros nacionales de perfeccionamiento	2	2.3
Total	88	100.0

Fuente: La Salud Pública en Cuba.Hechos y Cifras, Dirección Nacional de Estadística, MINSAP, Cuba,1999.

Los datos que aparecen en la tabla anterior pueden ser presentados mediante un gráfico de barras.

Figura 8.18

Estos datos también pueden exponerse a través de un gráfico de pastel, como el siguiente:



3. Notifica el libro Annual Health Statistics Report 1999 del National Statistics Bureau, del MINSAP en Cuba que el número de camas para cuidados médicos fue de 43 299 en 1975; 44 339 en 1980; 52 267 en 1985; 63 205 en 1990 y 66 116 en 1995 mientras que el número total de camas la cual incluye las de servicio social (en estas también se consideran las que se facilitan para uso privado) fue de 51244 en 1975; 53417 en 1980; 63784 en 1985; 77053 en 1990 y 80343 en 1995.

   a) Refleje la información en un cuadro estadístico y llévelo a un gráfico apropiado.

   El cuadro puede quedar conformado así:

   Camas para cuidados médicos y servicio social, Cuba, 1975-1995.

1975	43 299	7 945	57 244
1980	44 339	9 078	53 417
1985	52 267	11 577	63 784

1990	63 205	13 848	77 053
1995	66 116	14 227	80 343

Fuente: Health Statistic Report, 1999, National Health Statistic Bureau MINSAP, Cuba, 1999.

(1) Incluye las de uso privado.

Esta tabla pudiera confeccionarse sin la tercera columna si no es de interés, entonces el (1) se pondría en la columna de Total y en la nota explicativa: (1) El total incluye las de servicio social, donde a su vez se consideran las de uso privado.

La información de camas para cuidados médicos y total de camas puede ser reflejada mediante dos gráficos independientes o uno solo, a continuación un ejemplo de este último caso:

Figura 8.20

Se puede recurrir a una escala que no haya que interrumpir en el eje de las ordenadas como se vé en el gráfico siguiente:

Figura 8.21

4. Se cuenta con los datos correspondientes al nivel de colesterol de 30 personas que se realizaron dicho análisis el día 25 de marzo de 1999 en un hospital:

   3.4, 2.7, 6.4, 6.8, 7.9, 7.2, 4.5, 4.8, 3.7, 4.1, 3.9, 5.6, 5.2, 4.7, 7.2, 7.5, 6.3, 5.9, 4.0, 5.6, 3.9, 6.0, 5.4, 5.7, 4.2, 6.5, 7.3, 4.1, 4.6, 4.1

   a) Con los datos anteriores construir gráficos de: tronco y hoja, histograma, polígono de frecuencias, gráfico de frecuencias acumuladas, gráfico de caja y bigote.

   Gráfico de tronco y hojas.

   Nivel de colesterol. Pacientes que se realizaron análisis 25/3/99 Hospital X. País Y.

   Figura 8.22

   2 7

   3 4 7 9 9

   4 0 1 1 1 2 5 6 4 8

   5 2 4 6 6 7 9

   6 0 3 4 5 8

   7 2 2 5 5 9

   Fuente: Datos Laboratorio Hospital X

4. Día 25/3/99. País Y

Antes de contruir el resto de los gráficos, se debe confeccionar una distribución de frecuencias que pudiera ser la que aparece posteriormente:**Nivel colesterol**	**FA Absoluta**	**FAA Acumulada**	**Mc**
2.7-<3.4	4	1	3.05
3.4-<4.1	5	6	3.75
4.1-<4.8	7	13	4.45
4.8-<5.5	3	16	5.15
5.5-<6.2	5	21	5.85
6.2-<6.9	4	25	6.55
6.9-<7.6	4	29	7.25
7.6-<8.3	1	30	7.95
Total	30		

Figura 8.23

Figura 8.24

Figura 8.25

Figura 8.26

8.7 EJERCICIOS PROPUESTOS.

1. Se cuenta con los datos correspondientes al peso en Kg. del conjunto de personas del sexo femenino que han asistido a consulta la última semana en un consultorio del médico de la familia:

62.3 81.0 64.7 72.1 69.1 84.2 89.8

65.6 78.4 51.5 69.4 69.4 76.5 81.4

54.5	79.2	66.6	63.3	77.8	71.4	73.6
48.7	71.5	68.4	78.5	73.3	63.9	77.1
49.6	61.4	61.9	68.2	62.0	72.6	79.6
55.0	69.6	78.2	64.3	87.7	75.1	77.8
78.7	71.4	65.6	72.5	81.3	64.9	84.3
84.6	74.3	73.4	71.0	79.6	69.7	62.5
72.5	52.9	66.9	58.6	78.4	75.8	57.0
61.4	59.2	74.5	75.0	83.0	81.6	69.7
83.0	64.8	59.7	80.4	78.7	65.2	68.6

a) Construya una distribución de frecuencias completa con 5, 7 y 8 clases respectivamente.

b) Calcule a partir del conjunto de datos anterior las medidas de tendencia central y de dispersión que conoce.

c) Confeccione distribuciones de frecuencias completas con la cantidad de clases que Ud. desee.

d) Calcule las medidas de tendencia central y de dispersión estudiados en cada una de las clases encontradas en a) y c).

e) Calcule los percentiles 3, 5, 10, 15, 20, 50, 80, 85, 90, 95 y 97 del conjunto de datos anteriores.

f) Calcule las medidas descriptivas para datos agrupados en las variantes de a) y c).

g) Calcule e interprete para el conjunto de datos ofrecido los coeficientes de asimetría y curtosis estudiados.

2. Los resultados siguientes, pertenecen a los análisis de glicemia efectuados en el laboratorio de un policlínico, durante un mes:

7.5	3.9	5.6	6.6	6.9	6.5	6.1
2.8	8.4	6.3	4.5	5.4	5.8	5.2
6.3	6.7	5.1	9.6	5.1	5.4	5.9
4.7	7.9	5.4	9.8	4.8	3.2	4.3
3.6	5.3	4.9	7.6	4.2	8.8	7.8

3.8	4.5	7.6	4.3	4.1	6.7	7.7
5.6	3.9	7.1	4.9	4.6	8.1	4.6
5.4	3.8	5.8	5.6	4.4	5.5	5.7
4.7	4.9	5.8	4.7	5.6	6.4	

Utilizando estos datos:

a) Construya una distribución de frecuencias completas con 5 y otra con 7 clases

b) Escoja la cantidad de intervalos que Ud. desee y confeccione la distribuciones de frecuencia completa correspondientes.

Calcule:

c) Las medidas de tendencia central y de dispersión estudiadas.

d) Los percentiles 12, 19, 33, 48, 64, 87 y 98.

e) Los deciles 4 y 7, así como el cuartil 1.

f) Las medidas de tendencia central y de dispersión estudiadas para los datos de cada una de las clases obtenidas en los incisos a) y b).

g) Las medidas descriptivas para datos agrupados estudiadas en cada una de las variantes de los incisos a) y b). Compare las medidas obtenidas entre si y con los valores del inciso c).

h) Los coeficientes de asimetría y curtosis estudiados e interprete los resultados.

3. En una población de determinadas características se investiga la posible relación entre la hipertensión arterial y el hecho de contraer una enfermedad A: En aras de obtener una conclusión, se estudian 5 000 hipertensos debutantes y 5 000 no hipertensos durante 1 año para conocer cuántos desarrollan esa enfermedad A. Los resultados fueron los siguientes:

   Hipertensos:

   Desarrollaron la enfermedad 112

   No la desarrollaron 4 888

   No hipertensos:

   Desarrollaron la enfermedad 38

   No la desarrollaron 4 962

   Calcule el grado de asociación entre el factor de riesgo y la enfermedad A e interprete los resultados.

4. En un determinado país se quiere conocer si la edad está relacionada con la localización de la tuberculosis, para lo que se estudia una muestra de una población aquejada con dicha enfermedad, clasificada en grupos etáreos (menores de 15 años, entre 15 y 64 años y 65 años o más) y la localización (pulmonar y extrapulmonar). Calcule el grado de asociación entre las variables y el coeficiente de contingencia si los resultados fueron:

Grupos etáreos

Localización	<15	15-64	≥ 65
Pulmonar	13	461	128
Extrapulmonar	6	45	7

Calcule también el coeficiente de contingencia

5. Interesa saber si existe relación entre la tasa de tuberculosis en Cuba y la tasa de mortalidad, para lo que se cuenta con los datos de las 14 provincias. Halle los coeficientes de correlación lineal y por rangos.

   (Tasa por 100 000 hab.)

12.3	0.3
12.4	0.1
14.6	0.6
6.4	0.3
11.1	1.0
12.0	0.3
11.6	0.4
16.9	1.0
9.3	0.5
12.4	0.2
8.5	0.1
9.5	0.2
9.3	0.6
6.5	0.4

6. Se desarrolla una investigación con vistas a determinar si un método indirecto B menos costoso es capaz de detectar la hipercolesterolemia en igual medida que el método tradicional A. Para se

estudian 500 personas y se llega a los resultados siguientes:

Hay un total de 150 sujetos con hipercolesterolemia y:

	Método A	Método B
Falsos negativos	15	18
Falsos positivos	25	20

Calcule de ambos métodos la sensibilidad y la especificidad, y compare los resultados.

7. Un nuevo medicamento (B) para la cura de la tuberculosis pulmonar está siendo valorado por un equipo de investigadores. Pruebas de laboratorio han arrojado en un estudio con animales que de 800 tratados con este medicamento, 500 fueron curados, 200 mejoraron y 100 permanecieron iguales o empeoraron; utilizando el método tradicional (A), de 600 enfermos y tratados, 310 se curaron, 180 mejoraron y 110 se mantuvieron iguales o peor, calcule el coeficiente χ^2 y el de contingencia.

8. Se estudia la posible asociación entre contraer una enfermedad de las vías digestivas y haber practicado por espacio de más de tres meses una cierta dieta entre 3 y 5 años atrás. Con vistas a obtener una respuesta fueron estudiadas 140 personas que se habían sometido a esa dieta.

   - Enfermos, dieta menos de 3 meses 5
   - Enfermos, dieta más de 3 meses 15
   - Sanos, dieta menos de 3 meses 50
   - Sanos, dieta más de 3 meses 70

   Calcule el grado de asociación a través del coeficiente χ^2 y el de contingencia. Utilice la corrección de Yates.

9. En un estudio sobre pacientes aquejados de sicklemia en una región, se cuenta con los datos de 36 sujetos:

Paciente	**Ingreso**	**Edad**	**Veces consulta**	**Estadía (días)**
1	No	36	2	0
2	Si	29	5	15
3	Si	65	4	15
4	Si	48	3	12
5	No	18	3	0
6	Si	34	4	25
7	No	48	1	0
8	Si	23	5	28
9	No	26	5	0
10	No	29	4	0
11	Si	43	2	19
12	Si	21	3	19
13	Si	37	2	16

14	No	28	3	0
15	Si	26	4	36
16	Si	33	6	38
17	No	38	4	0
18	Si	45	4	12
19	No	66	2	0
20	Si	29	9	35
21	Si	28	5	12
22	Si	39	6	5
23	No	45	4	0
24	No	35	7	0
25	Si	43	2	24
26	No	24	1	0
27	Si	18	5	16
28	Si	14	3	24
29	No	35	4	0
30	Si	16	4	17
31	No	33	7	0
32	Si	26	3	16
33	Si	19	3	17
34	Si	23	4	23
35	No	26	3	0
36	Si	19	4	6

Veces consulta: número de veces que ha asistido a consulta en los últimos 12 meses.

Estadía: total de días que ha estado ingresado en los últimos 12 meses.

a) Construya una distribución de frecuencias para cada variable.

b) Calcule e interprete la proporción de pacientes que ingresaron en estos últimos 12 meses.

c) Calcule e interprete la razón de pacientes que no ingresaron con respecto a los que sí lo hicieron en estos últimos 12 meses.

d) Calcule e interprete el porcentaje de pacientes que tienen menos 30 años.

e) Calcule e interprete la razón de pacientes que asistieron menos de 3 veces a consulta con respecto a los que asistieron más de 4 veces en estos últimos 12 meses.

f) ¿Qué porcentaje de los pacientes que estuvieron ingresados permanecieron más de 20 días en la institución hospitalaria?

10. En el libro La Salud Pública en Cuba. Hechos y Cifras de la Dirección Nacional de Estadística

del MINSAP publicado en el año 1999 en Cuba aparece reflejado que el país tenía en 1998 como recursos humanos para atender a la población a 63 483 médicos, 9 873 estomatólogos, 82 527 enfermeras y 121 364 técnicos medios de la salud.

a) Construya una tabla estadística con la información anterior

b) Construya un gráfico de barras y un gráfico de sector.

11. Por medio del libro, La Salud Pública en Cuba. Hechos y Cifras, publicado por la Dirección Nacional de Estadística del MINSAP y algunos organismos internacionales en 1999, se conoce que Cuba disponía como unidades de asistencia médica y social en el año 1998 con 283 hospitales en número de, 440 policlínicos en, 166 clínicas estomatológicas, 161 puestos médicos, 12 institutos de investigación, 231 hogares maternos, 197 hogares de ancianos y 29 hogares de impedidos.

Construya, con esa información, una tabla estadística y los gráficos apropiados.

10. Según datos reportados en el libro La Salud Pública en Cuba. Hechos y Cifras de la Dirección Nacional de Estadística del MINSAP de 1999, en Cuba se notificaron en niños de 1 a 4 años de edad, 10 casos de fiebre tifoidea en 1990; 10 en 1991; 6 en 1992; 42 en 1993; 8 en 1994; 18 en 1995; 10 en 1996; 8 en 1997 y 7 casos en 1998.

Refleje los datos en un cuadro estadístico y construya gráficos adecuados.

11. En Cuba había 8 531 estomatólogo en 1993; 8 834 en 1994; 9 418 en 1995; 9 600 en 1996; 9 816 en 1997; 9 873 en 1998 y 9 918 en 1999; usando tasas por cada 10 000 habitantes había en 1993 una tasa de 7.8 odontólogos; en 1994 una de 8.1; en 1995 una de 8.3; en 1996 una de 8.7 y a partir de 1997 una de 8.9 (las cifras de 1999 son provisionales) estos datos fueron tomados del Annual Health Statistics Report 1999 del National Health Statistics Bureau del MINSAP en Cuba en 1999.

a) Construya una tabla estadística con los valores anteriores.

b) Grafique la información.

12. En Cuba la incidencia de hepatitis viral B y la viral sin especificar fue de 1 747 y 1 487 casos respectivamente en 1994; de 1 456 y 1 042 en 1995; 1 695 y 1 217 en 1996; de 1 345 y 657 en 1997; 1 026 y 426 en 1998 según publica el Annual Health Statistics Report 1999 del National Health Statistics Bureau del MINSAP de Cuba.

Presente:

a) La información en un cuadro estadístico construye los diferentes gráficos posibles.

13. Los nacidos vivos en hospital en el primer semestre de 1998 se comportaron de la forma siguiente: en enero 13 642, en febrero 11 143, en marzo 11 162, en abril 9 983, en mayo 10 087 y en junio 10 731.

a) Disponga en una tabla estadística los datos anteriores.

b) Refleje la información en gráficos adecuados

14. Se tienen los datos siguientes correspondientes al nivel de fisiológica de pacientes adultos que se atendieron por presentar enfermedad diarreica aguda en un consultorio médico, en el mes de marzo de 1999, según consta en su registro.

Nivel de la variable	Cantidad de pacientes
10.5–10.9	2
11.0–11.4	4
11.5–11.9	6
12.0–12.4	7
12.5–12.9	3
13.0–13.4	2
Total	24

a) Construya un histograma, un polígono de frecuencias y un gráfico de frecuencia acumulada

Si los valores en particular fueron:

10.7, 11.8, 13.1, 12.7, 10.6, 12.9, 11.6, 11.3, 12.7, 12.2, 12.3, 13.4, 11.2, 11.2, 12.1, 12.3, 12.1, 11.9, 11.6, 12.4, 11.2, 12.0, 11.8, 11.9

b) Construya un gráfico de tronco y hojas y otro de caja y bigote.

c) Construya otra distribución de frecuencia con un número diferente de intervalos de clase y dibuje los gráficos correspondientes.

17. Busque los ejemplos y ejercicios resueltos expuestos anteriormente en el capítulo y construya otros gráficos diferentes que sean adecuados. En el caso de las variables cuantitativas busque otras distribuciones de frecuencia con diferentes amplitudes y represéntelas gráficamente.

Capítulo 9. Nociones básicas acerca de la Teoría de las Probabilidades e Introducción a la Inferencia Estadística

9.1 INTRODUCCIÓN.

Las situaciones reales, usualmente, no adoptan la forma de un problema matemático. El mundo real es demasiado complejo para poder ser descrito con exactitud por medio de ecuaciones matemáticas. Esto se hace más evidente aún, cuando el objeto de estudio es un ser vivo. Por ello, para resolver la mayoría de los problemas que se presentan en la realidad, se acude a formulaciones simplificadas de los mismos que reciben el nombre de modelos matemáticos o estadístico-matemáticos. Es decir, se **idealiza** el problema.

El concepto intuitivo de probabilidad, como una medida de la posibilidad, es un concepto clave en el desarrollo de los modelos estadísticos. Por ello es fundamental el conocimiento de los principios básicos de la Teoría de Probabilidades: para poder aprender a utilizar e interpretar correctamente las técnicas y métodos de la Estadística.

La Estadística trabaja con un tipo muy particular de variables: las variables aleatorias. Asociado al concepto de variable aleatoria e indisolublemente unido a él, está el concepto de probabilidad.

Además, la solución de los problemas estadísticos se da en términos de afirmaciones probabilísticas. La magnitud de la probabilidad asociada a una conclusión representa el grado de confianza que se puede depositar en la certeza de esa afirmación.

Abordaremos de inicio el cálculo de probabilidades para las variables aleatorias que tienen un número finito de resultados posibles y el análisis de las variables aleatorias continuas se tratará más adelante.

9.2 CONCEPTOS BÁSICOS DE LA TEORÍA DE LAS PROBABILIDADES.

Cuando estudiamos algunas disciplinas de la Física, en particular la Mecánica Clásica, nos enseñan a expresar los fenómenos por medio de ecuaciones matemáticas. Quizás el caso más simple es el del movimiento rectilíneo uniforme, cuando se nos indica que:

$$V = S / T$$

Donde:

V : Velocidad S : Espacio recorrido T : Tiempo

Esta ecuación significa que la velocidad, en este tipo de movimiento, es igual al espacio recorrido en una unidad de tiempo. En realidad esto es un modelo en el que se ha despreciado, entre otras cosas, la fricción que se produce entre el móvil y la superficie de desplazamiento. Aún así, el nos permite predecir, con cierto grado de certidumbre, el espacio que puede recorrer un móvil que se mueve a velocidad constante durante un determinado período.

En el caso de que nuestro objeto de estudio sean seres vivos, es relativamente fácil darse cuenta de que no es tan predecible el resultado de una experiencia aunque conozcamos una serie de condiciones o características de los individuos involucrados en la misma. Veamos algunos ejemplos de resultados impredecibles:

1: Estatura adulta del hijo de una pareja.

Aunque tengamos toda la información antropométrica, de salud y socio-económica de ambos miembros de la pareja, resulta imposible conocer con exactitud cual será la talla final de un hijo.

2: Niveles de lípidos en el suero de un sujeto sano del sexo masculino.

No hay ningún mecanismo o procedimiento que nos permita conocer esas cifras, como no sea la extracción de una muestra de sangre y la valoración directa del nivel de dichas sustancias en el suero obtenido.

Ni todos los hijos (de igual sexo) de una pareja tienen la misma talla, ni todos los hombres sanos tienen lípidogramas iguales. Estas diferencias son el producto de factores que no podemos **controlar** o que no conocemos que influyen sobre esas características (talla adulta y nivel de lípidos), y que los resultados sean diferentes a pesar de realizar las observaciones en igualdad de condiciones.

Sobre la base de estos ejemplos, se analizaran algunos conceptos que resultan básicos para la teoría de las probabilidades y la estadística.

Es este tipo de característica (variable) la que le interesa a ambas teorias y recibe el nombre de **variable aleatoria**. El calificativo de aleatoria se emplea para indicar, como dijimos anteriormente, que sobre la variable actúan factores no controlables o desconocidos que provocan que los valores que tome la misma sobre un determinado sujeto no sean expresables por medio de una formulación matemática determinística. A ese conjunto de factores es al que, habitualmente, se denomina azar.

Si pensamos un poco más en estos ejemplos podemos darnos cuenta de que estas diferencias entre los valores de la variable aleatoria en individuos con características similares o iguales no son de cualquier magnitud. Es decir, no son fenómenos en los que reina la anarquía. En general, esos resultados se encuentran dentro de un determinado rango de valores posibles. Sabemos, por ejemplo, que:

a) Padres de talla elevada tienen, generalmente, hijos de alta estatura y padres de baja talla, también generalmente, tienen hijos de estatura pequeña.

b) Las cifras del perfil lipídico de un sujeto sano deben encontrarse dentro de determinados límites: cifras superiores pueden indicar la presencia de una enfermedad.

Cuando se analizan grandes cantidades de observaciones de variables de este tipo (aleatorias), se observa que las mismas muestran cierta regularidad en su comportamiento, regularidad que puede expresarse a través de leyes, las cuales reciben el nombre de **leyes estadísticas o probabilísticas**. Son precisamente estas leyes el objeto de estudio fundamental de la **teoría de las Probabilidades**.

El estudio y conocimiento de estas leyes, que resumen en si la forma de comportamiento de las variables aleatorias y muchas veces pueden expresarse de forma matemática, es de suma importancia, entre otras cosas para poder discernir el tipo de técnica o modelo estadístico que debamos usar para analizar los resultados de una variable aleatoria dada.

A continuación veremos, en forma más bien intuitiva, algunos de los conceptos elementales de la teoría de las probabilidades que nos servirán de base para el estudio de esas leyes, se refiere esto a la noción elemental de lo que es un evento o suceso y a diferentes tipos de eventos que resultan de interés.

Aunque las probabilidades se aplican a una amplia gama de situaciones prácticas, la comprensión del tema se hace más sencilla si se explica a partir de situaciones no prácticas, algo idealizadas, y en problemas aparentemente más simples, como son algunos juegos de azar. Es por esa razón que introduciremos los conceptos y las ideas intuitivas a partir de ese tipo de situaciones.

Consideremos un experimento repetitivo simple tal como tirar una moneda dos veces (o tirar dos monedas una vez). En esta experiencia hay cuatro resultados posibles:

Primera tirada	Segunda tirada
estrella	estrella
estrella	escudo
escudo	estrella
escudo	escudo

Solo si interesa de manera especial el orden de las tiradas (o de cual de las dos monedas es el resultado) se consideraran como resultados diferentes el segundo y el tercero. A cada uno de esos posibles resultados se les llama, en la jerga probabilística, eventos simples. En esta experiencia solo hay 4 resultados o eventos simples posibles y todos, teóricamente, tienen las mismas posibilidades de ocurrir (despreciando la posibilidad de que la moneda caiga de canto y se quede sostenida por el borde, sin caer a un lado u otro).

Al conjunto de todos los eventos simples (resultados posibles) de una experiencia se les denomina **espacio muestral** de dicha experiencia.

En el caso de la tirada doble de un dado (o la tirada de dos dados) el número de resultados posibles es mucho mayor:

11	21	31	41	51	61
12	22	32	42	52	62
13	23	33	43	53	63

14	24	34	44	54	64
15	25	35	45	55	65
16	26	36	46	56	66

donde el primer dígito de cada par representa la primera tirada (o el primer dado) y el segundo dígito la segunda tirada (o el segundo dado).

En este conjunto de resultados posibles pueden interesarnos cosas tales como:

- que la suma de las dos tiradas (caras) sea par

- que uno de los números sea el 6

- que la suma de las dos tiradas (caras) sea menor que un determinado valor

Es decir, puede interesarnos algo más que el simple resultado de las tiradas. Los resultados de este tipo, que pueden producirse por más de un evento simple, reciben el nombre de eventos compuestos.

Los conceptos de evento simple y evento compuesto son relativos, pues dependen de la definición inicial del problema que vamos a enfrentar.

En el caso de los eventos compuestos, es fácil comprobar que la probabilidad correspondiente se obtiene mediante la suma de las probabilidades de los eventos simples que lo componen.

Para poder expresar todo esto matemáticamente necesitamos una notación. Utilizaremos la siguiente:

- los eventos simples se denotarán como e_i y Las probabilidades asociadas a los eventos simples como $P(e_i)$.

- los eventos compuestos se denotarán con letras mayúsculas: A, B, C, D, etc, mientras que las probabilidades asociadas a los eventos compuestos como P(A), P(B), etc.

Usando esta notación podemos dar la definición de probabilidad de un evento compuesto:

Definición: la probabilidad de que un evento compuesto A ocurra es la suma de las probabilidades de los eventos simples que lo integran. Si el evento A ocurre al tener lugar cualesquiera de los eventos simples e_2, e_4 y e_6, entonces:

$$P(A) = P(e_2) + P(e_4) + P(e_6)$$

En muchas experiencias con juegos de azar se espera que todos los eventos simples ocurran con la misma frecuencia relativa. Esto puede ser cierto en las experiencias de tirar monedas o dados, pero no siempre lo es. Supongamos que tenemos una urna o bolsa donde se encuentran tres bolas rojas, dos negras y una verde, de forma tal que las bolas de idéntico color no puedan distinguirse una de otra. Si la experiencia consiste en

extraer una bola, solo tenemos 3 resultados posibles: o se saca una bola roja, o se saca una bola negra o se saca la bola verde. Cada uno se esos resultados tiene una probabilidad de ocurrir diferente a la de los otros dos, puesto que la cantidad de bolas de cada color es diferente.

Esta situación, de probabilidades diferentes asociadas a los elementos del espacio muestral es la más frecuente.

Consideremos solo dos eventos, A y B, asociados a una experiencia y supongamos que nos interesa conocer la probabilidad que tienen de ocurrir cuando se realice dicha experiencia. El evento conjunto se denota habitualmente (A y B) y su probabilidad por P(A y B). También podemos estar interesados en conocer la probabilidad de que ocurra alguno de los dos eventos (al menos uno) cuando se realice la experiencia. Este evento se denota (**A o B**) y su probabilidad por **P(A o B)**. Aquí el monosílabo 'o' quiere decir uno, el otro o ambos.

Si dos eventos A y B poseen la propiedad de que la ocurrencia de uno impide o previene la ocurrencia del otro, se dice que son eventos mutuamente excluyentes (o disjuntos**)**. Un ejemplo de eventos de esta clase, en la tirada de dos dados es:

- A: la suma de las dos caras es menor que 7.
- B: la suma de las dos caras es mayor que 10.

Otro tipo de eventos que resulta importante conocer son los conocidos como **eventos independientes**. Esto significa que la ocurrencia de uno de ellos no afecta en lo más mínimo la ocurrencia del otro. Por ejemplo, ¿cual es la probabilidad de que una mujer que ha tenido tres hijos varones tenga una niña en su cuarto embarazo?. No es difícil darse cuenta de que la probabilidad de tener una hija no varía en lo absoluto por el hecho de haber tenido embarazos anteriores, sean cuales sean los sexos de los productos de esos embarazos, puesto que cada mbarazo es totalmente independiente del anterior: son células germinales nuevas.

El concepto de independencia se hace extensivo a las variables aleatorias y en muchas ocasiones nos enfrentaremos al problema de comprobar si dos resultados posibles o dos variables aleatorias son independientes. Por ejemplo, cuando se investiga sobre posibles factores de riesgo para alguna enfermedad o condición que influye sobre la salud de un individuo lo que se busca, en sentido general, son los resultados o variables que están asociados a la enfermedad o condición, y son previos a esta última: se trata de probar la no independencia.

Cuando dos eventos no son independientes, sino que están relacionados de alguna manera, se utiliza el concepto de eventos condicionados o relacionados. Es decir, si la ocurrencia de uno de ellos condiciona de alguna manera la ocurrencia del otro se dice que los eventos están relacionados.

Ya hemos visto el primer paso a dar en la construcción del modelo matemático de la experiencia: identificar los eventos y resultados posibles. El segundo paso es asignar a cada evento del espacio muestral un número que nos indique en qué medida es posible que ocurra cada uno de ellos, y eso es lo que analizaremos en el acápite siguiente.

9.3 DEFINICIÓN CLÁSICA Y FRECUENCIAL DEL CONCEPTO DE PROBABILIDAD.

Supongamos que un evento E puede ocurrir de k formas diferentes de un total de n resultados igualmente posibles. Entonces la probabilidad de ocurrencia del evento se denota:

$P(E) = k/n = p$

Esta es la definición clásica del concepto de probabilidad, En el caso de la tirada de dos dados la frecuencia de presentación cada uno de los eventos simples (considerando 12 diferente de 21, al tener en cuenta el orden de tirada, o la diferencia entre los dados, quizás a partir del color) sería 1/36, es decir, una de las 36 respuestas posibles. La frecuencia de presentación de una suma de las caras que sea un numero par (evento compuesto, pues se produce de varias formas posibles) sería el número de eventos simples en que lo expresado se produce sobre los 36 resultados posibles, es decir, 18/36 = 1/2.

Estas respuestas serían absolutamente ciertas si los dados fuesen perfectos, es decir, si el peso de los mismos esta bien balanceado (en términos físicos, si coinciden los centros de gravedad, masa, etc.). Si el dado no esta cargado, aunque no sea perfecto, de acuerdo con lo que se conoce como Ley de Regularidad Estadística se obtiene un resultado similar si se calcula la frecuencia relativa cuando repetimos la experiencia un número suficientemente grande de veces, lo que nos lleva al concepto de probabilidad como límite de las frecuencias relativas.

Cuando decimos **Ley de Regularidad Estadística** se hace referencia a una propiedad de las frecuencias relativas de un evento observable A en un fenómeno aleatorio, es decir a, **la tendencia de las frecuencias relativas de A, calculadas sobre la base de diferentes conjuntos de observaciones de A, cada uno de ellos con un número suficientemente grande de las mismas, a mantenerse próximas a determinado valor constante**, propio o particular de cada evento dado A, conocido como **probabilidad o frecuencia esperada del evento A** y denotado por P(A)..

Veamos esto con un ejemplo.

Considere que se observa el sexo de nacidos vivos en muestras sucesivas, de mayor tamaño en cada ocasión. Intuitivamente se piensa que la mitad de los nacimientos serán varones y la otra mitad niñas, pero si se comprueba. Los resultados de dichas observaciones podrían ser los que se presentan en la tabla siguiente:

Tabla 9.1 Nacidos vivos del sexo masculino para diferentes tamaños muéstrales.

Número de	Total de nacimientos	Sexo masculino	
muestra	(n)	#	%
1	150	50	33.3
2	300	181	60.3

3	450	165	36.7
4	600	222	37.5
5	750	469	62.5
6	900	437	48.6
7	1 050	681	64.9
8	1 200	642	53.5

.

.

.

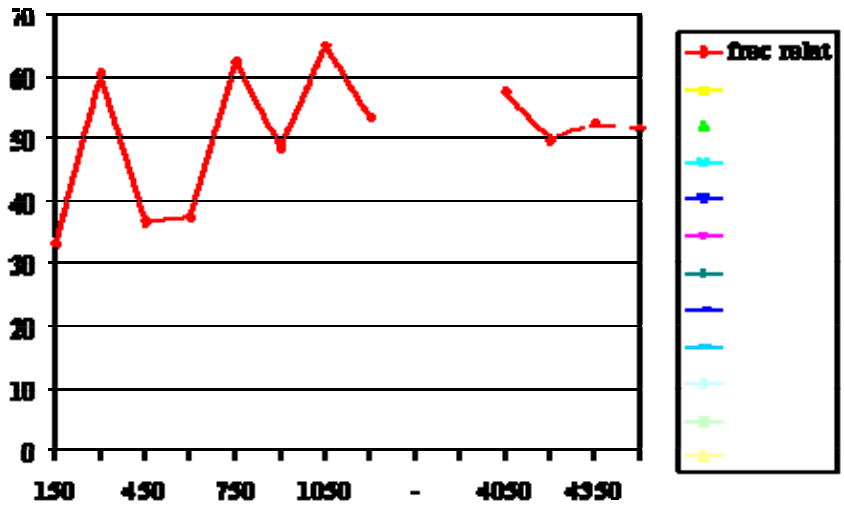
.

.

27	4 050	2329	57.5
28	4 200	2091	49.8
29	4 350	2288	52.6
30	4 500	2331	51.8

Si se hace un gráfico aritmético simple (fig. 9.1) que muestre el comportamiento de las frecuencias relativas de nacimientos de varones en función del tamaño de la muestra de que proceden, se tendría la representación siguiente:

Fig 9.1



Este gráfico da una idea bastante clara de que lo que se desea expresar cuando se habla de ley de regularidad estadística (tendencia observable en las frecuencias relativas a irse estabilizando alrededor de cierto valor, a medida que aumenta el número de observaciones). Esta es una cualidad intrínseca del fenómeno, totalmente independiente de quien realice la observación, siempre y cuando permanezcan fijas toda una serie de condiciones.

Estos valores pueden también estimarse a partir de un gran número de observaciones en momentos diferentes del tiempo. Por ejemplo:

Tabla 9.2 Nacidos vivos según sexo. Cuba 1991-1998.

		Sexo masculino		Sexo femenino	
	Total de nacidos vivos	#	%	#	%
1991	173 896	89 052	51.21	84 844	48.79
1992	157 349	80 610	51.23	76 739	48.77
1993	152 233	78 019	51.25	74 214	48.75
1994	147 265	75 503	51.27	71 762	48.73
1995	147 170	75 439	51.26	71 731	48.74
1996	140 276	71 905	51.26	68 371	48.74
1997	152 681	78 234	51.24	74 447	48.76

1998	151 080	77 413	51.24	73 667	48.76

Como puede verse, la probabilidad de que nazca un varón es 0.512 aproximadamente, y no 0.500 como se puede pensar y se hizo en tiempos pasados..

El concepto de probabilidad como se dijo, se refiere, a un valor numérico específico para cada evento (el que puede estar expresado por medio de una variable aleatoria), en él esta siempre implícita la idea de la realización de un número grande de observaciones. Por ello, este valor será útil cuando se pretende predecir o pronosticar lo que ocurrirá dentro del marco de muchas observaciones y no para resultados aislados. Es imposible, a punto de partida de las probabilidades asociadas a los resultados posibles de una variable aleatoria, predecir exactamente lo que ocurrirá en una observación futura.

Mediante una fórmula podemos calcular con exactitud la presión que ejercería un determinado gas noble sobre las paredes del recipiente que lo contiene, si conocemos el volumen del recipiente y la temperatura que tiene el gas. En los fenómenos aleatorios solo se puede tener, a lo sumo, una idea aproximada de lo que tiene mayores posibilidades de ocurrir.

Ejemplifiquemos estas ideas:

a) Si conocemos que la tasa de prevalencia de una enfermedad congénita "X" en la población cubana es de un 0,03% pensaremos que es poco posible que nazca un niño con la misma.

b) Al recibir a un paciente mayor de 50 años con sacrolumbalgia en la consulta, el médico piensa, en primer lugar, en las afecciones más frecuentes: artrosis, inestabilidad vertebral en la región lumbosacra, etc. Puede ser que el paciente tenga espina bifida, pero como esta es una enfermedad mucho menos frecuente y es de nacimiento, es difícil que a esa edad no haya sido descubierta. Insisto, puede ser que el paciente tenga espina bífida, pero las posibilidades de que la causa del dolor sean las otras afecciones son muchas más. Por ello es que se ha dicho que la probabilidad es una forma de expresar numéricamente la posibilidad de que ocurra un determinado resultado de una variable aleatoria.

Dado que las probabilidades o frecuencias relativas esperadas asignadas a los eventos son, frecuencias relativas esperadas basadas en consideraciones lógicas o límites de frecuencias relativas, estas deben cumplir con las propiedades de las frecuencias relativas. Estas propiedades son:

a) $0 \leq P(A) \leq 1$, para cualquier evento A que se considere.

b) $P(S) = 1$.

Por ejemplo, esto significa que si S contiene un número finito de eventos simples, la suma de todas sus probabilidades es 1. De otro modo si S esta vinculado a la realización de una variable aleatoria que tome solamente un número finito de valores posibles, entonces la propiedad se interpreta como que la suma de todas las frecuencias relativas esperadas o probabilidades de cada uno de los valores que asume la variable es la unidad.

c) Si dos eventos son excluyentes (o disjuntos), entonces P(A o B) = P(A) + P(B).

Esta propiedad expresa una importante característica de esta forma de medir o valorar las cosas, que es compartida por otras formas de medir conocidas como son, el medir longitudes (áreas, volúmenes), ya que cuando deseamos medir la longitud de algo que, por ejemplo sobrepase la longitud del instrumento de medición (una regla, una cinta, etc), entonces es usual, dividir lo que se desea medir en pedazos, medir estos y mediante una suma hallar el resultado pedido.

Poder sumar probabilidades, esto es lo esencial de la propiedad planteada, por ello cuando se quiere hallar, la probabilidad del evento A o B, como suma de P(A) y P(B), siempre se debe comprobar que los eventos A y B son excluyentes o disjuntos..

Como ejemplos de aplicación de las propiedades o axiomas anteriores veamos que: $P(\emptyset) = 0$ y que $P(A) + P(A^c) = 1$.

Como S y Ø son excluyentes, entonces P(S o Ø) = P(S) + P(Ø), pero por (a) P(S) = 1 y S o Ø = S, luego 1 = 1 + P(Ø), de donde P(Ø) = 0.

Se tiene que A y A^c, son eventos excluyentes, luego $P(A \text{ o } A^c) = P(A) + P(A^c)$ por (c), pero el evento A o A^c es una representación del evento S, luego $P(A \text{ o } A^c) = P(S) = 1$, luego $P(A) + P(A^c) = 1$.

d) No tiene unidades de medida.

Cuando el valor de probabilidad asociado a un evento es muy pequeño (esta muy cercano a 0), se considera que es poco posible que dicho suceso ocurra. Al contrario, si el valor esta muy cercano a 1, se espera que ese evento ocurra casi siempre.

El valor de probabilidad 0 es prácticamente imposible de obtener en una experiencia real. Cuando los valores que se obtienen son muy pequeños, para poder decidir si difieren realmente de 0 o no, se emplean técnicas estadísticas especiales, similares a las que se tratarán en el capítulo 10 cuando se traten los temas relativos a las **prueba de hipótesis**.

En el campo de las Ciencias Médicas el empleo de las probabilidades está muy generalizado. Veamos a continuación algunos de los usos comunes de las mismas en la práctica cotidiana de la medicina.

1. Al establecer un diagnóstico presuntivo (preliminar o impresión diagnóstica).
2. Si el paciente presenta tales síntomas y signos es muy probable que la enfermedad que padezca sea la X, pues la Y y la Z difieren en tales signos o síntomas, o porque aún teniendo cuadros similares, su incidencia en el lugar geográfico es nula o prácticamente nula.
3. Al valorar la posibilidad de que un paciente pueda presentar complicaciones, es usual que se valoren los tipos de complicaciones y que medidas tomar para evitarlas o disminuir la afectación al paciente.
4. En este caso el médico debe tomar en cuenta, en principio, cuáles son las complicaciones más frecuentes en esos casos, los antecedentes patológicos del enfermo que pueden coadyuvar a las complicaciones, etc.

5. Al decidir un tratamiento para un paciente en particular.

Aunque puedan existir muchos medicamentos para indicar, el médico debe valorar en cada caso en específico cuál de los medicamentos existentes puede resultar efectivo y producir menos reacciones adversas (o ninguna). Esto no solo tiene que ver con los antecedentes patológicos del sujeto, sino también con características de su personalidad, físicas, etc.

La interpretación o la valoración realizada por el médico en todos estos casos tiene una base probabilística, ya que en la misma este se apoya en el resultado de largas series de observaciones realizadas sobre gran número de enfermos y en la aplicación consecuente de la ley de regularidad estadística.

En resumen, la probabilidad puede interpretarse como:

a) el **valor teórico** (**frecuencia relativa teorica**) al cual tiende o se aproxima la frecuencia relativa observada de un evento, para un número elevado de repeticiones bajo condiciones estables.

b) una **medida de la posibilidad** de que el resultado asociado a dicho valor pueda ocurrir o no cuando se hace una sola observación del evento.

9.4 PROPIEDADES BÁSICAS DE LAS PROBABILIDADES.

Las aplicaciones de las probabilidades se realizan en general para un cierto número de eventos, usualmente compuestos, lo que hace necesaria la existencia de reglas que faciliten el cálculo de esas probabilidades. Veamos algunas de esas reglas.

9.4.1. Teorema de la Adición.

Cuando dos eventos A y B no son mutuamente excluyentes existen resultados elementales que se correspondan con la ocurrencia de ambos, es decir, los dos eventos tiene resultados comunes dentro del espacio muestral.

Cuando puede aplicarse la definición clásica de probabilidad, la probabilidad de (A o B) será el número de veces que ocurren A o B, sobre el número total de elementos del espacio muestral. Si denotamos por n(A y B), n(A) y n(B) y al número de veces que ocurre A y B, A, y B respectivamente, entonces

$$P(A \text{ o } B) = [n(A) + n(B) - n(A \text{ y } B)]/\ n = n(A)/n + n(B)/n - n(A \text{ y } B)/n$$

$$= P(A) + P(B) - P(A \text{ y } B)$$

Este resultado es el que se conoce como **teorema (o regla) de la Adición de probabilidades** y puede generalizarse a cualquier número de eventos.

Debe notarse que cuando los eventos A y B son mutuamente excluyentes entonces $P(A \text{ y } B) = P(\emptyset) = 0$ y en consecuencia, $P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$, se reduce a $P(A \text{ o } B) = P(A) + P(B)$, la propiedad o

axioma (c) de las probabilidades. Luego el teorema de la adición no es mas que una propiedad que amplia el uso de (c) al caso en que no se conoce nada sobre los eventos A y B.

9.4.2. Teorema de la multiplicacion.

En el caso de eventos independientes se hace equivalente a su definición la expresión siguiente: P(A y B) = P(A) . P(B).

Es decir, la probabilidad de que ocurran A y B es igual al producto de las probabilidades de ambos eventos, expresión que también es conocida como Regla de la Multiplicación o del Producto de las probabilidades, la que se generaliza también para cualquier número de eventos independientes.

Esta supuesta equivalencia se debe a que se define la probabilidad condicional entre dos eventos como: P(A/B) = P(A y B)/P(B), donde P(B) > 0.

Expresión que se lee como: probabilidad de que ocurra A dado que ocurrió B es igual a la probabilidad de que ocurran ambos entre la probabilidad de B.

Si sustituimos P(A y B) por el producto de probabilidades (bajo el supuesto de que A y B son independientes), tendremos: P(A/B) = (P(A).P(B))/P(B) = P(A).

Es decir, la probabilidad de que A ocurra no se ve afectada por la ocurrencia de B, que es precisamente la definición de independencia.

Veamos un ejemplo de uso de la independencia entre eventos:

La administración de un medicamento tiene una probabilidad de un 10% (0.1) de provocar una crisis hipertensiva. Un paciente requiere que se le administre 5 veces dicho medicamento. ¿Cuál es la probabilidad de que este sufra al menos una crisis hipertensiva?.

Se considera que el paciente cumple la condición si presenta 1, 2, 3, 4 ó 5 crisis hipertensivas. El método directo sería calcular las probabilidades asociadas a cada uno de esos cinco resulta, dos(que son excluyentes) y sumarlas, pero es mucho más cómodo hacerlo a partir de la relación complementaria.

Si denotamos x al número de crisis, entonces, $P(x \geq 1) = 1 - P(x = 0)$.

Como tenemos que, P(crisis) = 0.1, en consecuencia P(no crisis) = 0.9

Si cada administración del medicamento es considerada independiente de las otras: entonces, $P(x = 0) = 0.90 \cdot 0.90 \cdot 0.90 \cdot 0.90 \cdot 0.90 = 0.59$.

Luego $P(x \geq 1) = 1 - 0.59 = 0.41$ es la probabilidad de que el paciente sufra al menos una crisis hipertensiva en 5 administraciones del medicamento.

Estas reglas tienen un uso importante en el cálculo de probabilidades. Veamos, un ejemplo del uso de las probabilidades condicionadas y las reglas de probabilidades vistas. Dadas dos urnas, I y II, supongamos que la urna I contiene 4 bolas negras y 7 bolas blancas, y que la urna II contiene 3 bolas negras, 1 blanca y 4 amarillas. Si seleccionamos una urna al azar y extraemos una bola, ¿cuál será la probabilidad de que la misma sea negra?.

Denominando como U_1 al evento, seleccionar la urna I, y U_2 al evento, seleccionar la urna II, tenemos: $P(U_1) = P(U_2) = ½$.

Si denotamos N al evento, extraer una bola negra, se tiene que : $P(N/U_1) = 4/11$ y $P(N/U_2) = 3/8$.

El evento, extraer una bola negra, se puede escribir: N = (N y U_1) o (N y U_2), y como los eventos (N y U_1) y (N y U_2) son excluyentes (o disjuntos), se puede aplicar directamente la propiedad (c) teorema de la suma de probabilidades y escribir que:

$P(N) = P(N \text{ y } U_1) + P(N \text{ y } U_2)$

Y haciendo uso de la definición de probabilidad condicional cada sumando de la expresión anterior se puede poner como sigue:

$P(N/U_1) = P(N \text{ y } U_1)/P(U_1)$ y $P(N/U_2)$

$= P(N \text{ y } U_2)/P(U_2)$

Luego:

$P(N) = P(U_1).P(N/U_1) + P(U_2).P(N/U_2)$

$= 1/2 \,.\, 4/11 + 1/2 \,.\, 3/8 = 65/176 = 0.369$

valor que será la probabilidad de que ocurra (o de ocurrencia) del evento, extraer una bola negra.

9.4.3. Teorema de Bayes.

Veamos como se generaliza esa propiedad vista en el ejemplo, es decir, como calcular la probabilidad de un resultado que puede obtenerse por varias vías.

Se dice que tenemos una partición del espacio muestral S en n subconjuntos A_1 , A_2 , A_3 , ..., A_n , si:

a) la intersección de cualquier par de subconjuntos es vacía

b) ningún subconjunto es vació

c) la unión de los n subconjuntos da lugar al espació muestral, o sea:

A_1 o A_2 o A_3 o ....o A_n = S

En este caso cualquier evento E puede ser expresado entonces de la forma siguiente:

E = E y S = (E y A_1) o (E y A_2) o .....o (E y A_n)

Es fácil darse cuenta de que los eventos (E y A_i) son mutuamente excluyentes, ya que los A_i lo son, por tanto se tiene que: P(E) = P(E y A_1) + P(E y A_2) + ...+ P(E y A_n)

Si remplazamos cada sumando de la forma P(E y A_i) por su expresión equivalente que resulta de despejarla de la fórmula de la probabilidad condicional, entonces,

(P(E y A_i) = P(A_i).P(E / A_i)) y así obtenemos que:

P(E) = P(A_1).P(E /A_1) + P(A_2).P(E /A_2) +....+ P(A_n).P(E/A_n) Expresión que se conoce como, **formula de la probabilidad total**.

Una vez establecida la expresión anterior, se puede desarrollar el teorema de Bayes, el cual esta vinculado al cálculo de la probabilidad condicional , P(A_i /E) para cualquier evento A_i que se considere.

Esta probabilidad, P(A_i /E), puede expresarse por, P(A_i y E)/ P(E), según la definición condicional de probabilidad, pero a su vez puede plantearse que P(A_i y E) = P(A_i) P(E /A_i), luego sustituyendo se obtiene la expresión o formula siguiente,

P(A_i /E) = P(A_i) P(E /A_i) / [P(A_1)·P(E /A_1) + P(A_2)·P(E /A_2) +....+ P(A_n)·P(E /A_n)]

conocida en probabilidades como formula o regla de Bayes.

Veamos ahora como usar esta expresión matemática en un ejemplo concreto.

Se conoce que, entre los solicitantes de ingreso a una escuela de medicina solo son elegibles para entrar el 80% y que el 20% restante no lo es. Para ayudar en el proceso de selección de los candidatos se aplica un Test de admisión que esta diseñado para que un candidato elegible lo pase el 90% de las veces y un candidato no elegible en el 30% de los casos. Supongamos que un solicitante pasa el Test de admisión, ¿cual es la probabilidad de que el mismo esté entre los elegibles?.

Si denotamos por:

A_1 = solicitante elegible y A_2 = solicitante no elegible, tenemos dos eventos disjuntos (o excluyentes) cuya unión es el espacio muestral (todos los solicitantes), además,

P (A_1) = 0.8 y P (A_2) = 0.2

Si E es el evento, candidato que pasa el test de admisiones tiene que:

$P(E/A_1) = 0.9$ y $P(E/A_2) = 0.3$

Al final del enunciado, se nos pide calcular $P(A_1/E)$, luego por Bayes se tiene que

$P(A_1/E) = P(A_1) \cdot P(E/A_1) / (P(A_1) \cdot P(E/A_1) + P(A_2) \cdot P(E/A_2)$

$= (0.9).(0.8) / [(0.9).(0.8) + (0.3).(0.2)] = 0.72 / (0.72 + 0.06)$

$= 0.72 / 0.78 = 0.923$, o sea que: $P(A_1/E) = 0.923$.

Eso quiere decir que menos del 8% de los solicitantes que pasan el test son inelegibles, por lo que se puede considerar que dicho test es un instrumento razonablemente efectivo. A partir de este teorema es que se definen las probabilidades **a priori** y las probabilidades **a osteriori**. Veamos en que consisten estos conceptos a partir del mismo ejemplo del test de admisión en la escuela de medicina. La probabilidades $P(A_i)$ son las probabilidades de que un solicitante pertenezca a un grupo determinado (elegibles o no elegibles), antes de someterse al mismo, como es lógico. Estas son las probabilidades **a priori**.

La probabilidades $P(A_i /E)$ son las probabilidades de que un solicitante pertenezca a un grupo en particular, dado que ya paso el Test de Admisión. Estas son las probabilidades **a posteriori**. Solo la fórmula de Bayes nos proporciona una técnica para el cálculo de probabilidades a posteriori.

En la actualidad están muy en boga los llamados **Métodos Bayesianos**, que no son más que técnicas de Estadística Inferencial desarrolladas a partir de esta regla. Nosotros no vamos a estudiar ese tipo de métodos, pero si creemos necesario mencionarlos por la difusión que han alcanzado.

9.5 MODELO TEÓRICO DE DISTRIBUCIÓN O LEY DE UNA VARIABLE ALEATORIA.

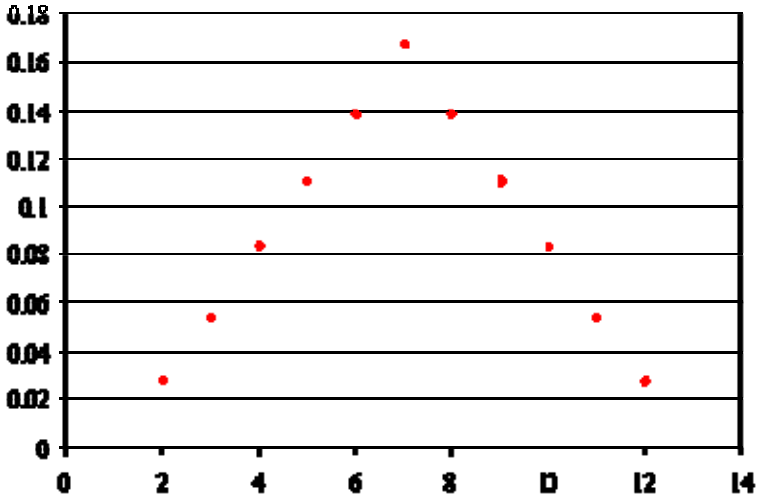
Las variables discretas no distribuyen la probabilidad total entre una cantidad innumerable de valores, sino que la concentran en determinados puntos. La función que en las variables discretas describe la probabilidad que le corresponde a cada valor de la variable, se nombra **función de cuantía** y se representa gráficamente con un diagrama de líneas. Por ejemplo, la función de distribución de probabilidades de la tirada de dos dados, si el resultado que nos interesa es la suma de las caras, sería:

Valor de x	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
P(x)	**1/36**	**2/36**	**3/36**	**4/36**	**5/36**	**6/36**	**5/36**	**4/36**	**3/36**	**2/36**	**1/36**

Note que es algo similar a una distribución de frecuencias empíricas, solo que en lugar de las frecuencias relativas observadas están las probabilidades.

El diagrama de líneas correspondiente se construye en un sistema de ejes coordenados donde sobre cada valor posible de la variable (en este caso suma de las caras de los dados) se plotea el valor de probabilidad correspondiente. Quedaría más o menos como se representa en la fig 9.2:

Fig 9.2



En una distribución de frecuencias empíricas de una variable aleatoria continua, los datos obtenidos pueden ser considerados como que pertenecen a una población grande, en la cual están disponibles muchas observaciones (en teoría, infinitas). Esto hace posible, al confeccionar un histograma o polígono de frecuencias, escoger intervalos de clase muy pequeños y que, no obstante, tengan números medibles de observaciones que puedan clasificar dentro de cada uno de ellos. Se podría esperar entonces que el polígono de frecuencias, para una población grande, estuviese dividido en pequeños segmentos lineales que pueden aproximarse por curvas suaves, las que llamamos distribuciones de frecuencias teóricas o función de densidad de probabilidades. Veámoslo gráficamente, mediante las figs 9.3, 9.4, 9.5 y 9.6:

Fig 9.3

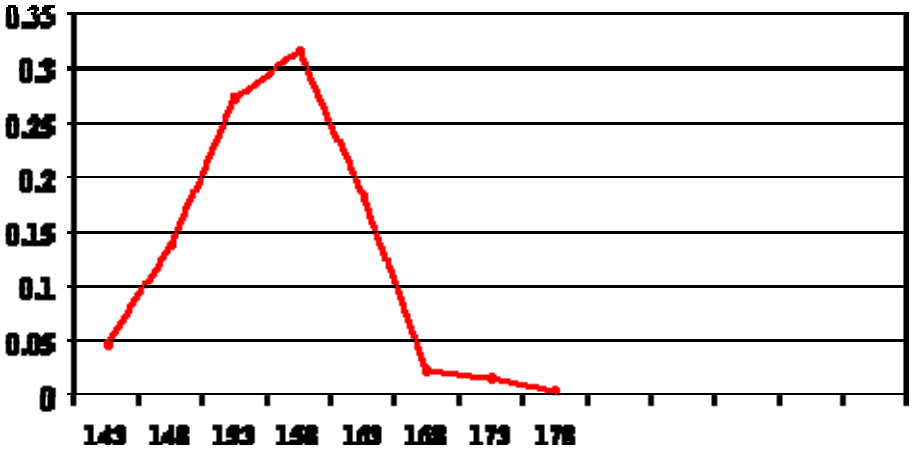


Fig 9.4


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
142 145 148 151 154 157 160 163 166 169 172 175 178

Fig 9.5

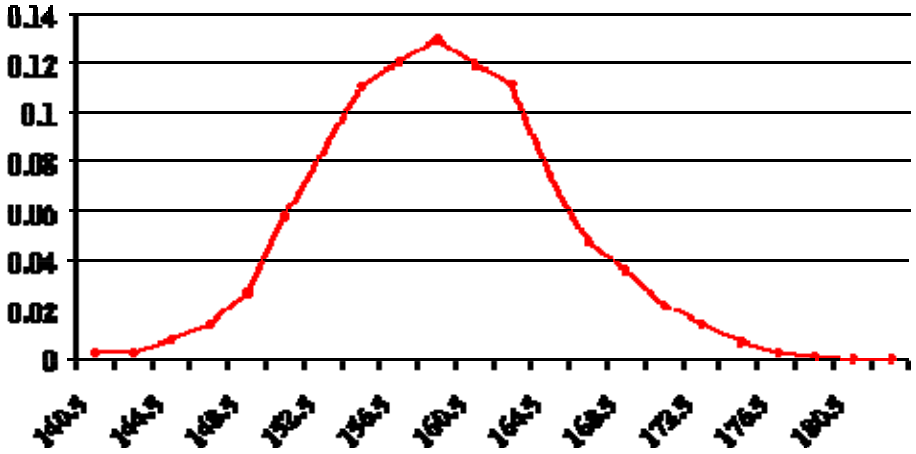
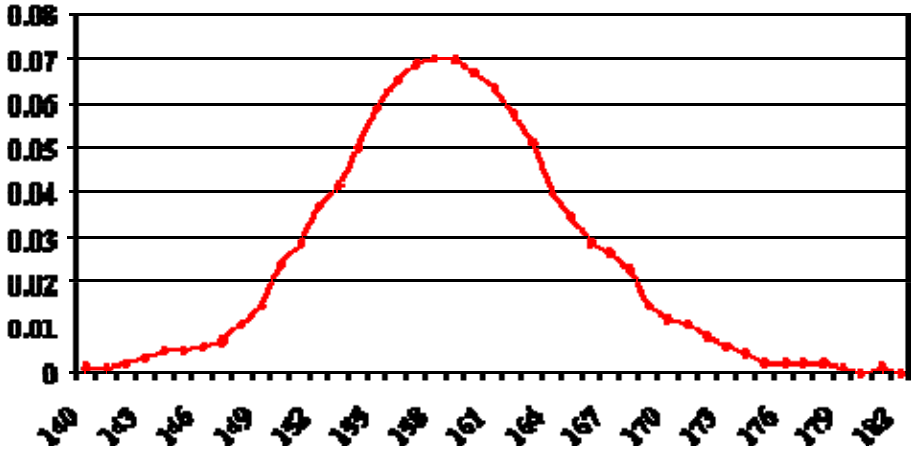

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0

Fig 9.6


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

Esta curva suave, **límite del polígono de frecuencias**, es característica para cada variable en particular y su existencia es el resultado de la acción de la ley de regularidad estadística.

Si analizamos un intervalo de clase en particular, digamos el intervalo (a_k, a_{k+1}), de acuerdo con la ley mencionada tendremos que la frecuencia relativa de ocurrencia de un valor de la variable aleatoria dentro de esos límites se irá estabilizando alrededor de un valor constante, único, en la medida en que aumenta el tamaño muestral, lo que provocará que la curva a la que se aproximan los sucesivos polígonos de frecuencia también sea única y característica de dicha variable aleatoria. Dicha curva se denomina **modelo teórico de distribución de probabilidades de la variable aleatoria** y, generalmente tiene una expresión matemática en la que intervienen, además de la variable en cuestión, algunas constantes numéricas y otros valores conocidos como **parámetros**, que no son mas que **cantidades numéricas que al asignársele valores particulares no modifican el tipo de distribución sino su forma y ubicación en los ejes coordenados**. También se les conoce como parámetros de las distribuciones o poblacionales.

En teoría, el número de elementos de una población se considera infinito, ya que tanto el concepto de probabilidad como el de ley de una variable aleatoria están asociados al concepto de límite cuando el tamaño muestral (n) tiende a infinito, pero en la realidad esto no es así: ni el número de niños que nace, ni la cantidad de sujetos adultos que habitan un país son infinitos, y las poblaciones que nos interesa analizar por medio de la Estadística son como estas de los ejemplos. Nos basta con que el número de elementos sea elevado. Para los casos en que el tamaño de la población objeto de estudio es relativamente pequeño también existen técnicas y métodos estadísticos de análisis, pero los mismos no se tratarán en este libro.

Como recordará, al explicar como se confecciona un histograma, se dijo que el área de los rectángulos es proporcional a la frecuencia que ellos representan. Si al construir el histograma (o polígono de frecuencias) usamos las frecuencias relativas (o probabilidades empíricas) tendremos que el área total del histograma (o bajo el polígono de frecuencias), que no es más que la suma de las áreas de los rectángulos, es decir, es igual a la unidad.

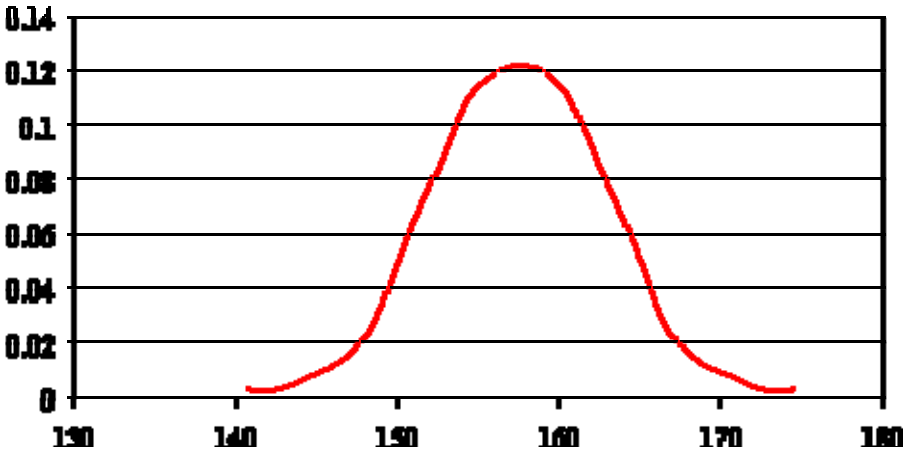
Cuando aumenta indefinidamente el número de observaciones, al irse aproximando cada vez más el histograma (o el polígono de frecuencias) a la curva ideal, el área del rectángulo correspondiente a un intervalo de clase dado se aproximará cada vez más a la frecuencia teórica correspondiente (probabilidad de que el valor de una observación de la variable se encuentre en dicho intervalo), es decir, al área bajo en la curva en ese mismo intervalo. Luego, el área bajo la curva en el intervalo (a, b) es la probabilidad de que la variable aleatoria tome valores entre a y b, y el área total bajo la curva también es la unidad.

En realidad el calculo de las probabilidades es una suma cuando se tiene un número contable de resultados posibles, porque estamos frente a una variable cualitativa o discreta, pero en el caso de una variable continua cuyos valores posibles están dentro del campo de los números reales, los resultados posibles son infinitos y se usa entonces el equivalente o límite de la suma, es decir, la integración.

Las curvas de frecuencias teóricas pueden tomar ciertas formas características, como las que se indican a continuación:

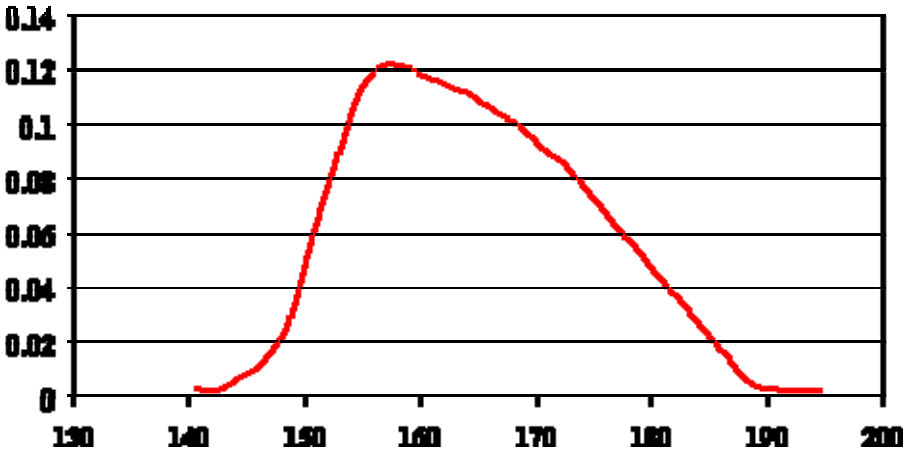
a) curvas simétricas y acampanadas (ver fig 9.7) que se caracterizan por el hecho de que las observaciones equidistantes del máximo central tienen la misma frecuencia.

Fig 9.7



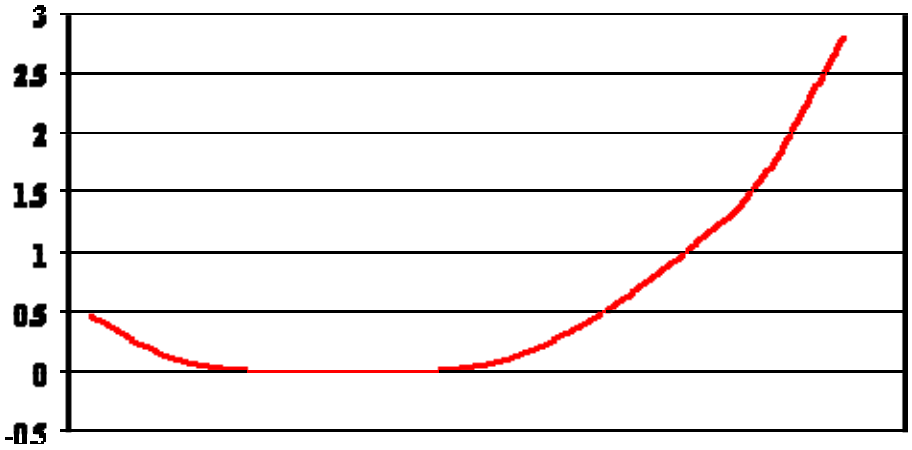
b) curvas de frecuencia moderadamente asimétricas o asimétricas (ver fig 9.8), donde la cola de las mismas a un lado del máximo central es más larga que al otro. Si la cola más larga está a la derecha, se dice que la curva es asimétrica a la derecha o que tiene una asimetría positiva. Si ocurre lo contrario, se dice que la curva es asimétrica a la izquierda o que tiene asimetría negativa.

Fig 9.8



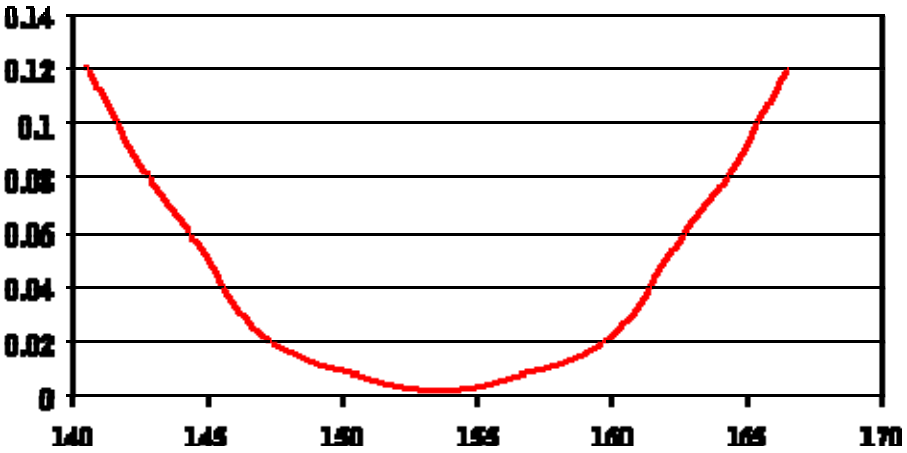
c) curvas en forma de J (o de J invertida, ver fig 9.9), en que la frecuencia máxima se encuentra en un extremo.

Fig 9.9



d) curvas de frecuencias en forma de U (ver fig 9.10), que tienen frecuencias máximas en ambos extremos.

Fig 9.10



c) curvas de frecuencia multimodales (ver fig 9.11), que tiene varios máximos.

Fig 9.11

Se conoce una serie relativamente grande de distribuciones o modelos teóricos de distribución de variables aleatorias, los que han sido muy bien estudiados y se aplican en muchas situaciones prácticas.

Ante un problema real particular debemos seleccionar el modelo que más se adecua el comportamiento de nuestra variable. Pero, ¿como saber si el **ajuste es realmente bueno**?. La selección del modelo teórico que más se ajusta a la variable en estudio implica la realización de una serie de cálculos, e incluso gráficos, que no son de objeto de estudio en este libro.

Ahora pasaremos a describir algunos de los modelos teóricos de distribución básicos en el desarrollo de las técnicas y métodos estadísticos. Solo han sido seleccionados cuatro de ellos por ser los que se relacionan directamente con los temas que se desarrollarán más adelante.

9.6 VARIABLE BERNOULLI Y LA DISTRIBUCIÓN BINOMIAL.

El modelo de probabilidad binomial es quizás uno de los modelos más sencillos y de amplio uso que existe. Responde a una secuencia de ensayos finita, que consiste en repetir n veces, una experiencia simple con solo dos resultados posibles, los que se designan habitualmente como éxito (E) y fracaso (F), para su concepción teórica se asume que el resultado de un ensayo o repetición no afecta el valor resultante de ningún otro ensayo, es decir, que los resultados de las distintas repeticiones son independientes y que la probabilidad de obtener éxito, o sea P(E) no cambia de un ensayo a otro, es decir, se mantiene constante. Este tipo de experiencia tan simple puede ser explicada (o caracterizada) completamente en términos de solamente dos números: la probabilidad de éxito P (E) y el valor de n o numero de repeticiones que se efectúan.

Como los eventos, éxito(E) y fracaso(F), son excluyentes y exhaustivos, ya que se trata de experiencias en que los resultados se pueden agrupar en estas dos categorías solamente , **la probabilidad de éxito**, **P (E)** si se denota por **p**, hace que **P(F) o probabilidad de fracaso** sea igual a **1 – p**, probabilidad que usualmente se designa por **q**. Es decir que dentro del contexto de la distribución binomial identificaremos **P(E)** con **p** y **P(F)** con **q = 1 – p**. Las ensayos aleatorios de este tipo o ensayos dicotómicos (de solo dos resultados posibles) recibe el nombre de **ensayos o pruebas Bernoulli**. Muchas situaciones en el mundo real se modelan según este tipo de variables. Por ejemplo: el sexo de un recién nacido; el contraer (o padecer) una enfermedad dada; si un producto cumple las especificaciones de calidad; etc.

Se llama **experiencia o ensayo binomial**, a aquel que consiste en observar los resultados de n observaciones (o repeticiones) independientes de un ensayo Bernoulli, bajo la condición de que, p (probabilidad de éxito) se mantenga constante durante todo el proceso de observación. Se insiste en que se observe que, un ensayo binomial, debe ser comprendido como una secuencia de n resultados como un todo.

En toda situación de esta índole, interesa observar el valor una variable aleatoria que da respuesta a la pregunta siguiente, ¿ **cuantos éxitos ocurren en la secuencia de n repeticiones que se ejecutan**?. Si denotamos este numero por **X** (equis mayúscula), no es difícil darse cuenta de que el valor de **X** puede ser cualquier valor entre 0 y n inclusive, ahora bien, dado el carácter aleatorio de los ensayos y su diverso comportamiento de ensayo a ensayo, es que no se puede predecir de antemano que valor tomara X en un ensayo binomial concreto, de ahí su carácter aleatorio.

En relación con la variable X, resulta de interés, no solamente conocer dentro de que rango o limites ella toma valores (0 a n), sino que, también se desea saber, ¿ **con que probabilidad (frecuencia) ocurre que la variable X sea igual a k éxitos, con k entre 0 y n** ?, el valor de probabilidad que da respuesta a la ultima pregunta es una formula de calculo que dentro de la teoría de probabilidades se conoce con el nombre función de distribución binomial.

Si denotamos por P(X = k) a la probabilidad de que se produzcan k éxitos en un ensayo binomial, entonces es conocido que:

$\mathbf{P(X = k) = C_{n,\,k} \cdot p^k \cdot (1 - p)^{n-k} = C_{n,\,k} \cdot p^k \cdot q^{n-k}}$, con k valor arbitrario entre 0 y n, donde $\mathbf{C_{n,\,k}}$ es un numero cuyo significado concreto es el de ser, **la cantidad de ensayos binomial concretos (secuencias de**

éxitos y fracasos de longitud n) que presentan exactamente k éxitos y (n – k) fracasos, también en matemáticas elementales se le conoce con el nombre, **numero de combinaciones de n objetos cuando se toman k de ellos**.

Hasta ahora no se ha indicado como calcular $C_{n,\,k}$ cuestión que se abordara ahora. Con este fin, considere que, m, es un numero entero positivo, entonces, al producto de todos los números del 1 al m, o sea al numero denotado por $1\cdot2\cdot3\cdot\ldots\cdot(m-1)\cdot m$ se le conoce como, **factorial de m**, y se denotara brevemente por **fact(m)**, por ejemplo: fact(5) = 1·2·3·4·5 = 120, fact(4) = 1·2·3·4 = 24, etc. Asumiremos que fact(0) = 1, por conveniencia cuando esto haga falta..

Note que, fact(m) se puede expresar de la manera siguiente, sea h un numero entero positivo menor que m, entonces, **fact(m)** = fact(h)·**[**(h +1)·(h + 2)·… (m – 1)·m]. , formula que facilita el trabajo con factoriales, al obviar su calculo en la determinacion de probabilidades.

Es conocido de la matemática elemental que entre las magnitudes de $\mathbf{C_{n,\,k}}$ **y fact(m)** existe un estrecho vinculo dado por: $\mathbf{C_{n,\,k} = fact(n) / (fact(k) \cdot fact(n - k)).}$

Esta relación hace que en la formula de cálculo de la ley binomial, el producto $\mathbf{C_{n,\,k}\cdot p^k \cdot (1-p)^{n-k}}$ adquiera su comprensión completa, pues ahora se indica como calcular a $C_{n,\,k}$ en el mismo, aunque en la práctica en diversas ocasiones se evita esto usando la Tabla A (ver anexos) donde aparecen los valores de $C_{n,\,k}$ para diversas combinaciones de n y k. Vale decir, que no es la única forma, pues también se logra lo mismo y mas aun, mediante el empleo de una tabla de la ley binomial. Estas dos estrategias serán ilustradas mas adelante. Por ultimo el nombre binomial de esta distribución proviene de que cada uno de los valores, $\mathbf{C_{n,\,k}\cdot p^k \cdot (1-p)^{n-k}}$ es igual a uno de los términos del desarrollo de $(p + q)^n$, que se hayan por medio de la formula del binomio de Newton.

Al escribir sobre las distribuciones de probabilidad, estas se representan mediante símbolos asociados a sus nombres. En la distribución o ley binomial esto se hace mediante la expresión, **b(n, k; p)** que sustituye al producto $\mathbf{C_{n,\,k}\cdot p^k \cdot (1-p)^{n-k}.}$

De este modo, la frase, **X es una variable con distribución binomial, con valores n y k dados**, admite que se refleje al escribir que, **P (X = k) = b(n, k; p)** si es lo que interesa, y si lo que se quiere es abreviar la misma**,** se hace mediante, **X ~ b(n, k; p)**, donde en esta combinación de símbolos, ~ debe comprenderse como un sustituto de, **es una variable con distribución** o **se distribuye según una ley** según sea el caso.

Se llaman **parámetros de la distribución binomial, a los valores de n y p**, ya que conociéndolos se puede singularizar a esta ley, y obtener todas las probabilidades de éxito deseadas.

Veamos un ejemplo: Sea un tirador que realiza 12 disparos, con probabilidad 1/3 de dar en el blanco cada vez.. ¿cual es la probabilidad de que acierte de 6 a 8 veces?.

La respuesta correcta se obtiene, por medio de la distribución binomial, ya que lo planteado se reduce al esquema de un **ensayo binomial con 12 repeticiones** y **evento éxito, dar en el blanco**, con **P (Éxito) = 1/3**. Note que el evento éxito, hay que definirlo en consonancia con el problema, en este caso es, dar en el blanco.

La probabilidad que se pide calcular tiene en cuenta al evento siguiente: **Acertar 6 ó 7 u 8 veces,** expresable mediante X (# de aciertos al blanco de 12 disparos) como, **(X = 6) o (X = 7) o (X = 8)**.

Se tiene que, P (**Acertar 6 ó 7 u 8 veces**) = P [**(X = 6) o (X = 7) o (X = 8)**]

= P[(X = 6)] + P [(X = 7)] + P [(X = 8)].

$P(X = 6) = C_{12,\,6} \cdot (1/3)^6 \cdot (2/3)^{12-6}$,donde $C_{12,\,6}$ se halla por medio de $C_{12,\,6}$ = fact(12) / (fact(6) · fact(6)) = fact(6)·7·8·9·10·11·12 / $fact(6)^2$.= 7·8·9·10·11·12 / 1·2·3·4·5·6 = 7·8·9·11 / 1·2·3 = 7·4·3·11 = 924. ya que 10·12 = 4·5·6 = 120,

En consecuencia, $P(X = 6) = 924 \cdot 2^6 / 3^6 \cdot 3^6$, donde $2^6 = 64$ y $3^6 = 729$.

Por lo tanto, $P(X = 6) = 924 \cdot 64 / 729^2 = 59136 / 531441 = 0.111$.

De manera semejante, se puede proceder con las restantes expresiones, tomando en cuenta claro esta los cambios en cada caso, así se obtienen los valores siguientes:

$P(X = 7) = C_{12,\,7} \cdot (1/3)^7 \cdot (2/3)^5 = 0.048$ y $P(X = 8) = C_{12,\,8} \cdot (1/3)^8 \cdot (2/3)^4 = 0.015$

Sustituyendo, se tiene: P(**Acertar 6 ó 7 u 8 veces**) = P(X = 6) + P(X = 7)+P(X = 8)

= 0.111 + 0.048 + 0.015 = 0.174.

Es decir, aproximadamente en el 17 % de las veces que este tirador realiza sesiones de 12 disparos, acierta entre 6 a 8 veces.

Como se ha podido observar, el uso de la operación fact(), para la obtención de los coeficientes de la formula de la distribución binomial implican un cálculo engorroso y en muchas ocasiones cuando n es relativamente grande, largo también. Para hacer mas fácil el trabajo del calculo de probabilidades con la distribución binomial, se dijo que, existen tablas que dan directamente el valor de $C_{n,\,k}$. Así la Tabla A (ver anexos) da directamente que: $C_{12,\,6} = 924$, para la pareja n = 12, k = 6; $C_{12,\,7} = 792$, con n = 12, k = 8 y $C_{12,\,8} = 495$, para n = 12, k = 8, sin necesidad de trabajar con factorial alguno, pero con la obligación de hallar el valor del producto: $C_{n,\,k} \cdot p^k \cdot (1-p)^{n-k}$, en cada caso.

Si se hace uso de la Tabla B (tabla de la distribución binomial) se busca dentro del cuerpo de la tabla **la cabecera, n = 12**, y dentro del grupo de segmentos de columnas de números, se busca **la cabecera** siguiente correspondiente a, **p = 1/3 ≈ 0.33** la que ubica un segmento de 11 valores de probabilidad distintos de cero (deben ser 13, pero los dos últimos valores son iguales a 0, por ello no se toman en cuenta), estos son :

1.0000 (k = 0) 0.1711 (k = 6)

0.9918 (k = 1) 0.0632 (k = 7)

0.9435 (k = 2) 0.0176 (k = 8)

0.8124 (k = 3) 0.0036 (k = 9)

0.5973 (k = 4) 0.0005 (k = 10)

0.3590 (k = 5)

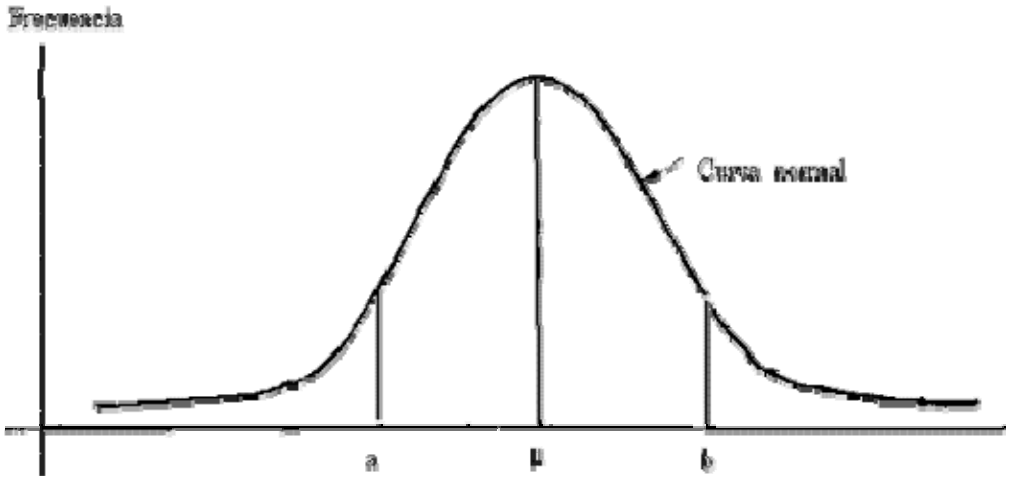
De acuerdo con la explicación que aparece en la tabla B, la probabilidad buscada en el problema viene dada por: P(**Acertar 6 ó 7 u 8 veces**) = P (X ≥ 6) – P (X ≥ 9) = 0.1711 – 0.0036 = 0.1675.

Nótese que el resultado de este cálculo es aproximado al anterior, es menor, debido a la aproximación usada para p = 1/3, como igual a 0.33, cuando en realidad el valor de p es 0.3333...Un decimal con infinitas cifras un poco mayor que 0.33.

9.7 MODELO DE LA DISTRIBUCIÓN NORMAL. PROPIEDADES.

Quizás el modelo teórico más relevante dentro de la estadística clásica, sea el conocido como **distribución normal o de Gauss**. Su nombre se debe a los trabajos de dicho matemático en relación con el análisis de los errores de medición en los cálculos astronómicos los que dieron a conocer esta distribución teórica, que a veces es llamado también **campana de Gauss**, porque la forma geométrica de la función matemática que la describe tiene forma simétrica y acampanada con respecto a un eje central. Su representación gráfica, se aprecia en la fig 9.12:

Fig 9.12(G-Normal)



Este modelo es sumamente importante por varias razones, entre las cuales tenemos:

- Toda la Estadística Inferencial clásica, prácticamente, ha sido desarrollada a partir de la hipótesis de que las variables en estudio se ajustan o siguen dicho modelo.

- Una gran cantidad de variables, especialmente biológicas, se rigen por este modelo.

Sin embargo, hay que tener en cuenta que no siempre se puede emplear como modelo de distribución para toda variable. En la práctica se debe indagar que modelo teórico hay que usar para cada una en particular, ya que de ello depende el tipo de técnicas estadísticas que se va a utilizar.

Las distribuciones de frecuencias teóricas, generalmente, se expresan por medio de funciones. Una distribución de frecuencias teóricas es del tipo normal si la función que la representa responde a la ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty, -\infty < \mu < +\infty, \sigma > 0$$

donde: p y e son constantes conocidas m es el punto en que el grafico de f(x) alcanza su valor máximo, $\frac{1}{\sigma\sqrt{2\pi}}$ s es el valor que altera o cambia el máximo de la curva en forma inversa a su magnitud, manifestándose esto en que la curva gana en altura (se hace mas apuntada) o baja de altura (se hace mas achatada), para valores pequeños o grandes de σ respectivamente.

Los valores de **μ** y **σ** representan constantes con valores específicos para cada distribución, estos permiten caracterizar o individualizar a la misma, por esta particularidad a dichos valores se les conoce como **parámetros de la distribución normal** de modo semejante a como que **n** y **p** significan lo mismo respecto al **modelo de la distribución binomial**.

El hecho de que, una variable aleatoria X tenga una distribución normal con parámetros m y s, se denotara en la escritura por: **X ~ N(m,s)**, denominándose a la misma como **variable aleatoria normal**.

Desde el punto de vista probabilístico o si se desea estadístico, se tiene que los valores **m** y**s**, tienen un significado preciso, el primero de ello representa el **valor medio teorico, poblacional o esperado** de la variable X, mientras que el segundo simboliza el **valor del desvio estandar o dispersión poblacional respecto del valor medio m**. Se aclara que si una variable presenta un comportamiento distinto del normal, esto no quiere decir que por ello sea **anormal**, sino, simplemente, que la ecuación que describe su curva de frecuencias es de otro tipo.

En la expresión de la función f(x) que caracteriza a la curva del modelo normal hay dos elementos que cambian de un caso a otro, m y s. Veamos como se comporta el gráfico de este modelo mediante ejemplos:

Distribución de la talla según el sexo:

a) del sexo masculino.

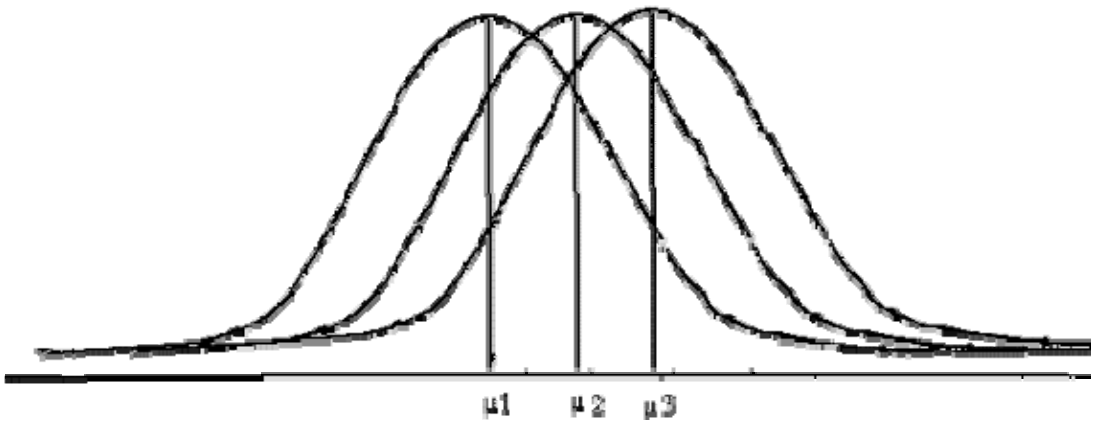
b) b) del sexo femenino.

Observemos gráficamente como influyen los valores de estos parámetros en la forma de la curva.

a) Consideremos tres valores de μ tales que: $\mu_1 < \mu_2 < \mu_3$, de tal modo que las tres distribuciones tengan el mismo valor de σ, o sea, $s_1 = s_2 = s_3$.

Si tenemos dispersiones o desvíos estándar poblacionales iguales los máximos de las campanas lo serán también, y como estarán centradas en μ_1, μ_2 y μ_3 respectivamente, se encontraran colocadas sobre el eje que representa a la variable aleatoria en esa forma escalonada, como se muestra a continuación en la figura 9.13..

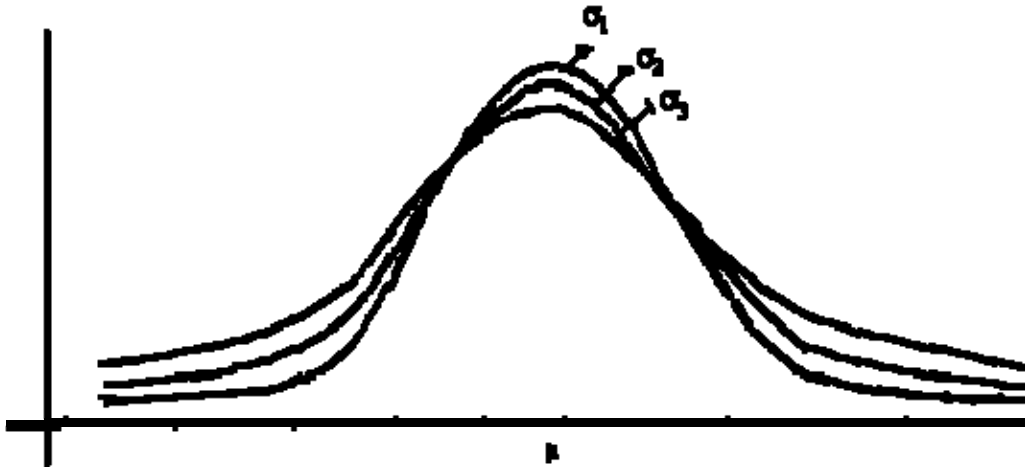
Fig 9.13



b) Consideremos tres distribuciones normales con igual media pero de tal modo que ellas tengan diferentes valores de σ, o sea, $s_1 < s_2 < s_3$.

Como tienen medias iguales, las tres están centradas en el mismo valor. Las dispersiones desiguales implican que las alturas máximas de las curvas que ellas representan, lo sean también, estableciéndose la siguiente relación: a menor valor de σ mayor altura. Lo expresado se muestra en la figura 9.14:

Fig 9.14



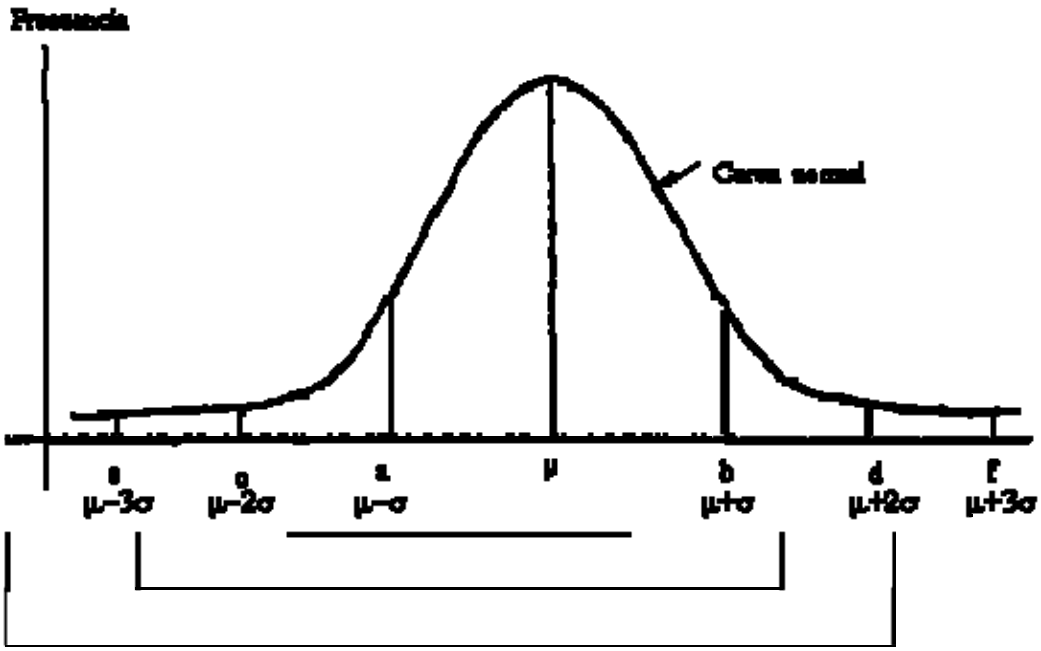
Una propiedad muy importante de este modelo y que se dice le caracteriza, es la siguiente: a) $P (\mu - s < X < \mu + s) = 0.6827$

b) $P (\mu - 2s < X < \mu + 2s) = 0.9545$

c) P ($\mu - 3s < X < \mu + 3s$) = 0.9973

Esta propiedad tiene un amplio uso en la práctica. La misma se ilustra mediante la figura 9.15 siguiente

Fig 9.15



Veamos algunos ejemplos:

1. En el conocimiento de como se comporta una característica (variable aleatoria) en los elementos de una población.

   Si sabemos que la talla adulta sigue una distribución normal y se nos dice que la media de talla del hombre en Cuba es de 167.3 cm con una desviación estándar de 7.8 cm podemos inferir que:

   - aproximadamente el 68.27% de los hombres cubanos tienen estaturas entre los 159.5 (167.3 – 7.8) cm y los 175.1 (167.3 + 7.8) cm.
   - aproximadamente el 95.45% de los hombres cubanos tienen estaturas entre los 151.7 (167.3 – 2(7.8)) cm y los 182,9 (167.3 + 2(7.8)) cm.
   - aproximadamente el 99.73% de los hombres cubanos tienen estaturas entre los 143.9 (167.3 – 3(7.8)) cm y los 190.7 (167.3 + 3(7.8) cm.
   - o lo que es lo mismo, la probabilidad de que al seleccionar aleatoriamente un hombre en la población este tenga una estatura entre:

   - 159.5 cm y 175.1 cm es de 0.6827

- 151.7 cm y 182.9 cm es de 0.9545

- 143.9 cm y 190.7 cm es de 0.9973

2. En la definición de rangos de normalidad de una variable.

   Si conocemos que una variable aleatoria sigue una distribución normal con media m y desviación estándar s, se pueden definir como límites del rango de normalidad (en su acepción corriente) los valores $\mu - 2s$ y $\mu + 2s$, por ejemplo, si sabemos que aproximadamente solo el 5 % de la población puede presentar valores 'anormales' o patológicos. Es decir, si el especialista conoce, ya sea por la experiencia o a través de resultados de investigaciones, tanto los parámetros de la distribución normal (que es el modelo teórico de su variable) como la proporción de sujetos enfermos en la población, puede encontrar un valor por el que debe multiplicar a la desviación estándar, para restar o sumar a la media, que le proporcione los límites de normalidad.

   En muchos casos se dan varios límites, que corresponden a decisiones diferentes: los límites de normalidad, que son dos valores entre los cuales debe encontrarse el resultado para un sujeto sano, y los patológicos, que pueden ser uno o dos valores que nos indican que resultados menores que el valor más pequeño o mayores que el valor más grande son patológicos. Los resultados que se encuentren entre un límite de normalidad y un límite patológico se reportan como dudosos o no conclusivos.

3. En la detección de errores accidentales o de casos atípicos.

Cuando se recoge información, siempre se esta sujeto a que se produzcan errores accidentales (de anotación, de copia, etc.). En estos casos siempre se realiza un chequeo de los datos recogidos con el objetivo de detectar los posibles errores. Generalmente se identifican los valores que resultan o muy grandes o muy pequeños con respecto al comportamiento promedio de la variable en el grupo y el especialista debe decidir entonces cuáles de esos datos deben ser eliminados del procesamiento partiendo del análisis conjunto de toda la información recogida sobre cada individuo en particular. Por ejemplo, en el proceso de validación de los pesos de sujetos de 11 años de edad, un peso de 75.2 Kg. se consideraría un posible error; solo del análisis de todos los datos antropométricos referentes a dicho sujeto podemos comprobar si es efectivamente un peso erróneo ó es que se ha incluido en la muestra un sujeto afectado por una obesidad masiva.

4. En el proceso de calificación de tests o pruebas de aptitudes.

   En muchas ocasiones se usan los intervalos de la normal con el sentido siguiente:

   Si el resultado es un valor que pertenece al intervalo:

   a) $(\mu - s\ ;\ \mu + s)$ se considera aprobado

   b) $(0\ ,\ \mu - s)$ se considera desaprobado

c) $(\mu + s, +\infty)$ se considera apto

9.8 DISTRIBUCIÓN NORMAL ESTÁNDAR.

¿Cual es la utilidad de conocer que una variable tiene una distribución normal?.

Supongamos que deseamos conocer la probabilidad de que un niño de 7 años tenga 115 cm o menos de estatura. Para ello deberíamos tener el número de niños con estatura de 115 cm o menos y conocer el total de niños de 7 años. En general, para obtener la probabilidad de que un niño de cierta edad tenga una estatura entre dos valores dados, digamos a y b, se debe tomar la cantidad de niños de esa edad con estaturas entre esos dos valores y dividirla entre el total de niños de la edad en cuestión.

Al analizar la situación a partir de la distribución de frecuencias teóricas, hallar la probabilidad de que un niño tenga una estatura que oscile entre a cm y b cm, equivaldría a calcular el área bajo la curva de frecuencias teórica que media entre esos valores. Esto se expresaría como la integral de la función que describe la curva entre los puntos a y b. Esto es:

$p(a<x<b)= \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

Afortunadamente, no es necesario resolver esta integral cada vez que se necesite calcular una probabilidad para una variable aleatoria normal. ¿Como es que se calculan entonces, estas probabilidades?. Para ello tenemos que trabajar con una distribución normal en particular: la distribución normal estándar.

Cuando m = 0 y s = 1, la expresión de la función que caracteriza a este modelo de distribución normal se simplifica, reduciéndose a:

$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $-\infty < x < +\infty$

En este caso se dice que la variable aleatoria tiene o sigue una distribución normal estándar. A esta variable normal se le acostumbra a denotar con la letra Z , y para expresar que ella tiene una distribución normal estándar se escribe: $Z \sim N(0; 1)$.

Para trabajar el calculo de probabilidades asociadas al modelo normal se utiliza el modelo normal estándar mediante la tabla C que aparece en los anexos Esta tabla ofrece los valores de las probabilidades (como áreas bajo la curva normal estándar desde 0 hasta un valor positivo de interés). Tenga en cuenta que, por ser la curva normal simétrica respecto al valor de la media, que en el caso de la estándar es cero, el área bajo la curva a la izquierda de cero es 0.5, la mitad del área total bajo la curva.

Por ejemplo la probabilidad (área) entre 0 y 1.21 según la tabla C es 0.3869. De este modo si el resultado de un cálculo de probabilidades conduce a tener que hallar la probabilidad de que la variable Z sea menor que 1.21, entonces a 0.5 se le debe sumar el valor 0.3869, para así poder hallar el resultado final 0.8869.

Entre las variables normales en general y la variable normal estándar existe una relación muy útil: se puede comprobar que el área bajo una curva normal con media μ y desviación estándar s entre los puntos a y b es igual al área bajo la curva normal estándar entre los puntos $(a - \mu)/s$ y $(b - \mu)/s$. A este procedimiento de restar la media y dividir ese resultado por la desviación estándar es lo que se conoce como procedimiento de estandarización.

Es conocido en teoría de probabilidades que siendo X una variable aleatoria cualquiera, entonces la variable estandarizada de X, o sea, $Z = (X - \mu) / \sigma$ tiene una media de valor 0 y una desviación estándar con valor 1.

Mostraremos el procedimiento que se debe seguir para el cálculo de las probabilidades (áreas) mediante algunos ejemplos.

1. Suponga que la estatura (talla) de niños de 7 años es una variable aleatoria normal con media de 119 cm y desviación estándar de 5 cm.

a) ¿cual es la probabilidad de que un niño de 7 años mida menos de 121 cm?.

b) ¿Que probabilidad hay de que mida menos de 116 cm?

Solución:

En el inciso (a), se desea hallar la probabilidad de que la talla sea menor que 121 cm, o sea, P(Talla < 121 cm), equivalente al área bajo la curva normal estándar a la izquierda del valor estandarizado correspondiente a 121, que es $(121 - 119) / 5 = 0.4$.

Se debe hallar ahora el valor de $P(Z < 0.4)$. En la tabla se aprecia que el área de 0 a 0.4 es 0.1554, luego, la probabilidad pedida es la suma de 0.5 (probabilidad a la izquierda de 0) más 0.1554 (valor de la tabla), o sea: P(Talla < 121 cm) = 0.5 + 0.1554 = 0.6556.

En el inciso (b), se desea hallar la probabilidad de que la talla sea menor que 116 cm, o sea, P(X < 116 cm), lo que es lo mismo, calcular el área bajo la curva normal estándar a la izquierda del valor estandarizado correspondiente a 116 cm, que es – 0.6. El signo negativo indica que, no podemos buscar directamente en la tabla, pero usando la simetría, el valor pedido es igual a la probabilidad bajo la curva normal estándar a la derecha de 0.6, es decir:

$$P(X < 116 \text{ cm}) = P(Z < (116 - 119) / 5) = P(Z < -0.6) = P(Z > 0.6)$$

$$= 0.5 - P(0 < Z < 0.6) = 0.5 - 0.2257 = 0.2743.$$

donde 0.5 representa el área bajo la curva normal estándar a la derecha de 0.

Mediante procederes análogos es posible obtener otros valores probabilidad, utilizando adecuadamente la tabla C y la propiedad de simetría. Una buena ayuda para estos fines es confeccionar un gráfico donde se visualicen o ilustren las áreas que intervienen en el cálculo.

Se debe señalar en los ejemplos anteriores, que las probabilidades pedidas en ellos, se pueden escribir de manera general en forma de ecuación como sigue:

$P(X < K) =$ Valor de probabilidad (p)

Donde :

§ $X \sim N(\mu, \sigma^2)$ con μ, σ valores conocidos.

§ K es un valor conocido

§ Valor de probabilidad (p), es un numero desconocido que se pide hallar en cada caso.

Además del empleo señalado, existe otro modo de usarse la distribución normal muy vinculada a la ecuación anterior, en el cual conociendo el valor de probabilidad (p) como dato, se pide entonces determinar el valor de K que le corresponde.

Por ejemplo, ¿cual será el valor de la talla K por debajo de la cual se encuentre el 95% de los niños cubanos de 7 años de edad ?.¿Como se calcularía?.

En la tabla de la normal estándar se ve que desde 0 hasta 1.65 se encuentra, aproximadamente el 45% del área, y por tanto a la izquierda de 1.65 el 95 %. Ahora bien ¿que valor después de restarle 119 y dividirlo entre 5 da 1.65?. Para ello basta resolver la ecuación: $(X – 119)/ 5 = 119$, la cual se resuelve a continuación:

$X = 119 + (5 \cdot 1.65)$

$X = 127.25 \approx 127.3$ cm

En consecuencia el 95% de los niños de 7 años mide menos de 127,3 cm, o lo que es lo mismo, 127.3 cm es el valor del 95 percentil poblacional para la talla en los niños cubanos de 7 años.

Varias investigaciones sobre desarrollo físico han demostrado que la distribución normal es un buen modelo para algunas dimensiones antropométricas como la estatura, los diámetros biilíaco, biacromial y bitrocantérico, la longitud del pie, la altura del sujeto sentado y otras que se apoyan en puntos óseos. Este conocimiento nos permite hacer el cálculo de los percentiles correspondientes, a partir de los percentiles de la distribución normal estándar.

Pero también muchas de las variables que miden desarrollo físico no tienen una distribución normal. El peso, los pliegues grasos, las circunferencias de brazo, muslo, pierna y tórax, y cualesquiera que comprenda algún componente graso, tienen curvas que presentan una cierta asimetría, es decir, una cola a un lado de la distribución. Esa asimetría, en función de las poblaciones de origen de los datos, puede resultar tanto positiva como negativa. Por ejemplo, el peso en Cuba tiene una asimetría positiva, y esa misma dimensión en un país con problemas serios de desnutrición debe presentar una asimetría negativa. ¿ Qué hacer en estos casos?

En algunas ocasiones se aplica una transformación a los valores de estas variables, (por ejemplo, la logarítmica), para obtener distribuciones no muy diferentes de la normal en los valores transformados. Este es un artificio muy utilizado cuando se desea hacer uso de los métodos de la Estadística Clásica.

9.9 APROXIMACIÓN DE LA DISTRIBUCIÓN BINOMIAL POR LA LEY NORMAL.

Los problemas de probabilidades relativos a la distribución binomial se resuelven fácilmente, cuando el número de repeticiones de la experiencia Bernoulli no es grande. Si ese número es grande, los cálculos a realizar usando las fórmulas correspondientes pueden llegar a ser muy largos y laboriosos; en consecuencia, es muy conveniente tener a mano una forma proceder, que sea de fácil aplicación , para resolver esos casos.

Tal aproximación existe y no es más que la misma distribución normal, considerando como parámetros de esta distribución, a los valores expresados por:

$$\mu = n \cdot p, \quad \sigma = n \cdot p \cdot q$$

donde: p = proporción de éxitos, q = 1 – p = proporción de fracasos

n = número de repeticiones (tamaño muestral) y p + q = 1.

La interrogante a responder ahora sería, ¿qué tamaño muestral se considera lo suficientemente grande como para poder usar esta aproximación?.

Esta pregunta tiene variadas respuestas, aquí se aplicara el criterio que plantea que:

si $n \cdot p \cdot q > 5$

entonces, a n se le debe considerar lo suficientemente grande como para aproximar la distribución binomial por medio de la normal. Haremos uso de este criterio por ser el más frecuentemente aplicado en la práctica. Veamos como emplear esa aproximación mediante el siguiente ejemplo. Calcular la probabilidad de que en 50 lanzamientos de una moneda balanceada, se obtengan 20 o más escudos en total. Solución: Este planteamiento conduce al empleo de la variable X, definida por:

X = Numero de escudos que se obtienen en los 50 lanzamientos.

La cual se conoce que tiene una distribución binomial con parámetros n = 50 y p = ½, o sea: $X \sim b(50, k; ½)$ y con valor de probabilidad pedida igual a $P(X \geq 20) = 0.9405$ según una tabla apropiada de la binomial. Este calculo seria realmente laborioso de llevar a cabo por otros medios. Como $n \cdot p \cdot q = 50 \cdot ½ \cdot ½ = 50/4 = 12.5 > 5$, entonces se puede aplicar la aproximación binomial – normal. Para ello debemos encontrar el área bajo la curva normal ajustada, centrada en $n \cdot p = 25$ y con varianza $n \cdot p \cdot q = 12.5$, que se encuentra a la derecha de 20. Para calcular esta probabilidad es necesario utilizar la estandarización y la tabla normal estándar del modo siguiente: $P(X \geq 20) = P(X \geq 19.5)$, por ser X una variable aleatoria discreta,

$= P(Z \geq (19.5 - 25)/\sqrt{12.5})$, ya que se aplica la aproximación binomial normal, $= P(Z \geq -1.56) =$ 0.94062, valor obtenido mediante el uso de la tabla normal estándar, que es una buena aproximación a la probabilidad Binomial correspondiente.

9.10 CONCEPTOS DE ESTADÍGRAFO Y DISTRIBUCIÓN MUESTRAL DE UN ESTADÍGRAFO.

Se define como estadígrafo a cualquier procedimiento sobre los resultados de una muestra de una variable aleatoria

Estos procederes generalmente se representan o expresan por medio de formulas, tal es el caso de la media aritmética , la varianza muestral y de otras valores muestrales que se encuentran detallados en el capitulo 8, aunque en ocasiones esto no sucede así, como por ejemplo ocurre con el calculo de la **moda de una muestra**.

Esta definición indica que no estamos frente a algo completamente nuevo, ya que todas las medidas de tendencia central, posición y dispersión que estudiamos antes corresponden a ella, es decir, son estadígrafos.

Por la propia definición de estadígrafo es relativamente fácil concluir que ellos a su vez son variables aleatorias, ya que el valor de estos depende de los valores muestrales, que son el resultado de observar el valor que toma una variable aleatoria en distintas ocasiones.

Como un estadígrafo es una variable aleatoria, podemos asimismo pensar que también el tiene una distribución teórica de probabilidad por medio de la cual se rige la ocurrencia de sus valores, s usual en estadística denominar a esta como la distribución muestral del estadígrafo y siempre va a depender del tamaño muestral n.

En este caso, una pregunta que nos podemos hacer es, ¿existe alguna relación entre el modelo teórico de la variable original y el modelo teórico del estadígrafo?, por ejemplo, ¿guarda alguna relación la distribución muestral de la media o la varianza muestral (estadígrafo) con el modelo de la variable original?.

Ahora pasaremos a analizar algunos casos particulares que resultan de interés para el desarrollo armónico de algunos de los temas que se trataran en próximos capitulos.

9.11 Distribución de la media muestral cuando la variable aleatoria original es normal. Error estándar de la media.

En primer lugar para que no haya dificultades de comprensión o de comunicación, se deben establecer algunas convenciones de notación. Se acostumbra denotar a la media poblacional de la variable aleatoria normal original con la letra griega μ (mu) y a la desviación estándar poblacional con la letra s (sigma).

Se usaran también esas letras griegas para los parámetros poblacionales de la distribución muestral del estadígrafo, pero añadiéndoles subíndices para diferenciarlas en su empleo de las de la población madre. Por ejemplo, si los estadígrafos $\bar{x}$ y S^2 son los de referencia, entonces

$\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$-μ_{S^2} y σ_{S^2} serán los símbolos que representaran la media y el desvió estándar poblacional $\bar{x}$ y S^2 respectivamente.

Estos símbolos y sus expresiones en términos de los valores de la población madre se muestran a continuación:

Valores de la media y varianza para los estadígrafos $\bar{x}$ y S^2

Media Varianza $\bar{x}$

$\mu_{\bar{x}} = \mu$ $\sigma^2_{\bar{x}} = \sigma^2/n$

S^2 $\mu_{S^2} = \sigma^2$ —

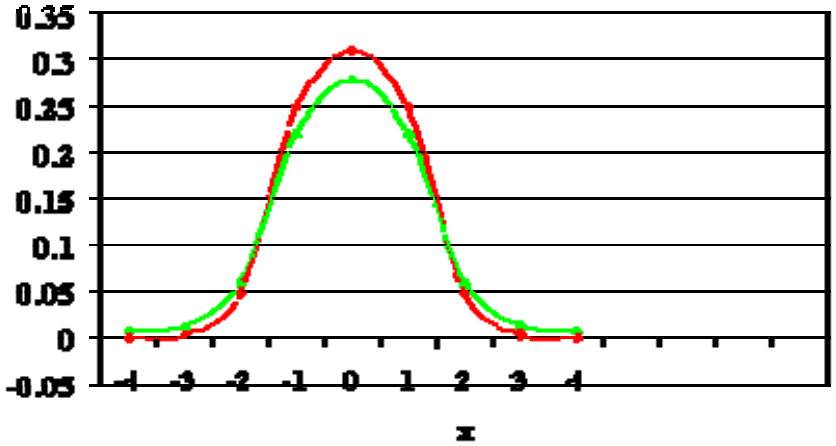
Se omite la expresión de la varianza poblacional de S^2, pues la misma depende de otros parámetros poblacionales de orden superior a los de la varianza, que no tiene importancia detallar pues su uso práctico no es de ningún interés en la estadística aplicada.

De las expresiones de $\mu_{\bar{x}}$ y $\sigma^2_{\bar{x}}$ se deduce que la distribución del estadígrafo $\bar{x}$, esta centrada en el mismo valor medio poblacional μ que la distribución madre de la variable X, mientras que su varianza poblacional es menor que la de la distribución madre en razón inversa al tamaño de la muestra. Esto ultimo significa que siempre vamos a tener que, la dispersión de los valores de $\bar{x}$ en torno de μ van a estar menos dispersos que los valores de X alrededor del mismo valor medio poblacional.

Cuando se conoce que X se distribuye normal con media μ y varianza σ^2, con representación simbólica mediante $X \sim N(\mu, \sigma^2)$, se sabe que $\bar{x} \sim N(\mu, \sigma^2/n)$, lo que equivale a afirmar que, el estadígrafo $\bar{x}$ también se distribuye normal con el valor de los parámetros poblacionales indicados anteriormente.

Lo expresado en el ultimo párrafo admite la representación grafica que se muestra a continuación:

Fig 9.16



Nótese que al ser la varianza poblacional del estadígrafo $\bar{x}$ menor que la varianza poblacional de la variable X, esto se expresa en que, la curva normal que representa la distribución de la variable $\bar{x}$ al estar centrada en μ es, geométricamente hablando mas alta o apuntada en los alrededores del valor medio μ que la distribución madre normal de X.

Si se habla en términos geométricos de área bajo la curva normal, lo expresado también puede interpretarse como sigue; al ser la varianza de $\bar{x}$ menor que σ^2, esta información pone de manifiesto que la curva normal de $\bar{x}$ tiene mas área cerca de la media poblacional μ que la curva normal de X, al menos para ciertos intervalos próximos a la media.

El valor $\sigma/\sqrt{n}$, que como se sabe, es la desviación estándar poblacional de los valores muestrales de $\bar{x}$, también se le conoce en estadística como error estándar de la media muestral o error estándar de $\bar{x}$ y se le denota por EE ($\bar{x}$).

Cuando σ es un valor desconocido, EE ($\bar{x}$) también lo es, pero se puede obtener un valor aproximado de EE ($\bar{x}$) por medio de $S/\sqrt{n}$, si recordamos que μ_{S^2} o media del estadígrafo S^2 tiene como valor a σ^2, significando esto que, la distribución muestral de S^2 esta centrada en σ^2 que por tanto para valores de n grandes, deberá esperarse que $S^2 \approx \sigma^2$.

Hasta el momento en el tratamiento del tema actual, no hemos hecho precisión alguna respecto al valor de n y su repercusión en el hecho de que con frecuencia tenemos que analizar variables aleatorias X que no poseen una distribución normal, pero, sin embargo la solución en la generalidad de los casos depende de la distribución del estadígrafo $\bar{x}$.

Cuando nos enfrentamos a esta situación, usualmente se recurre a la aplicación de un resultado de la teoría de las probabilidades conocido como, **teorema del limite central** el cual plantea que, la distribución o ley de probabilidad de $(\bar{x}-\mu)/\sigma/\sqrt{n}$, puede aproximarse por medio de una distribución normal estándar, cuando n se hace crecer ilimitadamente, o como se dice en ocasiones, cuando n es lo suficientemente grande.

Lo expresado se puede reflejar simbólicamente escribiendo que: $(\bar{x}-\mu)/\sigma/\sqrt{n} \sim$ N

(0, 1), cuando n es grande

O en forma equivalente que: $\bar{x} \sim N(\mu, \sigma^2/n)$, cuando n es grande, donde μ y σ son la media y desvió estándar de la población madre.

Ahora bien, nos podemos preguntar, ¿cuan buena es la aproximación normal que nos plantea el teorema del límite central?.

A este respecto muchos autores argumentan que es difícil dar una respuesta precisa a este cuestionamiento, hecho con el que se esta de acuerdo, y mas aun se pueden ofrecer algunos datos al respecto que avalan esta forma de pensar, a pesar de que se hace mención a las distribuciones no normales Poisson, Exponencial y

Uniforme sobre un intervalo que no se estudian en este libro, pero que sin embargo se justifica su cita aquí en aras de ejemplificar lo que se desea:

Leo Breiman, en su libro, Probability and Stochastic Proceses (La probabilidad y los procesos estocásticos), nos dice que cuando la distribución madre de X es una ley de probabilidad:

· Binomial con parámetro p = ½ , Poisson con parámetro λ = 1 y Exponencial con parámetro λ = 1, entonces la diferencia máxima entre las dos leyes de probabilidad es un valor inversamente proporcional a $\sqrt{n}$, o sea, existe una cierta constante c positiva, tal que, la diferencia máxima vale $c/\sqrt{n}$, con c = 0.7, 0.2 y 0.13 en cada caso respectivamente.

· Uniforme sobre el intervalo $[0,1]$, entonces la diferencia máxima entre ambos tipos de leyes de probabilidad es de orden inverso proporcional a n, o sea que, existe cierta constante c positiva, tal que, la diferencia máxima vale 0.02 / n.

para valores de n = 10, 20 y 80 en los casos en que la ley de X es la Binomial o la Poisson, n = 5, 10, y 20 para la Exponencial y n = 2 y 5 en el caso de la ley Uniforme citada.

Como puede observarse en los ejemplos citados, lo preciso de la aproximación esta dado en función o términos de n, pero también, en dependencia de la distribución que tenga X, la variable original. Por ello se conviene en la practica cotidiana, que para valores de n mayores o iguales a 30 ($n \geq 30$) y variables aleatorias X con distribución no normal, pero simétricas o aproximadamente simétricas respecto al valor medio poblacional respectivo (μ), la aproximación dada por el teorema del límite central se considere aceptable dentro de ciertos límites.

Para aquellos casos en que la distribución de X se hace muy asimétrica, como cuando se trabaja con problemas que involucran el uso de proporciones o porcentajes, en los que las proporciones poblacionales se alejan del valor ½, bien hacia la derecha con valores cercanos a 1 o por el contrario hacia la izquierda con valores cercanos a 0, en pago para hacer uso de la aproximación por medio del teorema anteriormente citado, se debe aumentar paulatinamente el valor de n a mdida que el valor de la proporción poblacional sea muy próxima a 0 o a 1.

9.12 Breve noción sobre las distribuciones Chi- cuadrado y t de student.

Cuando se estudian experiencias aleatorias, es usual encontrar discrepancias entre la frecuencia observada de un resultado y la frecuencia con que se espera que este ocurra, de acuerdo con el conocimiento que se tiene de las probabilidades asociadas a los mismos. Veamos dos ejemplos:

a) Distribución de 100 recién nacidos según sexo.

	Valor del sexo		
Frecuencia	F	M	Total

Observada	60	40	100
Esperada	51	49	100

b) Distribución de60 lanzamientos de un dado

Numero observado en la tirada del dado

Frecuencia	1	2	3	4	5	6	Total
Observada	11	8	12	9	11	9	60
Esperada	10	10	10	10	10	10	60

Si se observa, las dos distribuciones anteriores usan 2 y 6 casillas o celdas respectivamente, una por cada valor correspondiente de la variable. En lo que sigue y con el objetivo de exponer el asunto de manera general, se considera una distribución de una variable que tenga k valores posibles en total, las frecuencias observadas en una muestra se denotaran por O_i y las esperadas o teóricas por E_i.

Para medir la magnitud de las posibles diferencias entre O_i y E_i se usa un estadígrafo, que recibe el nombre, Ji-cuadrado o Chi-cuadrado en atención a que se usa la letra minúscula griega del referido nombre pero elevada a una potencia cuadrática, o sea χ^2, definido por la formula o expresión siguiente:

$$\chi^2 = (O_1 - E_1)^2 + (O_2 - E_2)^2 + (O_3 - E_3)^2 + ... + (O_k - E_k)^2$$

donde k significa el número de valores posibles que tiene la variable en cuestión.

En un experimento ideal, al coincidir las frecuencias observadas y las esperadas, las diferencias que aparecen dentro de los términos cuadráticos de la suma anterior son todos 0 y, por ende, $\chi^2 = 0$. Este es el valor más pequeño que puede tomar el estadígrafo.

Por un mecanismo similar, es decir, mediante el concepto de límite del histograma o polígono de frecuencias de esas medidas de discrepancia para un tamaño muestral dado, se llega a la distribución teórica de χ^2 . Ese histograma nos diría, aproximadamente, en qué porcentaje de tales experiencias se esperaría obtener ciertos rangos de valores del estadígrafo. Como existe solo un número finito de resultados posibles para las frecuencias, tanto observadas como esperadas, solo existe un número limitado de valores posibles de χ^2, por lo que dicha distribución es discreta. Como este tipo de distribuciones, cuando se tiene un número grande de repeticiones, requiere de cálculos largos y engorrosos, por consideraciones prácticas se emplea una distribución continua que es una aproximación a la distribución discreta y que es la única que se usa. Esa distribución aproximada recibe el nombre de distribución Ji-cuadrado o Chi-cuadrado.

Una característica interesante de esa distribución es que su forma depende solamente del número de celdas y se acostumbra a especificar la distribución por medio de un parámetro llamado grados de libertad y que está en función de dicho número de celdas. En este caso, ese parámetro sería k-1 (pues solo varían con la

repetición de la experiencia las frecuencias observadas). La frase grados de libertad se refiere al número de celdas independientes. En nuestro ejemplo, como son 60 tiradas del dado, la frecuencia observada de la sexta celda queda determinada en cuanto se especifican las frecuencias observadas de las primeras 5 celdas, luego los grados de libertad son 5.

En el gráfico que aparece a continuación (fig 9.17) se muestra cómo influye el parámetro grados de libertad sobre la forma de esta distribución.

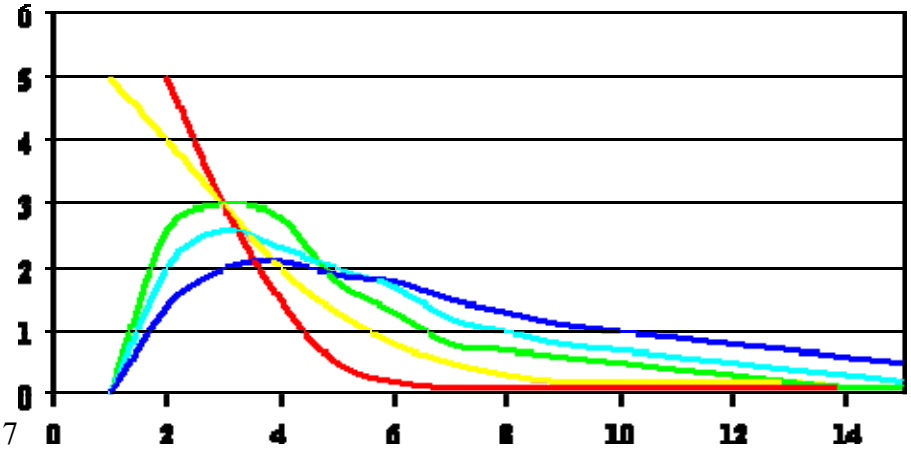


Fig 9.17

En la Tabla 2.C se muestran los valores de χ^2 para algunas probabilidades y algunos grados de libertad. En la primera columna de la misma aparecen los grados de libertad y en la primera fila o encabezamiento aparecen las probabilidades de que χ^2 exceda al valor en la tabla. Por ejemplo, para 6 grados de libertad la probabilidad de que χ^2 exceda a 12.5916 es menor o igual que 0.050.

No solamente es valido introducir la distribución Chi-cuadrado por la vía anterior, recordemos que fue con el objetivo de resolver el problema que surge cuando aparecen discrepancias entre las frecuencia observadas y esperadas o teóricas en determinados tipos de situaciones que hacen uso de variables aleatorias discretas, por medio del uso del estadígrafo:

$$\chi^2 = (O_1 - E_1)^2 + (O_2 - E_2)^2 + (O_3 - E_3)^2 + \ldots + (O_k - E_k)^2$$

para el cual de cierto modo se justifico que el estadígrafo χ^2, sigue aproximadamente una ley de probabilidad Chi-cuadrado con k – 1 grados de libertad.

También se puede introducir el uso de la distribución anterior por consideración de problemas que hacen uso de variables continuas.

Por ejemplo, supóngase que se tiene una variable aleatoria X con distribución normal con media μ y varianza σ^2, siendo σ un valor desconocido y consideremos que se desea obtener, dado un valor de probabilidad o frecuencia teórica $(1 - \alpha)$, con valor de α próximo a cero, que magnitud puede tomar un valor k para que, conocida una muestra de tamaño n de X, se tenga que:

$$P (S^2 \leq k \cdot \sigma^2) = 1 - \alpha$$

En otras palabras, se desea saber, que valor debe tomar k, para que, la frecuencia o probabilidad de que el estadígrafo S^2 sea menor o igual a $k \cdot \sigma^2$, en una muestra de tamaño n de X, tenga un valor predeterminado $1 - \alpha$, con α próximo a cero.

Para resolver este problema, por supuesto, se debe hacer uso de la expresión de S^2 y trabajar esta algebraicamente para transformarla y lograr con esto que se presente en escena una variable con la distribución Chi-cuadrado.

Como se sabe, $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$, luego se tiene que:

$(n-1)\cdot S^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2$, en donde si sumamos y restamos μ dentro de $(x_i - \bar{X})^2$, y se agrupa convenientemente los términos, se tiene que $(n-1)\cdot S^2 = \sum_{i=1}^{n}[(x_i - \mu) - (\bar{X} - \mu)]^2$, pero $[(x_i - \mu) - (\bar{X} - \mu)]^2 = (x_i - \mu)^2 - 2\cdot(\bar{X} - \mu)\cdot(x_i - \mu) + (\bar{X} - \mu)^2$, luego si se sustituye esto, en la expresión anterior y se aplican las propiedades conocidas de las sumas, se obtiene que, $(n-1)\cdot S^2 = \sum_{i=1}^{n}(x_i - \mu)^2 - 2\cdot(\bar{X} - \mu)\sum_{i=1}^{n}(x_i - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2$.

Ahora bien, $\sum_{i=1}^{n}(\bar{X} - \mu)^2 = n\cdot(\bar{X} - \mu)^2$, ya que dentro de la suma nada depende del índice de la suma. Pero además $\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu = n\cdot\bar{X} - n\cdot\mu = n\cdot(\bar{X} - \mu)$. Sustituyendo y agrupando términos comunes, se tiene que, $(n-1)\cdot S^2 = \sum_{i=1}^{n}(x_i - \mu)^2 - n\cdot(\bar{X} - \mu)^2$.

Si cada termino de esta ultima identidad se divide por σ^2, la igualdad no se altera, pero se obtiene la siguiente expresión, $\frac{(n-1)\cdot S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} - n\cdot\frac{(\bar{X} - \mu)^2}{\sigma^2}$, donde si se trabaja con n y σ^2 en los términos del miembro derecho, de manera que se introduzcan dentro de los paréntesis, se llega finalmente a la siguiente identidad,

$$\frac{(n-1)\cdot S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

Es posible observar ahora las características siguientes que se pueden deducir de esta última expresión. Se conoce que $X \sim N(\mu, \sigma^2)$, en consecuencia para cada termino de la forma, $\left(\frac{x_i - \mu}{\sigma}\right)$, se cumple que $\left(\frac{x_i - \mu}{\sigma}\right) \sim N(0, 1)$, pero también se tiene que $\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \sim N(0, 1)$. Luego se tiene que la variable aleatoria $\frac{(n-1)\cdot S^2}{\sigma^2}$ esta vinculada a una expresión que es, una suma algebraica de cuadrados de variables normales estándar.

Por esa razón, se conoce hoy día en estadística que la variable $\frac{(n-1)\cdot S^2}{\sigma^2}$ tiene una distribución Chi-cuadrado con $(n - 1)$ grados de libertad. Es común por ello representarla por medio del símbolo χ^2 y escribir que

$$\chi^2 = \frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2 (n - 1)$$

Luego la probabilidad planteada, $P(S^2 \leq k \cdot \sigma^2) = 1 - \alpha$, puede ser manipulada por medio de un proceso similar al de la estandarización en el caso de la normal, del modo siguiente:

$P(S^2 \leq k \cdot \sigma^2) = P((n - 1)S^2 \leq (n - 1) \cdot k \cdot \sigma^2)$, al multiplican ambos miembros por $(n - 1)$,

$= P((n - 1)S^2/\sigma^2 \leq (n - 1) \cdot k)$, luego se dividen ambos por σ^2,

$= P(\chi^2 \leq (n - 1) \cdot k) = 1 - \alpha$.

Y como $\chi^2 = \frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2 (n - 1)$, entonces disponiendo de una tabla de la Chi-cuadrado,

puede encontrarse el valor aquel que denotado por $\chi^2_{1-\alpha}$ logra que, $(n - 1) \cdot k = \chi^2_{1-\alpha}$, de donde se despejaría k, de modo que $k = \chi^2_{1-\alpha}/(n - 1)$.

Continuando adelante con la segunda parte del tema en curso, el de introducir el estudio de la distribución t-Student, se quiere expresar que, de manera semejante a como ya se vio en la segunda versión introductoria de la distribución Chi-Cuadrado, el hecho de que esta distribución este emparentada con la distribución normal estándar por medio del estadígrafo $(n-1)\cdot S^2/\sigma^2$, vuelve a repetirse, pero ahora a un cierto nivel de complejidad un poco mas elevado por medio de otro estadígrafo de muy amplio uso en los métodos estadísticos que posteriormente serán desarrollados dentro del capitulo dedicado a los métodos de la inferencia estadística, es decir, la estimación por medio de intervalos de confianza y las pruebas de hipótesis. Se trata de que en estos métodos que se han mencionado, se hace amplio uso como se vera posteriormente de la variable aleatoria, cuya expresión es:

$Z = \frac{X-\mu}{\sigma/\sqrt{n}}$ la cual, cuando se supone que $X \sim N(\mu, \sigma^2)$, se conoce que tiene una distribución normal estándar.

Como ya se expreso con anterioridad $\mu_{S^2} = \sigma^2$, es decir, el estadígrafo S^2 esta centrado en σ^2, lo que significa que S lo esta en σ, luego, ¿qué pasaría con la distribución de Z, si se sustituye σ por S en dicha expresión, si se continua suponiendo que $X \sim N(\mu, \sigma^2)$?.

Es decir, ¿continuaría en esta nueva situación, la variable $\frac{X-\mu}{S/\sqrt{n}}$, que habitualmente se denota por **t,** teniendo una distribución normal estándar?.

La respuesta a esto es negativa, ya que hoy día se conoce que $t = \frac{X-\mu}{S/\sqrt{n}}$, bajo la suposición hecha respecto de X, sigue un modelo teórico de distribución de probabilidades denominado como, t-Student, del cual se sabe que su forma geométrica es muy parecida a la distribución normal estándar, excepto para valores pequeños de n (menores o iguales a 30), es decir, su modelo está centrados en 0 y tiene forma acampanada, y para valores de n superiores a 30 muy próximo al de la curva normal estándar.

En la siguiente ilustración grafica se puede apreciar todo lo dicho al respecto:

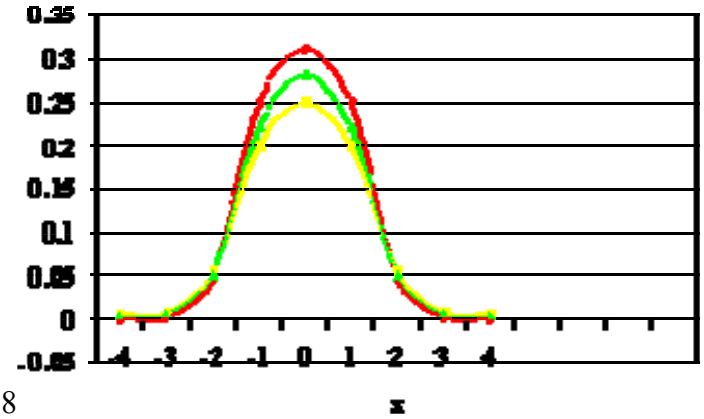


Fig 9.18

La distribución t-Student al igual que la Chi-Cuadrado depende del valor de n, y para ella se sigue usando la misma terminología de grados de libertad, por ello en la práctica de su uso, se dice que: el estadígrafo t ~ t-Student con (n-1) grados de libertad, siendo n en este caso, el tamaño muestral.

El porque de esta herencia de los grados de libertad, se puede apreciar, si se observa que la expresión original de uso de t, $\frac{X-\mu}{S/\sqrt{n}}$ puede transformarse en $z/\sqrt{\chi^2/(n-1)}$, por medio de simples cambios algebraicos que a continuación se muestran:

$t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/\sigma}{S/\sigma\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{S/\sigma}$, los cuales consisten en dividir numerador y denominador de la primera expresión por σ en primera instancia y luego en una segunda fase pasar la raíz cuadrada del tamaño muestral del denominador de la fracción inferior para el denominador de la fracción superior, cambios estos que como cualquiera puede comprobar dejan inalterada la expresión de t, pero que, sin embargo contribuyen a darle el aspecto necesario para lograr lo que se quiere hacer.

En la última expresión se puede identificar a $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ con Z, y de la expresión ya conocida, $\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ deducir la siguiente $\frac{S^2}{\sigma^2} = \frac{\chi^2}{(n-1)}$, y de esta última que, $\frac{S}{\sigma} = \sqrt{\frac{\chi^2}{(n-1)}}$.

Sustituyendo, se llega a la expresión para t que ya se anticipo, significando esta que, la variable t se puede escribir como el cociente de dos variables, de las cuales, la del numerador (Z) es siempre una variable normal estándar mientras que la del denominador depende de una variable χ^2 (chi-cuadrado) en forma de, raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad.

La tabla D de uso práctico de la t-Student, se muestra en el anexo. En ella los valores de los grados de libertad, (n – 1), aparecen en la primera columna del extremo izquierdo, mientras que los valores de probabilidad se muestran en la primera fila superior encabezando las columnas de la tabla, correspondiendo estos, a la frecuencia teórica o probabilidad de que ocurra que el valor de t sea superior en valor absoluto al numero que se encuentra en la intersección de la columna y una fila cualquiera.

Así para 15 grados de libertad, la probabilidad de que t sea superior en valor absoluto a 2.9467 es 0.01.

Ejercicios resueltos:

1.- Si A y B representan dos eventos mutuamente excluyentes tales que P (A) = 0.35 y P (B) = 0.50, calcule cada una de las probabilidades siguientes:

(a) P (A o B) (b) $P((A \text{ o } B)^C)$ (c) $P(A^c)$ (d) $P(B^c)$ (e) $P((A \text{ y } B)^c)$

Soluciones:

(a) P (A o B) = P (A) + P (B), por Ley de Adición, ya P (A y B) = 0

= 0.35 + 0.50 = 0.85

(b) $P((A \text{ o } B)^c) = 1 - P(A \cup B) = 1 - 0.85 = 0.25$, ya que $P(A \cup B) + P((A \cup B)^c) = 1$,

pues $(A \cup B)$ y $(A \cup B)^c$ constituyen una partición del espacio muestral.

(c) $P((A)^c) = 1 - P(A) = 1 - 0.35 = 0.65$

(d) P (B^c) = 1 - P (B) = 1 - 0.50 = 0.50

(e) P ((A y B)^c) = 1, pues P (A y B) = 0 por la suposición, es decir, por ser eventos mutuamente excluyentes)

2- Si la probabilidad de tener un hijo varón es 0.51, ¿Cual es la probabilidad de que en una familia que tenga 5 hijos, exactamente 2 sean varones?.

Solución:

Si se define la variable, X = Número de hijos varones que tiene una familia de 5 hijos.

Se tiene que por la condiciones del enunciado, X sigue una distribución binomial con parámetros n = 5 y p = 0.51 y evento, éxito = ser hijo varón.

La probabilidad pedida es, Prob.(Una familia de 5 hijos tenga dos hijos varones), la cual se interpreta en términos de X como P (X = 2).

Según la formula general de la distribución binomial, se tiene que:

$(P\ X = k) = C_{n,k} \cdot p^k \cdot (1 - p)^{n-k}$,

luego aplicando esta al caso presente: $P\ (X = 2) = C_{5,2} \cdot (0.51)^2 \cdot (0.49)^3$, donde, $C_{5,2}$ = fact(5) / (fact(2) · fact(3) = 5 · 4 · fact(3) / 1 · 2 · fact(3) = 5 · 2 = 10.

Luego $P\ (X = 2) = 10 \cdot (0.51)^2 \cdot (0.49)^3 = 10 \cdot (0.2601).(0.117649) = 0.306$

Este valor, se interpreta como que, aproximadamente en el 31% de los casos en que se observan familias con 5 hijos, encontraremos en ellas 2 varones.

3.- La probabilidad de que padres con cierto tipo de ojos azul- pardo tenga un hijo con ojos azules es 0.25. Si la pareja tiene 6 hijos, ¿Cual será la probabilidad de que al menos la mitad de ellos tenga los ojos azules?.

Solución:

Si se define la variable:

X = Número de hijos con ojos azules de 6 que tiene una familia en que ambos padres tiene ojos azul-pardo .

Se tiene que X sigue una distribución binomial con parámetros n = 6 y p = 0.25 = ¼ con evento, éxito = ser hijo de ojos azules

La probabilidad que se pide calcular se interpreta en función de la variable X, como P ($X \geq 3$), ya que “al menos la mitad de los hijos tengan ojos azules” se interpreta como que X es mayor o igual a 3.

Como X es discreta, con probabilidades concentradas en los puntos 0, 1, 2, 3, 4, 5, 6, entonces:

P (X ≥ 3) = P (X = 3) + P (X = 4) + P (X = 5) + P (X = 6), pero en realidad es mejor proceder del modo siguiente, que emplea un poco de menos calculo.

P (X ≥ 3) = 1 – P (X ≤ 2), pues los eventos (X ≥ 3) y (X ≤ 2) según lo descrito anteriormente resultan excluyentes y exhaustivos.

En consecuencia P (X ≥ 3) = 1 – (P (X = 0) + P (X = 1) + P (X = 2)), donde las probabilidades que aquí aparecen responden a la formula :

$(P\ X = k) = C_{n,k} \cdot p^k \cdot (1 - p)^{n-k}$

en la cual debe considerarse que k = 0, 1, 2 respectivamente con n = 6 y p = ¼.

Así se tiene que:

$P(X = 0) = C_{6,0} \cdot (1/4)^0 \cdot (1 - (1/4))^{6-0} = 1 \cdot 1 \cdot (3/4)^6 = 3^6 / 4^6 = (3^3)^2 / (4^3)^2$

$= 27^2 / 64^2 = 729 / 4096.$

$P(X = 1) = C_{6,1} \cdot (1/4)^1 \cdot (1 - (1/4))^{6-1}$

$= 6 \cdot (1/4) \cdot (3/4)^5 = 2 \cdot (3^6/4^6) =$

= 1458 / 4096.

$P(X = 2) = C_{6,2} \cdot (1/4)^2 \cdot (3/4)^{6-2} = (\text{fact}(6) / (\text{fact}(4) \cdot \text{fact}(2)) \cdot (3^4 / 4^6)$

$= (6 \cdot 5 \cdot \text{fact}(4)/2 \cdot \text{fact}(4)) \cdot (81/4096) = 15 \cdot (81/4096) = 1215 /4096.$

Sustituyendo se tiene que: P(X≥3) = 1– (729/4096 + 1458/4096 + 1215 /4096)

Luego P (X ≥ 3) = 1 – (3402 / 4096) = 694 / 4096 = 0.169

Este resultado muestra que la posibilidad de que en una familia de este tipo hayan 3 o más hijos de ojos azules es muy pequeña, ya que aproximadamente en 17 de cada 100 familias puede cumplirse tal condición.

4.- Los residentes de una comunidad son sometidos a un pesquizaje de cáncer. Los resultados del examen se clasifican en positivos (+) si hay sospecha de malignidad y de negativos (-) si no hay indicación alguna de malignidad, Si una persona tiene cáncer, la probabilidad de sospecha de malignidad es 0.98 y la probabilidad de sospecha donde este no existe es de 0.15. Si el 5% de los miembros de la comunidad padecen cáncer, ¿cual es la probabilidad de que una persona no tenga cáncer dado que el examen es positivo?.

Solución: Sean A_1, A_2 y E los eventos expresados por medio de:

A_1 = persona que padece cáncer, A_2 = persona que no padece cáncer y E = examen con resultado positivo

Por el enunciado del problema se puede identificar que:

$P(A_1) = 0.05$, $P(A_2) = 0.95$, $P(E/A_1) = 0.98$ y $P(E/A_2) = 0.15$

La probabilidad que se desea calcular se interpreta como $P(A_2 / E)$. Se tiene que,

$P(A_2 / E) = P(A_2 \text{ y } E) / P(E) = P(A_2) \cdot P(E/A_2) / P(E)$, por la regla de la multiplicación.

Además, $P(E) = P(A_1).P(E/A_1) + P(A_2).P(E/A_2)$, por la regla de la probabilidad total.

Luego, $P(A_2/E) = (0.95).(0.15) / ((0.5).(0.98) + (0.95).(0.15)) = 0.744$

Note que, esta probabilidad es superior a 0.5, por tanto aún dando positivo el resultado del examen, es mas probable que una persona examinada no tenga cáncer, que lo contrario.

5.- Supongamos que en una población de 1 000 sujetos hay 70 enfermos. En dicha población hay 520 mujeres, 40 de las cuales están enfermas. ¿Cual es la probabilidad de que un sujeto este enfermo dado que es mujer?.

Solución: Sean E y M los eventos expresados por, estar enfermo y ser mujer respectivamente. Se tiene que,

Total de enfermos = 70 Total de mujeres = 520

Total de mujeres enfermas = 40 Total de sujetos = 1000

A partir de estos datos puede identificarse que:

$P(E) = 70/1000$, $P(M) = 520/1000$, $P(E \text{ y } M) = 40/1000$ y que la probabilidad pedida es $P(E/M)$.

En consecuencia, $P(E/M) = P(E \text{ y } M)/P(M) = \frac{40/1000}{520/1000} = 40/520 = 1/13$

que es el mismo resultado que arrojaría la definición clásica, teniendo en cuenta solo las mujeres como espacio muestral.

6.- Consideremos un grupo integrado por 36 estudiantes. Supongamos que los eventos E y F son los siguientes:

E: tener los ojos azules F: ser del género masculino

Supongamos que la distribución conjunta de ambos eventos es la siguiente:

Ojos azules Total

	Si	No	
Masculino	6	6	12
Femenino	12	12	24
Total	18	18	36

¿Cual es la probabilidad de que un estudiante tenga los ojos azules dado que es del genero masculino?. Exprese a que criterio puede llegarse en relación con los evento E y F.

Solución: La probabilidad pedida se identifica como, P (E / F).

En consecuencia, las probabilidades involucradas en su calculo son:

P(E) = 18/36 = ½, P(F) = 12/36 = 1/3, P(E y F) = 6/36 = 1/6.

Luego, $P(E/F) - P(E \text{ y } F)/P(F) - 1/6/1/3$ = ½.

Como P(E / F) = ½ = P(E) los eventos son independientes.

7.- Un farmacólogo dice que una fórmula desarrollada por el para tratar cierta enfermedad es efectiva en un 80% de los casos. Se aplica el medicamento a 400 enfermos seleccionados aleatoriamente. Calcule la probabilidad de que se recuperen,

a) 240 pacientes de los 400.

b) De 300 a 325 pacientes de los 400.

c) A lo mas 350 pacientes de los 400.

Solución:

De lo planteado en el enunciado se tiene la variable:

X = Número de enfermos que se recuperan de 400 tratados con el medicamento.

Sigue una ley binomial con parámetros n = 400, p = 0.8 y evento, éxito = enfermo recuperado por uso del medicamento.

De este modo el planteamiento de los incisos viene dado en términos de X por:

a) P (X = 240) b) P ($300 \leq X \leq 325$) c) P ($X \leq 350$)

Como n es grande, se debe observar si $n \cdot p \cdot (1 - p) \geq 5$, para poder emplear la aproximación binomial-normal.

Se tiene que, $n \cdot p \cdot (1 - p) = 400 \cdot 0.8 \cdot 0.2 = 64 \geq 5$, luego es posible usar la aproximación, y entonces:

a.- $P(X = 240) = P(X \leq 240) - P(X \leq 239)$, ya que X es una variable aleatoria discreta cuyos valores de probabilidad se encuentran concentrados en 401 puntos, a saber, los valores del 0 al 400.

Pero $P(X \leq 240) = P(X \leq 240 + 0.5)$ y $P(X \leq 239) = P(X \leq 239 + 0.5)$, por ser X discreta.

En consecuencia: $P(X = 240) = P(X \leq 240.5) - P(X \leq 239.5)$.

Haciendo uso ahora de la aproximación binomial-normal, debemos considerar que:

X se distribuye aproximadamente $N(\mu = n \cdot p = 320, \sigma^2 = n \cdot p \cdot (1 - p) = 64)$

$$\textbf{Luego, } P(X \leq 240.5) = P\left(Z \leq (240.5 - 320)/\sqrt{64}\right) = P\left(Z \leq -\frac{79.5}{8}\right) = P(Z \leq -9.94)$$

$$= P(Z \geq 9.94), \text{ por la simetría de la curva normal.}$$

$$= 0.5 - P(0 \leq Z \leq 9.94).$$

$$\textbf{Análogamente, } P(X \leq 239.5) = 0.5 - P(0 \leq Z \leq 10.06)$$

Pero según la tabla normal estándar $P(0 \leq Z \leq 9.94) = P(0 \leq Z \leq 10.06) = 0$. Luego $P(X \leq 240.5) = P(X \leq 239.5) = 0$ y en consecuencia $P(X = 240) = 0$.

Esto significa que, con el poder de efectividad planteado por el farmacólogo para su medicamento, el evento, obtener 240 éxitos en 400 enfermos tratados, resulta ser un suceso prácticamente imposible de observar.

b.- $P(300 \leq X \leq 325) = P(X \leq 325) - P(X \leq 299)$, igualdad valida por ser X una variable discreta concentrada sobre los puntos del 0 al 400.

Nótese ahora que las probabilidades del miembro derecho son del mismo tipo del inciso (a), solamente que han variado los valores. Por tanto el procedimiento a usar es el mismo, por dicha razón solo se expondrán los resultados finales, así.

$P(300 \leq X \leq 325) = P(Z \leq 0.63) - P(Z \leq -2.63)$.

Pero $P(Z \leq 0.63) = 0.5 + 0.2357 = 0.7357$, según la tabla normal estándar y

$P(Z \leq -2.63) = P(Z \geq 2.63) = 1 - P(Z \leq 2.63) = 1 - 0.9957 = 0.0043$.

Sustituyendo, se tiene que

$P(300 \leq X \leq 325) = 0.7357 - 0.0043 = 0.7314$.

Este valor nos dice que, aproximadamente en el 73 % de las ocasiones el número de éxitos se debe encontrar en el intervalo de 300 a 325.

c.- P ($X \leq 350$) = P ($Z \leq 3.75$) = 0.5 + 0.4973 = 0.99973 ≈ 1, valor cuyo significado precisa, que en la practica de observar 400 pacientes a los cuales se les haya suministrado dicho medicamento es casi prácticamente seguro encontrar a lo mas 350 pacientes que se hayan recuperado.

Ejercicios para resolver:

1.- Sean A y B eventos de un espacio muestral S tales que: P (A) = 0.5, P (B) = 0.3 y P (A ? B) = 0.1. Calcule las probabilidades de cada uno de los eventos siguientes:

(a) A o B (b) A pero no B (c) B pero no A (d) ni A ni B

2.- (Problema de Chevalier de Mere). ¿Cuál de los eventos siguientes cree usted que es más probable que ocurra?.:

a) obtener al menos un as (uno) en la tirada simultánea de cuatro dados

b) obtener al menos un doble as en una serie de 24 tiradas de un par de dados

Suponga que los dados no están cargados y explique su respuesta.

3.- Tome un editorial del periódico Granma que contenga no menos de 1 000 palabras. Liste el número de veces que aparece cada letra del alfabeto. Basándose en esto, ¿cuál probabilidad de ocurrencia podría asignar usted a la aparición de una letra en un editorial?.

4.- Si el evento A es un subconjunto del evento B, demuestre que P (A) ≤ P (B) . Sugerencia: descomponga B en dos partes, las comunes a A y B y las que solo están en B, es decir: A y B-A.

5.- Supongamos que si a una persona con tuberculosis se le aplica un test de pesquizaje de TB, la probabilidad de que sea detectada su condición es de 0.90 y que si se lo aplican a una persona sana, la probabilidad de que sea incorrectamente diagnosticada es 0.3. consideremos además que el 11 % de los adultos residentes en una cierta ciudad padecen de tuberculosis. Si una de esas personas es diagnosticada como tuberculosa a punto de partida del resultado de dicha prueba , cual es la probabilidad de que la misma tenga tuberculosis?.

Interprete el resultado.

6.- Una persona olvidadiza debe darle a un familiar una píldora cada día. La probabilidad de que ella olvide darle la píldora al pariente es 2/3. Considere además que si familiar recibe la píldora, la probabilidad de que muera es de 1/3, mientras que si no la recibe, la probabilidad de morir se convierte en 3/4. El pariente muere. ¿Cual es la probabilidad de que la persona haya olvidado darle la píldora?.

7.- Una enfermedad infecciosa se propaga por contacto. Toda persona susceptible tiene la posibilidad de infectarse en cada contacto con una persona infectada, pero se vuelve inmune después de haber padecido la enfermedad y ya no la transmite. Responda las siguientes preguntas:

- ¿Cuánto tiempo durará la epidemia?

- ¿cuál es la probabilidad de que la **enfermedad se extinga**?.

8.- En una encuesta de 100 personas, clasificadas como bebedores y no bebedores y como poseedores o no de hígado graso, se obtuvieron los resultados siguientes:

	Hígado graso		Total
	Si	No	
Bebedores	52	18	70
No bebedores	8	22	30
Total	60	40	100

Sean los eventos E, E^c, F y F^c definidos por:

E: Ser bebedor E^c: No ser bebedor

F: Tener hígado graso F^c: No tener hígado

Diga cuales de estas parejas de eventos son independientes:

a) E y F b) E y F^c c) E^c y F

Capítulo 10.- Nociones elementales de Muestreo.

10.1 Introducción.

En la práctica médica se presenta, muy frecuentemente, la necesidad de estudiar algún hecho o fenómeno sobre un conjunto de individuos, al mismo tiempo que hay razones que imposibilitan o dificultan la obtención de información de todos los sujetos; por ejemplo como cuando se inicia un estudio experimental de un nuevo medicamento o cuando se quiere introducir una nueva técnica quirúrgica para el tratamiento de una enfermedad; o para conocer las características clínico-epidemiológicas de una enfermedad en un área de salud determinada.

Por problemas éticos, no podemos usar nuevos tratamientos indiscriminada- mente sobre todos los pacientes, ya sean estos medicamentosos o procedimientos quirúrgicos. También cuando se estudia una enfermedad en particular puede resultar imposible disponer de todos los recursos necesarios para la realización de todas las pruebas y examenes que resultarían imprescindibles, debido al hecho de que la morbilidad asociada a una enfermedad tiene 2 componentes uno visible constituido por los enfermos conocidos y otra que permanece oculta, generalmente importante y en ocasiones difícil de estimar.

Además de que el tiempo a emplear podría resultar excesivamente largo, aun en el caso en que se contara con una organización del trabajo adecuada y que permitiese llevar a cabo el mismo.

Incluso, hay estudios cuya realización lleva implícita la destrucción de los **sujetos** que se estudian, como es el caso de la inoculación de un cultivo o sustancia de la cual se sospecha que puede provocar una enfermedad o condición patológica o el control de la calidad de la producción. En el primer caso, hay que trabajar con animales de laboratorio para realizar la disección, a fin de analizar los cambios ocurridos a nivel de órganos y tejidos.

Como hemos visto, la imposibilidad de estudiar el fenómeno sobre todos los sujetos en los cuales este puede tener lugar, no solo se presenta cuando el grupo de individuos es muy númeroso o no se conoce a todos los elementos con la característica que interesa, sino que puede estar dada también por razones éticas, de tiempo o de costos.

¿Cómo, entonces, resolver esta situación? La solución es realizar el estudio solo sobre una parte del grupo de elementos y luego extender los resultados al resto del grupo, con cierto grado de confiabilidad. Pero, ¿qué puede decirse sobre el colectivo del cual solo se estudia un número limitado de elementos? Esto es precisamente la esencia de la **teoría del muestreo**: proveer métodos para la recolección adecuada de los sujetos que serán estudiados, de modo tal que lo observado en este grupo pueda generalizarse con un determinado grado de confianza. a la totalidad del conjunto del cual fueron extraidos.

Es sumamente importante enfatizar que aunque lo que se observa o estudia es solo una parte del todo, el objetivo es que las conclusiones a las que se arribe sean aplicables a este. Esto implica un proceso de generalización de los resultados obtenidos, que podría muy bien no tener sentido si los elementos o unidades que se seleccionaron para realizar el estudio no han sido escogidos adecuadamente.

Antes de continuar, precisemos un grupo mínimo de conceptos que son imprescindibles para poder expresar claramente las ideas del muestreo.

10.2 Conceptos básicos de la teoría del muestro.

La idea de tomar una pequeña parte de un todo para representarlo o caracterizarlo es de amplio uso en las disciplinas más disímiles, por ejemplo se usan muestras de tejidos y maderas para que el comprador decida cuál le gusta más para adquirirla; se hacen muestras pictóricas en las que se presenta un cierto número de

cuadros de un determinado autor para dar una visión panorámica de su trabajo; también cuando un especialista en literatura va a realizar el análisis crítico de toda la obra de un determinado autor, lo discute sobre la base de una parte de la misma (los trabajos más importantes o más significativos). El uso de tales **muestras** presupone que ellas evidencien todas las características más importantes del conjunto original.

Provisionalmente consideramos como población (no universo) para mas adelante precisarlo de un modo objetivo, a un conjunto de individuos que resulta de interés estudiar y al cual queremos extender los resultados observados en un subconjunto del mismo (muestra) de la cual se obtiene la información.

El **concepto de universo** es de **existencia teórica**, pues se refiere, en general, a todas las poblaciones a las cuales se pueden hacer extensivos los resultados muestrales, apoyándose en los conocimientos de la ciencia particular de que se trate, y se obvia la ubicación espacio-temporal del grupo de sujetos estudiados. Debemos aclarar que la Estadística solo permite hacer inferencias a la **población real** de la cual fue extraída la muestra que se estudió, cualquier otra generalización debe fundamentarse en el cuerpo teórico del campo de la ciencia particular en que se ubique el tema en cuestión.

Los conceptos de población y muestra están relacionados. Además, el concepto de población es relativo, y depende de los objetivos del estudio. Veámoslo a través de un ejemplo: si se desea estudiar la estatura promedio de niños de 7 años en Ciudad de La Habana en el 2003, la población estaría integrada por todos los niños de 7 años de la provincia en ese año, mientras que si se deseara estudiar esa misma característica en todo el país, la población sería la integrada por todos los niños de 7 años del país, en el mismo período.Es decir, al definir la población precisamos los sujetos sobre los cuales es válida la inferencia: en suma, delimitamos el alcance de los resultados. Por eso es muy importante definir la Población sin ambiguedades.

Hasta ahora, hemos estado hablando de sujetos o individuos que integran la población, y que, por ende, integrarán la muestra. Eso se debe a que no son solo pacientes lo que podemos estar interesados en estudiar. Podemos necesitar estudiar hospitales, servicios, áreas de salud, tabletas, de modo tal que nos interese conocer las características individuales de cada uno de ellos. Por ejemplo, en el control de calidad en la producción de medicamen- tos los individuos son las tabletas que se usan como muestra para evaluar si el medicamento cumple con los requisitos de producción establecidos, (cada tableta es un sujeto o individuo), y sobre ella se realizan una serie de observaciones y mediciones. Esto significa que una muestra puede estar integrada lo mismo por personas, que por familias, o instituciones, etc.

No siempre es posible o logísticamente fáctible hacer la selección de los sujetos que van a integrar la muestra de forma directa. Es decir, es posible que necesitemos llegar a la unidad básica (sujeto o individuo) por medios indirectos. Por ejemplo, si deseamos realizar un estudio sobre las carac- terísticas clínico - epidemiológicas del asma bronquial en el municipio Arroyo Naranjo, para hacer una selección directa de los pacientes asmáticos que integrarán la muestra sería necesario tener una lista con los nombres de todos esos pacientes en el municipio, labor que podría resultar bastante difícil de realizar. En esos casos se seleccionan de manera indirecta, ya sea por grupos (se selecciona cierto número de consultorios y se incluyen en el estudio todos los niños asmáticos atendidos en ellos) o por medio de selecciones sucesivas (se seleccionan áreas de salud en primera instancia), dentro de esas áreas se selecciona consultorios y dentro de los consul- torios elegidos, se hace una selección de los pacientes asmáticos a incluir en la muestra). Por esta razon se hace necesario considerar los conceptos de **unidad básica o de análisis , unidad de muestreo, unidad de observación y marco muestral**.

Concretando las ideas antes expresadas definamos estos conceptos básicos, como sigue:

Población: Es el conjunto de individuos (personas, animales, objetos, etc.) que presentan la(s) característica(s) que se desea(n) estudiar o son susceptibles de verse afectados por el fenómeno que se quiere investigar, en el marco de una región geográfica bien determinada y en un período o momento del tiempo dado.

Muestra: Es cualquier subconjunto de elementos de la población.

Unidad de análisis: Es cada uno de los elementos o sujetos de la Población sobre los cuales se recoge la información y su naturaleza está determinada por los objetivos que se persiguen.

Unidad de muestreo: Es la que se usa para realizar la selección Muestral, que en los casos más simples coincide con la unidad de análisis, pero generalmente no es así. Por ejemplo, una muestra de estudiantes de nivel secundario se puede obtener mediante una selección de grupos de clase, en lugar de a partir de una selección directa de alumnos; una muestra de pacientes con hipertensión arterial puede seleccionarse tomando alea- toriamente un cierto número de consultorios de médico de familia y considerando como integrantes de la misma a todos los pacientes dis- pensarizados como hipertensos en dichos consultorios. Las unidades elementales son los alumnos o los pacientes hipertensos, pero se usaron como unidades de muestreo los grupos de clase y los consultorios. Incluso, en un mismo diseño muestral se pueden usar varias unidades de muestreo diferentes, y establecer un modelo jerárquico entre las mismas. Por ejemplo, una muestra de habitantes de una provincia se puede tomar a partir de seleccionar, sucesivamente, municipios, áreas de salud dentro de los municipios, consultorios de médico de familia dentro de las áreas de salud, familias dentro de la población atendida por los consultorios y, finalmente, sujetos dentro de cada una de las familias seleccionadas.

Unidad de observación: Es aquella de la cual se obtiene la información. No siempre se obtiene la información a partir de la propia unidad de análisis. Por ejemplo, cuando nuestras unidades de análisis son niños pequeños, la información se obtiene generalmente de sus madres, que ni siquiera pertenecen a la población objeto de estudio.

Marco muestral: Es la lista de todas las unidades de muestreo susceptibles de ser seleccionadas para integrar la muestra.

Estas definiciones, aunque correctas, solo cobran un sentido práctico cuando estamos ante un problema real concreto, y resulta impres- cindible definirlas claramente y sin ambiguedades en todo trabajo en que se aplique alguna técnica de muestreo.

En particular, el concepto de población es relativo, es decir, depende de lo que se desea estudiar, donde se va a estudiar y cuando. Ilustrémoslo por medio de un ejemplo: si un ginecoobstetra quiere estudiar la fertilidad en el área de salud a la que brinda cobertura el hospital donde el trabaja, su población objeto de estudio estaría determinada por:

a) Sujetos en los que está presente la característica: mujeres en edad fértil

b) Área de salud que cubre el hospital: municipio(s), provincia(s), etc.

c) Período de tiempo en que se va realizar la fase de terreno o de recolección de la información.

Al definir la población objeto de estudio hay que dejar bien aclarados los aspectos siguientes: contenido o unidades básicas (1), unidades de muestreo (2), extensión (3) y tiempo (4), por ejemplo, la población a muestrear en el Estudio Nacional de Crecimiento y Desarrollo de 1982, se definiría como: todos los individuos menores de 20 años (1), en los núcleos familiares (2), en Cuba (3), en 1982 (4). Es decir, la población con la cual se va a trabajar debe ser definida con claridad, de modo tal que siempre se pueda saber si un individuo determinado pertenece o no a dicha población.

Veamos algunos ejemplos para ilustrar este último planteamiento. A continuación aparecen tres definiciones de poblaciones, en las cuales podremos detectar deficiencias:

1: Todos los pacientes asmáticos menores de 15 años residentes en el municipio Cerro, que han sido atendidos en el Hospital Pediátrico de Centro Habana en los últimos 5 años.

2: Todos los pacientes asmáticos residentes en el municipio Arroyo Naranjo que han hecho uso de los servicios de urgencia en el último año transcurrido.

3: Las mujeres en edad fértil del municipio Centro Habana

En la primera definición queda bastante claramente delimitada la población, si sé explicíta más tarde que se trata de pacientes dispensarizados como asmáticos y que se incluyen en el estudio tanto la consulta externa como el servicio de urgencia del hospital. Si solo se van a incluir en el estudio los casos atendidos en consulta externa, eso debe formar parte de la defini- ción de la población.

En la segunda definición no se precisa qué institución de salud nos interesa. En estos momentos, además de los hospitales provinciales y municipales, se brinda este tipo de atención en los policlínicos de urgencia.

La tercera definición no tiene ubicación espacio-temporal, por lo que no define una población, sino un universo.

Al referirnos a la muestra decíamos que era un subconjunto de individuos de la población. Ahora bien, ¿cuán grande debe ser ese subconjunto?, o mejor aún, ¿cuál es el menor número de sujetos que se deben seleccionar para tener una muestra cuyos resultados puedan servir para hacer inferencias a la población que la origino?. Este número, la cantidad de unidades básicas a seleccionar en la población, es otro de los temas que se tratan dentro de la teoría del muestreo. Pasaremos ahora a discutir su importancia.

10.3 Tamaño de la muestra o tamaño muestral.

Como dijimos, una muestra no es más que una parte de la Población, con un cierto número de elementos. A ese número se le llama tamaño de la muestra o tamaño muestral y es de importancia vital. Generalmente se denota al tamaño poblacional con la letra mayúscula N y al tamaño muestral con la letra minúscula n.

En la determinación del tamaño muestral intervienen los factores siguientes:

a) La variabilidad poblacional de la característica en estudio, o la frecuencia de aparición o presentación del fenómeno que se estudia.

b) La confiabilidad de los resultados requerida.

c) La precisión deseada de los resultados.

d) El tipo de estudio.

e) Las técnicas estadísticas que se van a utilizar, etc.

Veamos como actúa la variabilidad poblacional de la característica sobre el tamaño muestral. Mientras que para dar una idea clara de las características principales de la obra de un artista, pintor o escritor, hace falta analizar varias obras, dada la diversidad de temas tratados, la madurez obtenida por el autor con el paso del tiempo, etc., para estimar los niveles de hemo-globina o cualquier otra sustancia en sangre solamente le hacen falta al laboratorista unas pocas gotas de la misma, dado que esta es un líquido homogéneo y bien mezclado. Es decir, a menor homogeneidad de los elementos en la muestra, se hace necesario tener una mayor cantidad de ellos.

Como ejemplo de la repercusión que tiene la frecuencia de presentación del fenómeno que se investiga, supongamos que se va a realizar un estudio de factores de riesgo de mortalidad infantil en un lugar donde la tasa de mortalidad infantil es de 8,3 por 1 000 nacidos vivos. Si se estudian 1 000 nacimientos solo se tendrán

unas 8 defunciones, lo que no permite hacer un análisis de este tipo. Para poder realizar el estudio, sería necesario que se estudiaran varias decenas de defunciones y por ende, varios miles de nacimientos.

La precisión en los resultados (grado de aproximación con que se quiere expresar o calcular algo), así como la confiabilidad de los mismos (proba-bilidad de no equivocarse al dar esos resultados por ciertos), tienen efectos similares: a mayor precisión o mayor confiabilidad, mayor tamaño de la muestra. Estos conceptos de precisión y confiabilidad se formalizaran en el próximo capítulo de este libro, donde además se deducira una fórmula para el cálculo del tamaño de muestra para algunos casos particulares de técnicas estadísticas.

También el tipo de estudio, como veremos más adelante, y las técnicas de procesamiento estadístico que se van a emplear se relacionan con el tamaño de la muestra, pues la forma en que van a ser recogidos y procesados los datos a su vez impone requisitos y limitaciones.

Por el momento no profundizaremos más en este tema, pues el mismo será retomado más adelante en el texto. Por ahora es suficiente que se comprenda la importancia del tamaño muestral para poder alcanzar los objetivos que se propone una investigación.

10.4 Ventajas y desventajas del uso del muestreo.

Como hemos visto, hay varias razones que hacen necesario o preferible el estudio por medio de muestras al estudio de una poblacion. Además de las causas vistas anteriormente, hay otras consideraciones de orden práctico por las que resulta ventajoso el uso de muestras. Las principales ventajas que ofrece el uso de muestras se relacionana continuación:

1.- Reducción de los costos: Al ser menor la cantidad de sujetos a estudiar, los gastos de salarios, materiales gastables, etc., resultan proporcionalmente menores.

2.- Mayor rapidez en la obtención de resultados: Al tener que recolectar y procesar datos de un grupo más pequeño de individuos estas tareas pueden realizarse en un tiempo menor, lo que garantiza una obtención de resultados más rápida.

3.- Mayor alcance o esfera de acción: Cuando se trabaja con un número muy grande de sujetos, resulta casi imposible utilizar equipamiento muy especializado, realizar pruebas complejas, etc., por razones de disponi-bilidad. En una muestra, sin embargo, hay una mayor flexibilidad y posibilidad de obtener información de muy diversos tipos o de carácter y especializada.

4.- Mayor precisión en los datos: Al recoger la información de un grupo relativamente pequeño de sujetos es posible utilizar el personal más calificado, mejor entrenado y llevar a cabo una supervisión más cuidadosa y eficiente del trabajo de terreno y del procesamiento de datos.

Como desventaja se debe señalar, que, siempre que se utilizan muestras para estudiar el problema que nos interesa en la población se espera que las mismas sean un modelo **reducido** o a **pequeña escala** de la misma y, en la medida en que el comportamiento de estas se diferencie del que se pretende modelar (el poblacional), se está sujeto a errores de apreciación. Esto puede verse de la forma siguiente, es común que muestras de igual tamaño seleccionadas de la misma población difieran entre sí en cierta medida; algunas se asemejan más a la población original y otras menos, entonces, ¿cómo se podría estar seguro de que la muestra elegida **realmente representa** a la población de la cual fue extraída?. Si la característica en estudio se comporta de forma bastante homogénea en la población que se muestrea, la posibilidad de tener una muestra que no sea representativa de la población es relativamente pequeña; pero si el comportamiento de la misma es muy variable, el riesgo de seleccionar una muestra que conduzca a resultados erróneos puede ser grande. Es a ese error en que se incurre, al menos en teória, al inferir la situación de la Población por medio de lo observado en la muestra, al que se denomina **error de muestreo**.

Veamos, a través de un ejemplo, cómo se produce este error. Supongamos que tenemos una población de solo 5 individuos, cuyos valores respectivos de hemoglobina (g / l) son los siguientes: 117, 120, 125, 125, 130.

Los valores se han ordenado en forma creciente para facilitar la comprensión del ejemplo.

Si para esta población hipotética quisiésemos estimar el valor medio de hemoglobina mediante el empleo de muestras de tamaño 3, tendríamos un total de 10 muestras posibles y una media muestral para cada una de ellas. Hagamos los cálculos solo para las 4 muestras siguientes:

I) 117, 120, 125 (los tres más bajos):

$\overline{X}$ = (117+120+125) / 3 = 362 / 3 = 120.7 g / l.

II) 125, 125, 130 (los tres más altos)

$\overline{X}$ = (125+125+130) / 3 = 380 / 3 = 126.7 g / l.

III) 117, 125, 130 (los tres diferentes)

$\overline{X}$ = (117+125+130) / 3 = 372 / 3 = 124.0 g/l.

IV) 120, 125, 125 (los tres valores centrales)

$\overline{X}$ = (120+125+125) / 3 = 370 / 3 = 123.3 g/l.

La hemoglobina media de esta población es:

μ = (117+120+125+125+130) / 5 = 123.4 g/l.

NOTA: recordemos que en la realidad desconocemos la media poblacional y no tenemos forma de calcularla, por esra razon es que se hace uso del muestreo.

Al comparar la media poblacional con las medias muestrales obtenidas y las medias muestrales entre sí, es fácil constatar que son diferentes. Las medias de las muestras con los valores más a los extremos (los más elevados y los más bajos) difieren más de la media poblacional que la calculada a partir de los valores **centrales**. A estas diferencias es a lo que hace alusión el **error de muestreo**.

La magnitud de este error está estrechamente vinculada con el tamaño muestral. Para confirmarlo se hara una laboriosa experiencia, al tomar la misma población hipotética que se uso para ejemplificar el error de muestreo. La tabla que a continuación se muestra, presenta la información relativa a todas las muestras posibles de tres tamaños diferentes: 10 muestras de tamaño 2; 10 muestras de tamaño 3, y 5 muestras de tamaño 4.

Tabla 10.1. Muestras de tamaños 2, 3 y 4.

Tamaño Muestral	**Datos que** la integran	**Datos** muestrales	**Media** Muestral
2	0 y 1	117 120	118.5
	0 y 2	117 125	121.0
	0 y 3	117 125	121.0
	0 y 4	117 130	123.5
	1 y 2	120 125	122.5
	1 y 3	120 125	122.5
	1 y 4	120 130	125.0
	2 y 3	125 125	125.0

	2 y 4	125 130	127.5
	3 y 4	125 130	127.5
3	0, 1 y 2	117 120 125	120.7
	0, 1 y 3	117 120 125	120.7
	0, 1 y 4	117 120 130	122.3
	0, 2 y 3	117 125 125	122.3
	0, 2 y 4	117 125 130	124.0
	0, 3 y 4	117 125 130	124.0
	1, 2 y 3	120 125 125	123.3
	1, 2 y 4	120 125 130	125.0
	1, 3 y 4	120 125 130	125.0
	2, 3 y 4	125 125 130	126.7
4	0, 1, 2 y 3	117 120 125 125	121.8
	0, 1, 2 y 4	117 120 125 130	123.0
	0, 1, 3 y 4	117 120 125 130	124.0
	0, 2, 3 y 4	117 125 125 130	124.3
	1, 2, 3 y 4	125 125 125 130	125.0

El cálculo del número de muestras posibles de un tamaño dado, como el mismo sujeto no puede ser seleccionado más de una vez, resulta relativamente sencillo: es el número de combinaciones de los n elementos tomados en subgrupos de tamaño k, o sea, $C_{n,k}$.

Para las muestras de tamaño 2:

$$C_{5,2} = \frac{fact(5)}{fact(2)\cdot fact(3)} = \frac{5\cdot 4\cdot fact(3)}{1\cdot 2\cdot fact(3)} = 10$$

Para las muestras de tamaño 3:

$$C_{5,3} = \frac{fact(5)}{fact(3)\cdot fact(2)} = \frac{5\cdot 4\cdot fact(3)}{1\cdot 2\cdot fact(3)} = 10$$

Para las muestras de tamaño 4:

$$C_{5,4} = \frac{fact(5)}{fact(4)\cdot fact(1)} = \frac{5\cdot fact(4)}{fact(4)\cdot fact(1)} = 5$$

Veamos, qué pasa con las medias de las diferentes muestras que se pueden seleccionar para esos tamaños dados.

Al comparar el valor medio poblacional, 123.4 g/l, con las medias muestrales se observa que todas difieren de esta (las más próximas son 123.3 y 123.5 g/l). La comparación entre las medias muestrales arroja dos resultados interesantes: muestras diferentes (con igual o diferente tamaño) pueden dar lugar a medias iguales, diferentes a la poblacional, y, a medida que aumenta el tamaño muestral, es menor el rango de valores posibles de la media muestral. Esta ultima afirmación puede comprobarse buscando los valores mínimo y máximo de la media muestral para cada tamaño de muestra: se tiene que, **el valor mínimo crece y el valor**

máximo disminuye con el aumento del tamaño muestral. Esto hace que las diferencias entre las medias muestrales y la media poblacional se hagan más pequeñas en la medida en que crece el tamaño de la muestra. Como el error de muestreo no es más que la magnitud de la diferencia entre el valor real de la media y el valor obtenido por medio de la muestra, podemos concluir que este es inversamente proporcional al tamaño muestral, es decir, disminuye según se incrementa el número de elementos en la muestra.

10.5 Condiciones de una buena muestra. La muestra representativa.

Como ya dijimos, el objetivo de las técnicas de muestreo es brindar métodos que permitan la selección de muestras, de modo tal que, a partir de los resultados de las mismas, se puedan hacer inferencias válidas para toda la población en su conjunto, por consiguiente, las bases para declarar a una muestra como **buena** debe ser precisamente su capacidad de satisfacer el propósito para el cual fue diseñada. Esta condición, que no se puede medir de ninguna manera, es la que se llama **representatividad de la muestra** y significa que ella reproduce, como un modelo a pequeña escala, las características más importantes de la población, y recibe el nombre de muestra representativa.

El investigador nunca puede saber con certeza si la muestra seleccionada es representativa de la población de origen o no, pues para saberlo tendría que comprobar su representatividad a través de la comparación de los resultados muestrales con los valores poblaciónales, lo que implica el conocimiento de la población, cuando, paradójicamente, si tuviésemos ese conocimiento, no sería necesario el uso de muestras para el estudio de la misma.

Para evaluar la **bondad** de una muestra, los dos factores más importantes que se deben tener en cuenta son:

- El tamaño muestral
- Las condiciones de selección de la misma

La importancia del tamaño muestral se analizó en el acápite 10.3 y la influencia del mismo en el error de muestreo se hizo evidente en el último ejemplo desarrollado en el epígrafe anterior, por lo que no consideramos necesario insistir en ello.

Ahora bien, sobre lo relativo a la forma en que se seleccionan los sujetos sí resulta imprescindible hacer una breve explicación .

Dado que se desea que la muestra sea un modelo **a pequeña escala** de la población original, esta debe ser seleccionada de modo tal que refleje las características fundamentales de la población. Es decir, que en la muestra se debe reproducir tanto la composición como la situación de la población. Por ejemplo, si vamos a estudiar la prevalencia de una enfermedad que tiene un comportamiento diferencial entre sexos, de alguna manera tenemos que garantizar que ambos sexos aparezcan representados adecuadamente en la muestra, de forma que se puedan alcanzar los objetivos que se persiguen. Y utilizamos la expresión **adecuadamente**, porque la composición de la muestra esta muy ligada a los objetivos: si en nuestro ejemplo lo que quisiéramos conocer es la prevalencia general de la enfermedad, los sexos deben estar representados en la muestra en la misma proporción en que lo están en la población; si lo que nos interesa es comparar el comportamiento de algunas características clínicas de la enfermedad entre sexos, deben ser iguales los tamaños de las dos sub- muestras (masculina y femenina). Eso significa que, para la selección de la muestra se deben tener en cuenta determinados principios o criterios, que respondan a las condiciones y objetivos específicos del problema que se desea estudiar en una población dada.

Una buena opción de apoyo a la **bondad** de la muestra es la obtención de algunas evidencias colaterales de la posibilidad que la misma tiene de serlo. Con esto queremos decir, obtener información sobre la muestra que

nos haga confiar en que sus características o su composición no difieren demasiado de las de la población.
¿Cómo se puede hacer esto?. Fundamentalmente, describiendo de forma detallada la composición de la muestra en función de las variables que puedan tener alguna asociación con el fenómeno en estudio: distribuciones atendiendo a sexo, edad, zona de residencia, grupo racial, etc., y comproban- do que algunos de los hechos relacionados con ese fenómeno se comportan de manera similar en la muestra y en la población, por ejemplo, la morbili- dad por una causa en particular.
En resumen, el tamaño de la muestra por sí solo no sirve de garantía para que se pueda hacer una inferencia a la población de los resultados obtenidos en la muestra. Es necesario, además, que la misma sea seleccionada según determinados principios o criterios que nos auxilien en la tarea de búsqueda de representatividad de la misma. De no hacerlo así, lo más probable es que no resulte válido el uso de la muestra obtenida o, por lo menos, que las inferencias que se hagan a partir de la misma resulten cuestionables.
Para obtener una **buena muestra**, una que probablemente sea representativa de la población según los fines de una investigación en particular, se utilizan métodos o técnicas más o menos sofisticados para la selección. Estos métodos de muestreo se abordan a continuación.

10.6 Tipos de muestreo. Esquemas muestrales básicos según el tipo de muestreo. Muestras complejas.

Al explicar el error de muestreo no se hizo alusión a la forma de selección de los individuos, por lo que debe estar claro que este error ocurre siempre que se trabaja con muestras. En consecuencia, para el que está obligado a usar esta forma de trabajo, sería muy conveniente tener alguna manera de evaluar la magnitud de dicho error, aunque fuese solo de manera aproximada.
Existe una clasificación de los esquemas muestrales en dos tipos, que atiende a si ellos permiten la valoración del error de muestreo o no. Esta es la siguiente:

- Muestreo no probabilístico.
- Muestreo probabilístico
Los esquemas de muestreo que **no permiten la valoración del error de muestreo se ubican en el grupo correspondiente al muestreo no proba- bilístico** y los que **si en el muestreo probabilístico**. La razón para que el calificativo haga referencia a las probabilidades es la siguiente: se dice que el muestreo es probabilístico, si la selección de las unidades muestrales se realiza utilizando un esquema muestral basado en las probabilidades (medida de las posibilidades) que tienen los sujetos de la población en formar parte de la muestra, y, que es no probabilístico, si se emplea otro tipo de criterio, siendo precisamente esta la razón por la cual no se puede valorar el error de muestreo.
Existen diferentes formas de muestreo no probabilístico entre los que mencionaremos los cuatro siguientes, como formas usuales de obtener muestras de esta clase:
- Muestras fortuítas o de voluntarios.
Se trabaja mucho con este tipo de muestras en arqueología, historia, etc,. e incluso en medicina. Consiste en estudiar los casos que de manera fortuita llegan al investigador o en usar sujetos que se presten para realizar el estudio. Su uso está completamente justificado, aun en el campo de la medicina, por las razones siguientes: ¿es acaso posible obtener una muestra de un tamaño adecuado de enfermos cuyo padecimiento tiene una prevalencia de 1 por cada 10 000 habitantes o más?, ¿es ético o aceptable por parte de los sujetos el que se les inocule un virus que produce una enfermedad con la finalidad de comprobar una vacuna?, ¿cualquier persona

se prestaría a ser observada mientras se entrega a actividades sexuales, aunque conozca que el objetivo de la observación es estudiar el comportamiento o la respuesta de su organismo?. En tales situaciones no hay otra opción posible.

- **Muestreo por selección de experto (muestreo opinático o al juicio)**.

Es una técnica usada por expertos con la intención de seleccionar **especímenes típicos o representativos** de un fenómeno en particular, con fines fundamentalmente experimentales. Su debilidad radica en que, varios expertos pueden tener puntos de vista diferentes sobre la mejor manera de seleccionar o caracterizar a esos casos **típicos o representativos**. En la medida en que los criterios empleados se aparten de la subjetividad, los resultados que se obtengan serán más confiables.

- Muestreo por cuotas.

Consiste en la selección de cantidades específicas de sujetos, proporcio- nales al tamaño de la población, sobre la base de algunas características demográficas. Este tipo de muestreo se aplica con frecuencia en estudios de mercado, opiniones, gustos, etc.

- Muestreo de poblaciones móviles (método de captura - marcaje-recaptura de insectos, peces, aves, etc).

Es muy utilizado en algunas ramas de la Zoología, sobre todo para estudiar el comportamiento migratorio de los animales y la estimación del tamaño poblacional. Existen modelos teóricos muy ingeniosos para justificar este método.

Es cierto que estos métodos adolecen de varios defectos, el fundamental es la imposibilidad de estimar el error de muestreo, pero siguen siendo de gran utilidad en el caso de la existencia de problemas cuyo estudio solo es posible realizarlo mediante muestras de este tipo. Además, si la selección muestral se organiza y realiza con rigor, sobre una base teórica bien justificada, los resultados son válidos, por eso es justo decir que muestra no probabilística no es equivalente a muestra no representativa.

Los esquemas o diseños muestrales probabilísticos permiten la valoración del error de muestreo, por hacer uso de la teoría de la probabilidad (la que ya se trato en el capítulo 9), y es por eso que son los más usados en los trabajos de carácter investigativo o científico. Sobre ellos se ha escrito una cantidad enorme de libros. Aquí solo se trataran los cuatro esquemas básicos siguientes:

- Muestreo Aleatorio Simple (m.a.s).

- Muestreo Sistemático (m.s).

- Muestreo Aleatorio Estratificado (m.a.e).

- Muestreo por Conglomerados (m.c).

De cada uno de ellos se veran, sus características fundamentales y la forma de empleo.

Muestreo Aleatorio Simple (m.a.s).

Este es el esquema muestral más simple, aunque se debe aclarar que él, por sí solo, se emplea muy raras veces.

La esencia de este diseño muestral radica en, que todos los sujetos en la población deben tener la misma probabilidad de ser seleccionados y que no existan diferencias marcadas, para la característica en estudio, entre subgrupos poblacionales. Los pasos para llevar a cabo este esquema son los siguiente:

1.- Enumerar consecutivamente todos los elementos de la población a muestrear.

2.- Seleccionar los elementos que componen la muestra empleando algún mecanismo aleatorio.

Tratemos de explicarlo usando un ejemplo: Supongamos que se desea conocer la estatura media de una población compuesta de 260 atletas. En primer lugar, necesitamos una relación con los nombres de dichos atletas y a cada uno de ellos se les asigna un número de orden. Si el tamaño de muestra (que se calcula por la

formulación correspondiente a este tipo de muestreo y las características de la estatura en estos sujetos), resulta ser de 30, se realiza un **sorteo** a partir del cual se obtienen 30 números entre 1 y 260 (ó 0 y 259). Los individuos cuyos números hayan salido en el sorteo serán los que integrarán la muestra.

Este diseño muestral tiene el gran inconveniente de que es necesario tener como marco muestral una lista de todos los elementos de la población, cosa que rara vez es posible.

El **sorteo** en realidad no presenta grandes dificultades de realización, aunque si puede resultar laborioso. Para facilitarlo se usan las llamadas **tablas de números aleatorios**, de las cuales tenemos un ejemplo en la **Tabla F** en los anexos.

Para utilizar esas tablas el procedimiento es el siguiente:

Se consideran todos los números de identificación de los individuos del mismo **tamaño**, es decir, que poseen la misma cantidad de dígitos, comenzando por el cero. (Hay que comenzar numerando con el 0 porque en las tablas de números aleatorios aparece el 0).

Se selecciona aleatoriamente alguna hoja de la tabla. Esto puede hacerse por medio de una tirada de dados, abriendo indistintamente el libro que la contiene por cualquier pagina, etc. Esto significa dejar la respuesta al azar.

En dicha hoja se selecciona, también al azar, alguna columna de números.

Se comienza a leer la columna, atendiendo a números de tres dígitos (en este caso), y los primeros 30 números que aparezcan, con valores entre 000 y 259 son los seleccionados. Se puede continuar la lectura en la misma columna de la página siguiente a la seleccionada o en la columna siguiente de la misma página, haciendo la elección al azar.

Cuando la cantidad de individuos que se desea seleccionar para integrar la muestra es grande, este procedimiento resulta muy engorroso, pero en la actualidad esto puede realizarse en forma automatizada empleando programas computacionales que generan números aleatorios.

Muestreo Sistemático (m.s).

Este diseño viene a ser una variante del anterior. En este caso se parte de que, además de **conocer** a todos los individuos que integran la población, se puede obtener una lista o marco muestral en el cual el ordenamiento de los mismos guarda cierta relación con la característica en estudio, o que existe un **desorden total** en relación con dicha característica.

Por ejemplo, si lo que nos interesa es el salario medio, el ordenamiento podría ser una lista de los trabajadores por categorías ocupacionales, (las que están relacionadas con el salario). En esa lista aparecerían, primero, los grupos de menor salario y, por último, los grupos de salarios mayores. En ese caso, garantizariamos tener en la muestra todos los salarios si selec- cionamos los sujetos de la forma siguiente: si el tamaño de la población es 280 y el tamaño de muestra calculado es de 35, se calcula el cociente del tamaño de la población entre el tamaño muestral 280/35, que es 8, se selec- ciona aleatoriamente un número entre 1 y 8, ambos inclusive, al que se llama raíz o arranque aleatorio. Ese será el número del primer individuo selec- cionado y para obtener los 34 números restantes, se suma a la raíz, sucesivamente, el número 8. Supongamos que la raíz que se seleccionó fue 4, los otros números serían: 4+8=12, 12+8=20, etc., el último número seleccionado en este caso sería el 276.

Garantizar el **desorden total** en relación con la característica en cuestión resulta algo bastante más difícil. Quizás podría resolverse organizando la lista por orden alfabético, por lugar donde reside, etc. Este requerimiento se debe a que si en el ordenamiento de los sujetos existe algún tipo de sistematicidad vinculada con lo que se desea estudiar, esta forma de selección puede inducir a error. Veámoslo con un ejemplo: si se

quiere estudiar el número promedio diario de consultas de Pediatría y las causas de las mismas en cierto tipo de servicio y se dispone de la lista de consultas brindadas diariamente durante todos los meses del año, este método de selección podría conducir a que todos los días (o la mayoría de ellos) que integren la muestra sean el mismo día de la semana, y esto sesgaría la información, ya que se sabe que la cantidad de pacientes que acude a la consulta varía de acuerdo con los días de la semana.

Este diseño, además del inconveniente de la lista de todos los elementos de la población, requiere el orden o desorden total de los mismos de acuerdo con alguna característica relacionada con el fenómeno en estudio, por lo que su utilidad, por si solo, resulta muy limitada.

Muestreo Aleatorio Estratificado (m.a.e).

Para casos en los cuales se conoce o se sospecha la existencia de diferencias significativas entre subgrupos poblacionales para la característica en estudio, fue que se creó este tipo de diseño de muestras.

La idea básica en la que se apoya este esquema es la posibilidad de dividir la población en subgrupos tales que, dentro de cada subgrupo, los elementos que lo integran sean similares entre sí, mientras que de un subgrupo a otro hay diferencias notables, en cuanto a la característica en estudio. **A estos subgrupos se les llama estratos**. Un ejemplos de esta forma de división de la población es el siguiente:

a) Se conoce que los índices de parasitismo y EDA están relacionados estrechamente con las condiciones sanitarias de vida, luego la población a estudiar se podría clasificar en dos grandes grupos: **estrato urbano** y **estrato rural**.

En el ejemplo, la subdivisión en estratos se puede hacer de antemano, es decir, antes de realizar la selección de los individuos, pero existen casos en que, eso no es factible. Cuando no se puede estratificar **a priori**, se realiza esta estratificación **a posteriori**, con ciertas limitaciones.

El procedimiento que se debe seguir en este diseño es como sigue:

Se divide la población en estratos y para cada estrato se hace una lista de todos los elementos (marco muestral por estratos).

El tamaño de muestra calculado se divide entre los estratos formados, atendiendo a algún criterio de asignación, de modo tal que todos los estratos aporten sujetos a la muestra.

Se aplica el m.a.s (o el m.s) en cada uno de los estratos.

Los criterios de asignación de los tamaños de muestra a los estratos en el m.a.e son varios y su selección depende del interés y las necesidades del estudio. Aquí hablaremos del más sencillo y **lógico** de acuerdo con el sentido común: **la asignación proporcional**.

Esta forma de asignación consiste en:

Primero: Hallar el peso relativo de cada estrato en la población, es decir, si N_1, N_2, N_3,.....y N_k, representan los tamaños de los estratos en la población, y N el tamaño de la misma, entonces el peso relativo de cada estrato vendría dado por:

$$P_j = \frac{N_j}{N} \text{ para } \quad j \text{ desde } \quad 1 \text{ hasta } \quad k$$

Segundo: Si el tamaño de muestra calculado es n, entonces en cada estrato se tomará como tamaño muestral (n_j) el resultado del producto:

$n_j = Pj \cdot n$ para j desde 1 hasta k

Al hacer las aproximaciones en estos tamaños muestrales se debe tener cuidado de que al final se cumpla que:

$n <= n_1 + n_2 + ...+ n_k$

ó sea, el tamaño de muestra total calculado debe ser menor o igual a la suma de los tamaños de muestra en los estratos. Por ello, es aconsejable aproximar siempre por exceso esos tamaños.

Este esquema muestral también tiene el inconveniente de la lista de elementos en la población, dificultándose aún más esto con la división en estratos de los mismos. Imagine cuan difícil puede ser obtener listas de individuos según la edad, por ejemplo.

Muestreo por Conglomerados (m.c).

Hasta ahora, todos los diseños muestrales vistos parten del conocimiento de todos los elementos de la población, y esto no sucede con mucha frecuencia. Aun en el caso de querer estudiar la población de un lugar en el que recientemente se ha realizado un censo de población, resulta muy difícil obtener una listade los habitantes del mismo. Este diseño viene a solucionar, en cierta medida, esa dificultad.

Para utilizar este esquema también se parte de una división de la población en subgrupos, pero en este caso, en lugar de buscar homogeneidad dentro de esos subgrupos lo que se pretende es todo lo contrario: que todos y cada uno de esos subgrupos muestren la misma variabilidad de la característica en estudio que la población, es decir, que cada uno sea **representativo** de la población original, en relación con esa característica. **A estos subgrupos se les da el nombre de conglomerados**. El tamaño de muestra total se divide entre los conglomerados que se seleccionen para la muestra, por medio de criterios de asignación similares a los que son utilizados en el muestreo aleatorio estratificado (m.a.e).

Para este diseño el procedimiento a seguir sería:

Dividir la población en conglomerados (generalmente zonas geográficas).

Seleccionar un cierto número de conglomerados, por algún procedimiento aleatorio.

Seleccionar los **sujetos**, dentro de los conglomerados elegidos, según los tamaños de muestra asignados a cada uno de ellos, empleando m.a.s o m.s.

Dentro de este esquema muestral se pueden realizar subdivisiones sucesivas de la población. Por ejemplo, dividir el país en provincias, las provincias en municipios y los municipios en áreas de salud. Si seleccionamos provincias, dentro de estas municipios y en los municipios seleccionamos áreas de salud, como se han hecho tres selecciones antes de llegar a los individuos, el diseño recibe el nombre de por conglomerados trietápico. Si solo seleccionamos las provincias y dentro de estas los municipios, antes de seleccionar a los individuos, sería por conglomerados bietápico.

La ventaja que esto ofrece es que no necesitamos una lista de toda la población, sino solo de una parte de ella, lo que resulta mucho más fácil de lograr. En el ejemplo de muestreo por conglomerados trietápico del párrafo anterior, solo sería necesario obtener la lista con los nombres de las provincias, las de los municipios en las provincias seleccionadas y las listas de las áreas de salud de los municipios seleccionados en la segunda etapa de selección. El marco muestral se reduciría a las listas de las áreas seleccionadas.

Muestras complejas.

En la práctica, ocurre muchas veces que no se justifica la aplicación de estos esquemas simples de muestreo por si solos, ya que en la mayor parte de los casos, la complejidad del trabajo es tal, que se requiere tener en cuenta no solo la característica que se desea estudiar, sino también la logística del estudio. Este es el caso de los Estudios Nacionales de Crecimiento y Desarrollo Humano, que requieren de diseños muestrales más complejos.

Veamos brevemente, el esquema muestral utilizado en la investigación de este tipo realizada en Cuba en 1982.

Dada la importancia vital de la determinación de las características en estudio para ambos sexos por separado, teniendo en cuenta la edad, y por el carácter transversal del mismo, se trabajó con 58 clases determinadas por los dos sexos y 29 grupos de edad (subpoblaciones de edad y sexo). Como se sabe que las condiciones socioeconómicas y el nivel de urbanización están relacionados con el crecimiento y desarrollo, a fin de asegurar la presencia en la muestra de las distintas provincias del país y las zonas urbanas y rurales de estas, se realizó la estratificación de dichas subpoblaciones según provincia y zona (urbana o rural), de acuerdo con la división político-administrativa existente en el país en 1972 (para que se pudiesen realizar comparaciones con los resultados del estudio realizado en esa fecha). Estos estratos fueron, a su vez, subdivididos en conglomerados a los que se dió el nombre de macrodistritos y que estaban integrados por 5 ó 6 distritos de los demarcados por el Censo de Población y Vivienda. La selección de los macrodistritos se realizó por muestreo sistemático dentro de cada estrato, con asignación proporcional al tamaño.

En cada macrodistrito seleccionado se confeccionó una lista de los individuos menores de 20 años y las embarazadas de más de 20 semanas de gestación (para poder recoger datos de recién nacidos) y con ellos se confeccionaron los marcos muestrales de los 58 estratos de sexo y edad, a partir de los cuales se realizó la selección de los individuos, aplicando una variante de muestreo sistemático, de modo tal que dentro de cada clase se obtuviera un tamaño muestral acorde con los requerimientos de precisión establecidos.

La necesidad de utilizar muestras complejas es frecuente. Solo en contadas ocasiones se puede hacer uso de los esquemas básicos del muestreo probabilístico por sí mismos. Por ello, es tan importante que el investigador, una vez definido su problema de investigación y los objetivos que persigue, busque asesoría sobre todo en relación con la selección muestral y el procesamiento posterior de los datos. De otra forma corre el riesgo de recoger información sobre un grupo de sujetos, (con los gastos consiguientes de tiempo y materiales), que luego no le sea real- mente útil para los fines de su estudio.

No siempre una muestra probabilística es mejor que una no probabilística, además de que, en muchas ocasiones, la única forma de realizar el estudio es a través de muestras del segundo tipo. Por ejemplo, si quisiéramos describir las características y condiciones en que se producen los nacimientos en un determinado pueblo, ciudad, región o país, parece ser un diseño mejor y más factible registrar los datos de todos los nacimientos que se produzcan en un determinado período relativamente breve (lo que nos permitiría concentrar los recursos y esfuerzos), que seleccionar una muestra aleatoria de **nacimientos**, durante un lapso de tiempo más largo, lo que conllevaría más problemas logísticos y de recursos, sin aportar una mayor **credibilidad** a los resultados.

La ventaja de la muestra aleatoria radica, fundamentalmente, en que nos permite evaluar a **posteriori** los errores o desviaciones que pudieran producirse o presentarse al **estimar** las caracteristicas poblacionales a través de la muestra. Sin embargo, debemos tener muy en cuenta que siempre el criterio del investigador está presente en mayor o menor medida; ya sea al seleccionar la muestra o al determinar el alcance en tiempo y espacio de sus resultados. De todas formas, el que una muestra sea adecuada o no va a estar determinado por la utilidad (práctica, teórica o ambas) de la infor- mación que ella brinda.

Tanto los muestras probabilísticas como las no probabilísticas pretenden ser representativas de la población objeto de estudio, pero muchos investigadores califican sus muestras como representativas, cuando quieren expresar que son probabilísticas, y hacen equivalentes, por error, ambos términos. Obviamente, esto sería un logro para el investigador y garantizaría, en cierta medida, la validez de los resultados de la investigación, pero tal afirmación solo podría hacerse si se conocieran las características de toda la población, lo cual no es posible. Además, si conocemos la población, ¿para que estudiámos una muestra de la misma?. No podemos

asegurar que una muestra es representativa jamás. Lo más que podemos hacer es comparar algunas características generales conocidas de la población (pirámide poblacional, composición racial, etc.) con las de la muestra, para decir si la muestra estudiada tiene un comportamiento similar al de la población para esas características, y, por ende, puede ser que la represente adecuadamente.

En muchas ocasiones un criterio de expertos puede lograr una muestra con una mejor representatividad de la población objeto de estudio que una muestra aleatoria diseñada y seleccionada por personas que no conocen con profundidad el tema que se realiza, o que conociéndolo, hacen la selección de manera puramente aleatoria, sin tener en cuenta los conocimientos que se tienen sobre el asunto, y corren así el riesgo de perder en representatividad.

Si se realiza un experimento con un pequeño grupo de pacientes enfermos de cierta modalidad de cáncer, seleccionados de acuerdo con el criterio de experto del investigador, para comprobar la efectividad de un nuevo medi- camento o un nuevo tratamiento, la representatividad o no de la muestra depende de la rigurosidad y objetividad con que se usó el criterio del experto y de su fundamento teórico.

Además, por encima de todos estos criterios técnicos se encuentra la factibilidad o no de un estudio con un tamaño de muestra prefijado y utilizando un esquema muestral aleatorio más o menos complicado, dadas las condiciones materiales existentes y el nivel del conocimiento científico del problema en cuestión, en ese momento.

10.7 Ejercicios propuestos.

Nota: en todos los casos explique el por qué de su decisión.

1.- Sugiera cómo tomar una muestra aleatoria de 100 sujetos entre los estudiantes de un determinado centro escolar de nivel primario, con el objetivo de conocer su desarrollo psicomotor.

2.- Dé un ejemplo de una población real para la cual usted crée que el m.a.e resulte considerablemente más barato y mejor que el m.a.s o el m.s.

3.- ¿ Qué tipo de esquema muestral emplearía usted para tomar una muestra de farmacias en una provincia, con el objetivo de conocer cuáles son los medicamentos más indicados por los médicos?.

4.- El director de un área de salud desea hacer un estudio clínico- epidemiológico de los asmáticos menores de 15 años. Se sabe que estos pacientes están dispensarizados, por lo que resultaría relativamente fácil obtener una lista con el nombre y dirección de los mismos. ¿Qué tipo de esquema muestral le aconsejaría usted que utilizara?.

CAPITULO 11.- INFERENCIA ESTADISTICA

11.1.- INTRODUCCION

La Inferencia Estadística es la parte de la Métodos Estadísticos relacionados con los problemas de orden práctico que tienen que ver con el hecho de tomar decisiones en situaciones de incertidumbre. Por consiguiente, el objetivo fundamental de este capítulo será el de introducir a los lectores en el estudio de dos técnicas básicas de la Inferencia Estadística, de amplio uso en el campo de la biomedicina.

Generalmente, como se dijo, el proceso de toma de decisión se produce en presencia de la incertidumbre, y esto, como se vera mas adelante, sólo será posible a través de un razonamiento inductivo relacionado con valores de probabilidad. Es decir, se afirma que se pueden hacer inferencias inseguras, en las que el grado de incertidumbre es susceptible de medición siempre que se haya realizado un experimento regido por determinados principios.

Mostremos a continuación tres situaciones básicas que sirven como ilustración de lo que mas adelante desarrollaremos:

Situación problemica 1. Supongamos que se descubre una vacuna que cura una enfermedad *x*. El gobierno de un país está dispuesto a adquirirla, pero desconoce el número de enfermos de *x* en su población. No existen los recursos materiales y humanos para hacer un censo, y se decide realizar una encuesta por muestreo probabilístico. A partir del número de enfermos en la muestra se infiere el número de enfermos en la población total y se determina la cantidad de vacunas que se deberá comprar.

Este constituye un ejemplo de **inferencia inductiva**; a partir de una muestra probabilística de la población objeto, se generaliza el resultado a dicha población. Este proceso de generalización es totalmente inútil si no abarca también la evaluación de la validez que tiene el conjunto de datos, para ofrecer la información deseada.

Para cumplir con tales objetivos la Inferencia Estadística utiliza técnicas que, tradicionalmente, se agrupan en dos categorías:

a) las relacionadas con la estimación de parámetros y
b) las dedicadas a la prueba o contraste de hipótesis.

Situación problemica 2. Supongamos que estamos interesados en realizar un estudio para describir las características del desarrollo físico en niñas cubanas entre 8 y 8.9 años de edad, por medio de la observación de algunas dimensiones antropométricas.

Una forma adecuada y útil de describir dichas características sería seleccionar distribuciones de probabilidad teóricas tan cerca de las observadas como sea posible. Al proceder estadístico, mediante el cual, se llevan a cabo los cálculos necesarios, se le llama **estimación**.

Situación problemica 3. Consideremos de nuevo la situación anterior y asumamos, ahora, que se conoce que la distribución de la talla en la población base es normal con parámetros μ_0 = 126.9 cm y que la talla promedio en una muestra simple aleatoria de n niñas de esa edad que residen en la provincia de Ciudad de La Habana es 128.8 cm. Este resultado sugiere al investigador al menos dos preguntas de interés: ¿la diferencia encontrada de 1.9 cm a favor de las niñas de la provincia Ciudad de La Habana estará indicando que las niñas de 8 años de la capital del país son más altas, producto quizás de un medio que les permite expresar mejor su potencial de crecimiento? o sencillamente, ¿la diferencia será producto del error del muestreo?

Para dar respuesta a estas interrogantes habrá que seguir una técnica estadística que se conoce como **prueba o contraste de hipótesis**.

11.2.- ESTIMACION. ESTIMACION PUNTUAL

Ya establecimos que el procedimiento básico para poder estudiar cierta propiedad o variable aleatoria tal como la talla, el coeficiente inteligencia, el padecer o no de tuberculosis, etc, en una población objeto sin hacer un censo, es tomar una muestra probabilística de tamaño n (desde ahora, siempre que hablemos de muestras deberá entenderse que nos referimos a muestras de este tipo, pues con ellas se puede medir el error de muestreo. De otra forma no se cumple con la condición para poder aplicar las técnicas de la Inferencia Estadística) y basándonos en la inspección de los elementos de la muestra, generalizar lo que deseamos decir acerca de la propiedad de interés a toda la población. Se sabe que tal afirmación no puede ser cierta con certeza, pero al menos lo será en un sentido probabilístico. Es decir, que podemos hablar de una afirmación razonable.

Usualmente, si la variable aleatoria a estudiar es continua, una ley de distribución que puede describir adecuadamente su comportamiento es la normal.

Es conocido que los parámetros que identifican la función de distribución normal son μ y $\square^2$ y que su expresión gráfica es una curva acampanada y simétrica; a cada par de valores de los parámetros corresponderán curvas diferentes. Entonces, la estimación consiste en seleccionar cuál de las infinitas curvas se ajusta mejor al comportamiento de los datos de la muestra que estamos analizando.

Para cumplir con este objetivo habrá que encontrar expresiones que permitan, mediante el proceso de sustituir en ellas los valores observados de una variable aleatoria X en una muestra, el calcular una cifra que sirva como una buena aproximación al valor desconocido del parámetro o de los parámetros en la población. Dicho de otra forma, hay que encontrar funciones matemáticas en las que, al ser sustituidos los valores muestrales, se hallen cifras que estimen de forma satisfactoria el valor del parámetro en cuestión.

Definición 1. Se llama **estimador** $\hat{\theta}$ de un parámetro desconocido θ, a cualquier función que dependa únicamente de los elementos de una muestra.

Definición 2. La cifra numérica o valor observado del estimador, obtenida por sustitución de los valores muestrales en la expresión de $\hat{\theta}$, es lo que en la terminología estadística se denomina como **estimación** del parámetro θ.

Ejemplo 1: Consideremos como continuación la segunda de las tres situaciones problemicas iniciales:

Asumamos con bastante seguridad que la variable X, talla, se distribuye en la población de acuerdo con una ley de distribución normal cuyos parámetros μ y $\square^2\square$, se suponen desconocidos, lo expresado es común escribirlo en la notación habitual estadística como sigue: $X \sim N(\mu, \square^2)$.
Supongamos para continuar que se ha tomado una muestra de tamaño n = 90 y queremos estimar la talla media y la desviación estándar. Denotemos por $x_1, x_2, ..., x_{90}$ los valores correspondientes a la talla en centímetros de cada una de las 90 niñas de la muestra, de estadística descriptiva conocemos que las fórmulas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{y} \quad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (1)$$

se utilizan para calcular la media y la varianza muestral. Estos son precisamente los **estimadores** de los parámetros μ y σ^2 ya que tanto $\bar{x}$ como s^2 se obtienen a través de funciones que dependen de las x_i , es decir, dependen de la talla de cada una de las 90 niñas seleccionadas en la muestra. Si al realizar los cálculos apropiados se obtiene que $\bar{x}$ = 126.9 cm y s = 6.15 cm, entonces esas cifras son las **estimaciones** de la media y la desviación estándar poblacionales, o sea, de μ y σ.

Notemos que la primera suposición que se hizo fue sobre el tipo de ley de distribución de la variable aleatoria talla en la población. Sin hacer esa suposición no hubiese sido posible resolver el problema de estimación. Después se hizo la selección de la muestra y se sustituyeron los valores en (1). Este es usualmente el procedimiento a seguir para hacer una estimación.

Ejemplo 2. Supongamos que en un determinado municipio de la Ciudad de La Habana donde residen 25,000 habitantes, se quiere estimar la prevalencia[3] de asma bronquial en una fecha dada. Con ese fin se selecciona una muestra de tamaño 100, detectándose 8 asmáticos. Halle la estimación de la prevalencia, suponiendo que la **variable asmático** se distribuye según la ley de distribución binomial.

Solución.

La prevalencia estaría dada por el número total de habitantes del municipio enfermos de asma bronquial en la fecha dada. Por dato tenemos que N = 25,000, luego, si tuviéramos el valor del parámetro π de la ley de distribución binomial, que en este caso es la probabilidad o frecuencia teórica de estar enfermo de asma bronquial, podríamos conocer la prevalencia a través del producto $N\pi$. Para estimar π en base a los datos de la muestra se usa la expresión $\hat{p}$ = k/n, donde :

k- número total de enfermos de asma bronquial en la muestra

n- tamaño de la muestra

Como la muestra es de tamaño 100 y el número de asmáticos 8,

$\hat{p}$ = 8/100 = 0.08

es la probabilidad estimada que tiene un habitante del municipio, de ser asmático. La prevalencia estimada es 25,000 · 0.08 = 2,000 habitantes.

Debe notarse que la utilidad práctica del estimador, o estadígrafo como también suele llamarse, radica en que por medio de un proceder de cálculo se obtiene un valor único, la estimación. En este sentido, tanto uno como lo otro constituyen valores **puntuales**. Constituye, en este esquema, un aspecto esencial la selección de muestra, con la que, por sustitución de los valores observados en la expresión del estimador, hallamos un valor numérico (una estimación) que debe corresponder a un parámetro poblacional bajo estudio, descriptor de una propiedad de interés. Luego , por el momento lo que tenemos son **estimaciones puntuales**.

Este hecho, unido a lo que conocemos acerca de la incertidumbre que se produce en el proceso de selección de muestras aleatorias, deja en dudas la utilidad de la estimación puntual, ya que continuamos sin ninguna información en relación con cuán cerca está el valor encontrado del verdadero valor desconocido del

[3] La prevalencia de una enfermedad x es el número de enfermos de x en una población y en un momento dados.

parámetro poblacional. Es decir, sabemos que va a existir una diferencia entre la cifra estimada y la verdadera, pero no conocemos todavía si tal diferencia es admisible o no.

Una propiedad del estimador, que mejora su utilidad, se enuncia a continuación.

Definición 3. El estimador $\hat{\theta}$ de un parámetro desconocido θ se dice que es **insesgado,** si al extraer infinitas muestras, todas de tamaño n, el promedio aritmético de las estimaciones coincide con θ, el valor del parámetro desconocido.

Es decir, que un estimador cumple con la propiedad de ser insesgado si al repetir un número grande de veces la extracción de muestras aleatorias de tamaño n de la población base, la media de las estimaciones, tomada sobre todas las repeticiones del experimento **extracción de muestras aleatorias de tamaño n**, coincide con el valor del parámetro en la población.

Luego, si un estimador puntual de un parámetro desconocido es insesgado, al menos esperamos que, en promedio la estimación coincida con el verdadero valor del parámetro. Significa, además, que la desviación del estimador con respecto al parámetro es, en promedio, pequeña.

Los estimadores $\bar{x}$, s^2 y $\hat{p}$, hoy día se conoce, son estimadores insesgados de μ, σ^2 y π, respectivamente. El ejemplo siguiente permitirá al lector una mejor percepción acerca de la utilidad de trabajar con estimadores insesgados.

Ejemplo 3. En los meses de marzo a mayo de 1973, se realizó en Cuba la Investigación de Mortalidad Perinatal que abarcó, entre otros, a todos los nacidos durante la semana del 1 al 7 de marzo. Con los sobrevivientes a los 7 días de edad, producto de partos simples, se decidió formar una cohorte y medirla en etapas claves de su desarrollo físico, educacional e intelectual.

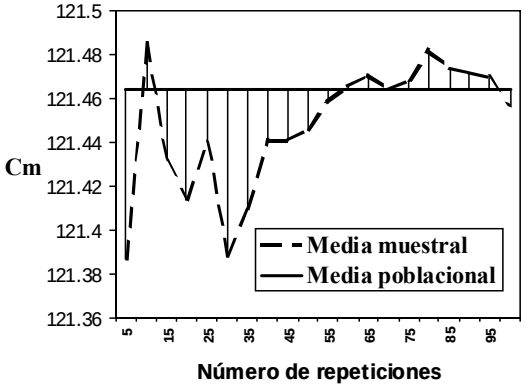
Del seguimiento realizado en 1980, cuando los niños de la cohorte tenían 7 años, vamos a trabajar con los datos correspondientes a la talla de las 1760 niñas. La media de esta dimensión antropométrica en toda la población fue de 121.46cm.

Para estudiar el comportamiento del estimador, $\bar{x}$, se tomaron 100 muestras aleatorias de tamaño 200.

En la figura 1 que aparece mas adelante se muestran los promedios de las medias muestrales agrupadas en repeticiones de 5 en 5 hasta 100. Es decir, se seleccionaron las 5 primeras muestras y se computó el promedio de sus medias; después se repitió la operación pero añadiendo las segundas 5 muestras, para obtener el promedio sobre la base de 10 repeticiones y así sucesivamente hasta tener 100 muestras de tamaño 200. La figura muestra que las desviaciones de los promedios de las medias muestrales con respecto a μ disminuyen en la medida en que el número de repeticiones se incrementa; a partir de 40 repeticiones esas desviaciones se mantienen en el orden de ± 0.02cm.

Por otra parte, cuando se realiza una estimación se hace trabajando con una sola muestra, por lo que el hecho de que un estimador puntual sea, además, insesgado sigue sin brindar ninguna información respecto a la seguridad de cuán cerca o lejos se está, para esa muestra en particular, del verdadero valor del parámetro.

Figura 1. Desviaciones entre la media de la estatura de niñas de 7 años y el promedio de las medias muestrales. Muestras de tamaño 200 extraídas de la población de 1760 niñas.



Ejemplo 4 (continuación). Consideremos que las medias correspondientes a las 5 primeras muestras fueron: 121.633, 121.272, 120.908, 121.838, 121.283 y sus diferencias con respecto a μ: 0.169, −0.192, −0,556, 0.374 y 0.181 respectivamente; si se hubiera trabajado con la muestra #1 se habría sobreestimado el valor desconocido en aproximadamente 2 milímetros, si la escogida hubiese resultado la muestra #3 entonces se habría subestimado el valor del parámetro en más de medio centímetro, pero la situación es hipotética. En la práctica tendríamos un desconocimiento total acerca del monto de esas diferencias.

Entonces, podemos concluir que una estimación puntual de un parámetro desconocido, aún y cuando sea insesgada, no nos proporciona todavía suficiente información, pues hay que tener en cuenta el error que se comete en el proceso de selección de la muestra. Una solución mejor del problema, que incluye dicho error, se brinda en el próximo epígrafe.

11.3.- ESTIMACION POR INTERVALO DE CONFIANZA.

De acuerdo con lo visto hasta aquí, es necesario acompañar la estimación de su error muestral. Parece bastante natural construir un intervalo centrado en la estimación $\hat{\theta}$, de la forma: $\hat{\theta}$ – d, $\hat{\theta}$ + d, donde d incluya el error debido al muestreo. Es decir, deseamos construir un intervalo $(\hat{\theta}_1, \hat{\theta}_2)$ con $\hat{\theta}_1 < \hat{\theta}_2$, como función de los valores observados en la muestra y afirmar que el parámetro desconocido se encuentra en ese intervalo.

Por supuesto, si el parámetro no se encuentra en ese intervalo estaremos cometiendo un error y lo que deseamos es que dicho error ocurra raramente, es decir, con muy baja frecuencia o probabilidad. Por otra parte, la longitud del intervalo es una medida de cuán bien puede ser estimado el parámetro a través del intervalo, es decir, que habla de la utilidad de nuestra afirmación.

Así, por una parte, queremos que el parámetro se encuentre con gran probabilidad en el intervalo $(\hat{\theta}_1, \hat{\theta}_2)$ y, por otra, que la longitud del intervalo sea bastante pequeña.

Es importante notar por un lado, que el parámetro se considera desconocido, con valor fijo, mientras que, al menos uno de los límites del intervalo debe ser aleatorio, pues ambos dependen de los valores de una muestra aleatoria.

Definición 4. Se conoce como **intervalo de confianza** para estimar un parámetro desconocido $\hat{\theta}$, al intervalo aleatorio de la forma $(\hat{\theta}_1, \hat{\theta}_2)$ donde $\hat{\theta}_1 < \hat{\theta}_2$, tal que esperamos que contenga a $\hat{\theta}$, con una probabilidad dada.

A $\hat{\theta}_1$ se le llama límite inferior del intervalo y a $\hat{\theta}_2$ límite superior.

La estimación por intervalo de confianza puede ser, pues, aplicada para estimar cualquier parámetro de interés.

En este texto, nos ocuparemos de la construcción de intervalos de confianza para la media μ de una variable aleatoria X con distribución normal y para la proporción π de una variable aleatoria X con distribución binomial.

11.3.1.- INTERVALO DE CONFIANZA PARA μ CON □ CONOCIDA.

Vamos a aplicar las nociones anteriores para hallar un intervalo de confianza para μ cuando □ es conocida.

Aunque la suposición acerca del conocimiento de □ es difícil de alcanzar en los problemas con que nos enfrentamos en la práctica, comenzaremos por este caso, que es el más sencillo y después se introducirán situaciones más próximas a la realidad.

Ejemplo 5. Vamos a suponer que se extrae una muestra de tamaño 50 de la población de tallas de niñas de 7 años a la que se hizo referencia en el ejemplo 3 (ver anexo 1). Se conoce además que □□= 5.53 cm. Los 50 valores de talla aparecen en la tabla que sigue a continuación, considere adicionalmente que la muestra es simple aleatoria:

TABLA #1

131.5, 115.0, 125.5, 119.0, 123.0, 125.0, 124.5, 122.5, 118.0, 125.0

125.5, 119.0, 124.5, 118.0, 117.0, 128.0, 119.5, 124.5, 132.0, 122.5

120.5, 120.0, 126.0, 128.0, 120.5, 120.0, 119.0, 117.0, 129.5, 124.0

121.0, 119.0, 120.0, 128.0, 128.0, 118.5, 118.0, 124.0, 118.0, 118.5

131.0, 117.0, 118.0, 116.5, 121.0, 122.0, 124.0, 120.5, 114.5, 121.0

La media en este caso es = $\bar{x}$ 122.05 cm.

Hasta aquí ¿qué podríamos decir?. Como sabemos, nuestro estimador es un 'buen estimador', en el sentido de que pertenece a la clase o tipo de los insesgados, por tanto la estimación 122.05 nos parece una buena aproximación para el verdadero valor poblacional desconocido μ. Hagamos uso ahora de lo que ya conocemos, sabemos, en primer lugar que el estimador $\bar{x}$ de μ se distribuye normal con media μ y

desviación estándar $\sigma/\sqrt{n}$ y en segundo lugar que para toda variable aleatoria X con distribución N(μ, □^{2}) se cumple que:

$$\Pr(\mu - 1.96\ \square < X < \mu + 1.96\ \square) = 0.95 \qquad (2)$$

Sustituyendo en (2), X por $\bar{x}$ tendremos:

$$\Pr(\mu - 1.96\ \sigma/\sqrt{n} < \bar{x} < \mu + 1.96\ \sigma/\sqrt{n}) = 0.95 \qquad (3)$$

Con la doble desigualdad presente en el miembro izquierdo de la igualdad anterior, se pueden realizar dos transformaciones algebraicas que conducen cada una de ellas a desigualdades equivalentes a la de partida, y por tanto ambas con idéntico valor de probabilidad 0.95.

Son estas, las transformaciones siguientes:

- ✓ Si en $(\mu - 1.96\ \sigma/\sqrt{n} < \bar{x} < \mu + 1.96\ \sigma/\sqrt{n})$, restamos μ en todos los términos, esto conduce a $(-1.96\ \sigma/\sqrt{n} < \bar{x} - \mu < 1.96\ \sigma/\sqrt{n})$. Si ahora dividimos todos los términos por el valor $\sigma/\sqrt{n}$, se obtiene una de las transformaciones anunciadas:

$$(-1.96 < \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} < 1.96) \qquad (4)$$

- ✓ Para obtener la otra doble desigualdad, debemos dividir la inicial en dos, y realizar en ambas por separado, los cambios algebraicos necesarios.
- ✓ Para ellos comencemos con, $\mu - 1.96\ \sigma/\sqrt{n} < \bar{x}$, y en ella pasemos la expresión, $-1.96\ \sigma/\sqrt{n}$, del miembro izquierdo hacia el derecho, obteniéndose, $\mu < \bar{x} + 1.96\ \sigma/\sqrt{n}$.

  De modo semejante se puede proceder con, $\bar{x} < \mu + 1.96\ \sigma/\sqrt{n}$, para obtener que, $\bar{x} - 1{,}96\ \sigma/\sqrt{n} < \mu$. Combinando ahora ambas desigualdades, estas permiten plantear que:

$$(\bar{x} - 1.96\ \sigma/\sqrt{n} < \mu < \bar{x} + 1.96\ \sigma/\sqrt{n}) \qquad (5)$$

Como ya se dijo, $\Pr(-1.96 < \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} < 1.96) = 0.95$, donde el termino, $\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}}$, usualmente identificado por Z, es la expresión aritmética necesaria para transformar una variable aleatoria X con distribución N(μ, □^{2}) en una variable aleatoria normal estándar, por lo que 1.96 y –1.96 no son más que los valores entre los que se acumula el 95 % del área bajo la curva normal estándar; dicho de otra forma, fuera de ese intervalo sólo está el 5 % del área bajo la curva.

Como la curva normal es simétrica, ese 5 % quedará igualmente repartido hacia cada extremo de la curva, hacia las llamadas colas; por debajo de –1.96 tendremos un 2.5 % y por encima de 1,96 el otro 2.5 %. Sabemos, de estadística descriptiva, que los percentiles son aquellos valores de la variable que se corresponden con valores dados de la distribución de frecuencias acumuladas relativas, por lo tanto –1.96 es, en este caso, el percentil 2.5 de la distribución normal estándar mientras que 1.96 es el percentil 97.5. Se puede poner 0.95 = 1−□, de este modo, □□= 0.05. En consecuencia, 1.96 = $z_{0..975} = z_{1-\square/2}$ y –1.96 = $z_{0.025} = z_{\square/2} = -z_{1-\square/2}$

Lo expresado en combinación con (5) permite plantear que:

$$Pr(\bar{x} - z_{1-\square/2} \sigma/\sqrt{n} < \mu < \bar{x} + z_{1-\square/2} \sigma/\sqrt{n}) = 1-\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square$$

que es exactamente lo que necesitábamos; una estimación para μ, que ya no es un valor único, donde incorporamos finalmente, al utilizar la probabilidad, una medida de confiabilidad o confianza.

Si sustituimos en (6) los datos del ejemplo 5 tendremos:

$$Pr(122.05 - 1.96 \cdot 5.53 / \sqrt{50} < \mu < 122.05 + 1.96 \cdot 5.53 / \sqrt{50}) = 0.95$$

$$Pr(122.05 - 1.96 \cdot 0.782 < \mu < 122.05 + 1.96 \cdot 0.782) = 0.95$$

$$Pr(120.52 < \mu < 123.58) = 0.95 \qquad (7)$$

Es importante en este momento dar una interpretación correcta a la expresión (7).

Hemos hallado un intervalo que podemos afirmar contiene a μ con probabilidad 1−□□= 0.95, pero como sabemos, al escoger una y sólo una de las infinitas muestras de tamaño 50, tengo que **multar** esa decisión. La forma en que se ha construido el intervalo de confianza garantiza que si se toman 100 muestras de tamaño 50; 95 de ellas aproximadamente producirán intervalos que contienen el valor real de μ mientras que alrededor de 5 producirán intervalos que no lo contienen. Ese es el tipo de afirmación que se hace cuando se habla de que tenemos un 95 % de confiabilidad o confianza de que el intervalo (120.52, 123.58) contenga el verdadero valor del parámetro μ. En términos del ejemplo utilizado diremos que, basados en una muestra aleatoria de tamaño 50 y conociendo que la desviación estándar poblacional de la talla de niñas de 7 años es 5.53 cm se estima, con un 95 % de confianza, que la media poblacional tome algún valor en el intervalo de 120.5 cm a 123.6 cm.

Con la definición siguiente el resultado anterior se generaliza:

Definición 5. Sea X una variable aleatoria normal con media poblacional μ desconocida y desvió estándar □ conocido. Se llama intervalo de confianza para μ con nivel de confiabilidad del (1-□□□·□□□□□□□ a la expresión:

$$(\bar{x} - z_{1-\square/2} \sigma/\sqrt{n},\ \bar{x} + z_{1-\square/2} \sigma/\sqrt{n}) \qquad (8)$$

donde:

$z_{1-\square/2}$: percentil de orden 1−□/2 de la distribución normal estándar,

$\bar{x}$: valor observado de la media muestral, en una muestra de la variable X y

n: tamaño de muestra;

Los valores $\bar{x} \pm z_{1-\alpha/2} \sigma/\sqrt{n}$ son, respectivamente, los límites superior e inferior de confianza del intervalo.

El valor $1 - \alpha$ se da de antemano; por supuesto, $0<1-\alpha<1$ y α pequeño. Al percentil $z_{1-\alpha/2}$ de la distribución normal estándar se le denomina **coeficiente de confianza**.

La longitud del intervalo de confianza nos da una idea acerca de la proximidad con que estimamos μ; por esta razón la cantidad $d = z_{1-\alpha/2} \sigma/\sqrt{n}$ se conoce con el nombre de **precisión** de la estimación por intervalo de confianza.

En el caso particular del ejemplo anterior, si calculamos el valor de la precisión por medio de la expresión, d, esta equivale aproximadamente a 1.5 cm. S Por ello si queremos buscar una estimación más precisa de μ, para un nivel de confianza fijo, esto equivaldría a encontrar un intervalo más estrecho, lo cual, según la citada expresión significaría incrementar el tamaño de la muestra (n).

En el ejemplo siguiente vamos a practicar el cálculo del intervalo de confianza y al mismo tiempo, vamos a comparar el intervalo obtenido con el que se da en (7).

Ejemplo 6 . Se extrajo una muestra aleatoria de tamaño 200, a partir de la población de tallas del anexo 1. Calcule un intervalo de confianza del 95 % para la media poblacional, sabiendo que: $\bar{x}$ = 121.633 cm y σ = 5.53 cm..

Compare el intervalo de confianza que halle con el obtenido en (7).

Solución:

Debemos calcular un intervalo de la forma: ($\bar{x} - 1.96 \sigma / \sqrt{n}$, $\bar{x} + 1.96 \sigma/ \sqrt{n}$).

Para ello debemos calcular el valor de $1.96 \cdot \sigma/\sqrt{n} = 1.96 \cdot 5.53 / \sqrt{200}$.

Pero $= 5.53 / \sqrt{200} = 0.391$, luego $1.96 \cdot \sigma \sqrt{n} = 1.96 \cdot 0.391 = 0.766$

En consecuencia, 121.633 ± 0.766 = (120.867, 122.399) es el intervalo en cuestión.

En este caso, decimos que los límites del intervalo de confianza del 95 % para la media de la talla poblacional son, aproximadamente, de 120.9cm y 122.4cm.

Mientras que el intervalo hallado con la muestra de tamaño 50 tenía una longitud de 3.1cm, éste tiene una longitud de 1.5cm; es decir, se redujo a la mitad. La precisión ahora es mayor, pero, a cambio, el tamaño de la muestra se cuadruplicó.

Lo que acabamos de ver en el ejemplo anterior se cumple siempre; es decir, si se conoce la desviación estándar poblacional y se mantiene un nivel de confianza fijo, para aumentar la precisión de la estimación por medio del intervalo de confianza es necesario incrementar el tamaño de la muestra.

Si, por otra parte, solamente se cambia el nivel de confianza, digamos al 99 % que se corresponde con $z_{0.995}$ = 2.58 (ver tabla de la distribución normal), el intervalo será más amplio que el correspondiente al 95 %, puesto que hay que sumar y restar a la media muestral una cifra mayor. Entonces, si mantengo fijas las condiciones del problema, mayor confiabilidad siempre va a implicar menor precisión y viceversa.

Ello es bastante comprensible; mientras que en el caso del 95 % de confiabilidad se admite que de cada 100 muestras de tamaño n, cinco aproximadamente pueden no contener el verdadero valor del parámetro desconocido, al elevar el nivel de confianza al 99 %, se espera que aproximadamente sólo una de cada 100 no contenga el verdadero valor del parámetro. Bajo las mismas condiciones esto se logra con un intervalo más amplio.

11.3.2.- INTERVALO DE CONFIANZA PARA □ CON □² DESCONOCIDA.

Es el caso que más se encuentra en la práctica. Como en la construcción del intervalo de confianza para □ siempre va a estar implicado el valor de □², será necesario estimar este último valor.

Como ya vimos $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ es un estimador puntual de □². Es, además, su estimador insesgado. Entonces, parece bastante plausible sustituir □² por s^2 en (5) y así obtener el intervalo de confianza de interés. Sin embargo, tenemos que responder a una nueva interrogante, ¿seguirá siendo apropiado usar el percentil correspondiente a la distribución normal estándar? Puesto en otra forma, vimos que la interpretación de 1.96 dependía del conocimiento que teníamos acerca del estadígrafo $(\bar{x} - \square) / (\square/\sqrt{n}) = Z$, ¿qué sucede con esta expresión cuando se sustituye □ por su estimador s?

Se conoce que en este caso Z ya no tiene una distribución normal con parámetros 0 y 1, sino que $(\bar{x} - \square) / (s / \sqrt{n})$ tiene ahora la distribución t de Student con n-1 grados de libertad.

En la tabla de la distribución t se puede ver que cuando los grados de libertad son mayores de 30, o lo que es lo mismo, para muestras de tamaño n > 31 los percentiles de la distribución t y de la N(0,1) son muy parecidos, entonces, es común considerar intervalos de confianza diferentes en el caso de □² desconocida en dependencia de si n > 30 ó n ≤ 30.

Primer caso (n > 30).

Definición 6. Sea X una variable aleatoria normal con parámetros desconocidos □, □². Si n es lo suficientemente grande, mayor que 30, el intervalo:

$$(\bar{x} - z_{1-\square/2}\, s/\sqrt{n}\ ;\ \bar{x} + z_{1-\square/2}\, s/\sqrt{n}) \qquad (9)$$

donde s es el estimador de la desviación estándar poblacional, **es el intervalo de confianza** para □, con nivel de confiabilidad $(1-\alpha)\cdot 100$ %.

Es decir que cuando tenemos una muestra de tamaño mayor que 30, lo único que debemos hacer es utilizar la estimación de □ □para construir el intervalo de confianza.

Ejemplo 7. Calcular un intervalo de confianza del 95 % para la media de talla de niñas de 7 años, a partir de los datos de la tabla 1, asumiendo varianza desconocida. Suponga que el calculo de la desviación estándar muestral fue s = 4.383 cm.

Solución:

Debemos calcular un intervalo de la forma: $(\bar{x} - z_{1-\square/2}\, s/\sqrt{n}\ ;\ \bar{x} + z_{1-\square/2}\, s/\sqrt{n})$

Conocemos que, $\bar{x}$ = 122.05cm y s = 4.383 cm, y además como el nivel de confianza = 1 – α = 0.95, entonces $z_{1-\square/2} = z_{0.975} = 1.96$.

Calculemos ahora el valor de $z_{1-\square/2}\, s/\sqrt{n}$. Sustituyendo los valores conocidos se tiene que:

$z_{1-\square/2}\, s/\sqrt{n} = 1.96 \cdot 4.383 / \sqrt{200} = 1.215$.

Ahora restando y sumando 1.215 al valor de $\bar{x}$, obtenemos los límites del intervalo de confianza. Estos son: (120.84, 123.27).

La afirmación que podemos hacer en este caso es la misma que para □ conocida: para un nivel de confianza del 95 % la media poblacional de la talla de niñas cubanas de 7 años fluctúa, aproximadamente, entre 120.8cm y 123.2cm,

Similarmente, si extrajéramos r muestras de tamaño 200 es de esperar que aproximadamente 0.95 · r de ellas (es decir, el 95 % de r), produzcan intervalos de confianza que contengan el valor de □, mientras que aproximadamente el 0.05 · r no producirán intervalos con tal propiedad.

Segundo caso ($n \leq 30$).

Definición 7. Sea X una variable aleatoria normal con parámetros desconocidos □, $\square^2$. Para $n \leq 30$, el intervalo:

$$(\bar{x} - t_{n-1,1-\square/2}\, s/\sqrt{n}),\ \ \bar{x} + t_{n-1,1-\square/2}\, s/\sqrt{n}) \qquad (10)$$

donde:

$t_{n-1,1-\square/2}$: percentil de orden $1-\square/2$ de la t de Student con $n-1$ grados de libertad y

s: estimador de □,

es **un intervalo de confianza** para □ con nivel de confiabilidad (1 – α) · 100 %.

Es decir que para todo caso en que, n, no sobrepase la cifra de 30, tendremos que utilizar la tabla de la distribución t de Student para computar los límites del intervalo de confianza.

Ejemplo 8. Construya un intervalo del 95 % de confianza para la media poblacional de la talla de niñas de 7 años, asumiendo que se seleccionó una muestra de tamaño 20, con una media y varianza muestral de 121.93cm y 24.95 cm^2, respectivamente.

Solución:

Los datos que tenemos son:

n = 20, $\bar{x}$ = 121.93cm, $s^2 = 24.95\ cm^2$

Como n < 30 hay que usar la expresión $(\bar{x} - t_{n-1,1-\square/2}\, s/\sqrt{n}),\ \ \bar{x} + t_{n-1,1-\square/2}\, s/\sqrt{n})$.

De acuerdo a la tabla de la distribución t con □ = 0.05, pues el nivel de confianza deseado es del 95 %, se tiene que $t_{(20-1),(1-\square/2)} = t_{19,\,0.975} = 2.09$;

Tenemos que como conocemos s^2 , entonces $s = \sqrt{24.95} = 4.995$.

Sustituyendo los datos en la formula prevista, se tiene entonces que

$121.93 - 2.09 \cdot 4.995 / \sqrt{20} < \square < 121.93 + 2.09 \cdot 4.995 / \sqrt{20}$

$121.93 - 2.33 < \square < 121.93 + 2.33$

Luego (119.60, 124.26) es el intervalo de confianza buscado.

La interpretación es como siempre; tenemos un 95 % de confianza de que el valor de $\square$ se encuentre entre los límites aproximados de 119.6 cm y 124.3 cm.

11.3.3.- ESTIMACION POR INTERVALO DE CONFIANZA PARA π.

Como ya se sabe el estimador $\hat{p}$ del parámetro π de la distribución binomial es un estimador puntual insesgado de π y como tal, presenta las mismas deficiencias ya señaladas en el caso de $\bar{x}$. Sería entonces muy útil obtener estimaciones por intervalo de confianza para el parámetro π de la binomial.

Ejemplo 9. Asuma que la variable que nos interesa es "presentar estomatitis subprótesis" y que se quiere hallar un intervalo de confianza del 95 % para la proporción en la población, π, de enfermos de estomatitis subprótesis. Se realiza un pesquizaje en portadores de prótesis estomatológicas de Ciudad de La Habana, efectuándose para ello, la selección de una muestra aleatoria de 50 portadores, encontrándose que, 25 padecían de la citada enfermedad. Entonces el estimador puntual de π es $\hat{p} = 25/50 = 0.5$.

Si asumimos que se cumple la aproximación de la distribución binomial a la normal[4] entonces la proporción de enfermos en la muestra se distribuye de acuerdo a una normal con parámetro π; o de otra forma $\hat{p} \sim N(\pi, \pi(1-\pi)/n)$.

Similarmente al caso de μ, bajo las suposiciones anteriores, podemos plantear que:

Definición 8. Sea X una variable aleatoria binomial con parámetro desconocido π. Para n tal que se cumpla la aproximación de la distribución binomial a la distribución normal, el intervalo

$$(\hat{p} - z_{1-\square/2} \sqrt{\hat{p}(1-\hat{p})/n} \;\; \hat{p} + z_{1-\square/2} \sqrt{\hat{p}(1-\hat{p})/n}) \qquad (11)$$

es un **intervalo del (1- $\square$)·100 % de confianza** para π.

Ejemplo 10 (continuación). Vamos a calcular el intervalo del 95 % de confianza para la proporción de enfermos de estomatitis subprótesis en la población. Como $\hat{p} = 0.5$ y $1-\hat{p} = \hat{q} = 0.5$; computamos $n \cdot 0.5 \cdot 0.5 = 12.5 \geq 5$, lo que implica que es válido usar (11) para obtener una estimación de π por intervalo de confianza. Si hacemos las sustituciones de los datos del problema en:

$$(\hat{p} - z_{1-\square/2} \sqrt{\hat{p}(1-\hat{p})/n} \;;\; \hat{p} + z_{1-\square/2} \sqrt{\hat{p}(1-\hat{p})/n})$$

tendremos que, lo anterior se traduce en,

$$(0.5 - 1.96 \sqrt{0.5*0.5/50} < \pi < 0.5 + 1.96 \sqrt{0.5*0.5/50})$$

$$(0.5 - 1.96 \cdot 0.07 < \pi < 0.5 + 1.96 \cdot 0.07)$$

---

[4] Recordemos que para que dicha suposición sea válida $n \cdot \pi \cdot (1-\pi) \geq 5$

Como 1.96 · 0.07 = 0.14, los límites de confianza para π, o proporción de enfermos de estomatitis subprótesis en la población, son (0.36, 0.64).

Al igual que cuando construimos los diferentes intervalos de confianza para la media μ de una población normal, es de esperar que si se seleccionan un número suficientemente grande de muestras de tamaño n, aproximadamente el $(1-\square)\cdot 100$ % de las muestras produzcan intervalos de confianza que contengan el verdadero valor del parámetro π.

El intervalo (11) para la proporción poblacional es fácilmente aplicable al caso del porcentaje poblacional.

Como se sabe el porcentaje poblacional no es más que $\pi \cdot 100$, su estimación puntual será $\hat{p} \cdot 100$, por lo que si $\hat{P} = \hat{p} \cdot 100$, el intervalo (11) puede ser reescrito sustituyendo $\hat{p}$ por $\hat{P}$; y en lugar de hablar de un intervalo de confianza para la proporción poblacional se hablará entonces del intervalo de confianza para el porcentaje poblacional.

Ejemplo 11. Calcule el intervalo de confianza del 95 % para el porcentaje poblacional de enfermos de estomatitis subprótesis, con los datos del ejemplo 9.

Solución:

$\hat{P} = 0.5 \cdot 100 = 50\ \%, \qquad 100 - \hat{P} = \hat{Q} = 50\ \%$

Los límites inferior y superior del intervalo serán, respectivamente:

$LI = 50.0 - 1.96\sqrt{50.0 * 50.0/50} = 50.0 - 13.9 = 36.1$

$LS = 50.0 + 13.9 = 63.9$

El intervalo resultante es (36.1 %, 63.9 %); o sea, como era de esperar, es igual al de la proporción pero multiplicado por 100.

Por último, aunque es perfectamente posible que exista un problema empírico de interés donde $n \cdot \hat{p} \cdot \hat{q} < 5$. Este caso no es objeto de estudio en este libro, para la solución de tal caso, habría que aplicar el cálculo exacto de las probabilidades acumuladas por medio de la distribución binomial. Un tratamiento del tema puede verse en (1).

11.4.- TAMAÑO DE LA MUESTRA EN LA ESTIMACION POR INTERVALO DE CONFIANZA.

Al final del ejemplo 5 hicimos un análisis de la relación entre confiabilidad y precisión; donde también, de forma breve, se hizo referencia al tamaño de la muestra.

A partir de la fórmula de la precisión $d = z_{1-\square/2} \sigma/\sqrt{n}$, vamos a despejar n.

Multiplicando ambos lados de la igualdad previa por $\sqrt{n}$, se obtiene que

$\sqrt{n}\, d = z_{1-\square/2}\, \square$

Elevando ahora ambos lados de la igualdad al cuadrado y dividiendo luego por d^2, se deduce que : $n = (z_{1-\square/2}\, \square\square / d)^2$ (12)

Es decir que, si la desviación estándar es conocida y se fija el valor del coeficiente de confianza, podemos garantizar la precisión deseada usando (12) para calcular el tamaño de la muestra.

Ejemplo 12. Suponga que se quiere hacer una estimación por intervalo de confianza para la media de la población de tallas del anexo 1, pero que se desea alcanzar una precisión de 1 cm, con una confiabilidad del 95 %. ¿Cuál sería un tamaño de muestra adecuado?

Como ya sabemos que la desviación estándar de la talla en la población es 5.53cm y que a una confiabilidad del 95 % corresponde un coeficiente de confianza de 1.96, podemos sustituir los datos en (12) y así obtener el valor de n.

$$n = (1.96 \cdot 5.53/1)^2 \approx 118$$

Luego, se puede esperar que, si seleccionamos aleatoriamente 118 niñas, se alcance una precisión de 1cm en la estimación de la media poblacional.

11.5.- PRUEBA O CONTRASTE DE HIPOTESIS

Diariamente el ser humano se enfrenta a la toma de decisiones. Las hay muy simples; por ejemplo, si vamos a comer una fruta seleccionamos de acuerdo a nuestra experiencia anterior, **según el color, olor y consistencia**, la que consideramos mejor o más apetecible a nuestro gusto. Otras decisiones son algo más complejas como cuando un estudiante de preuniversitario en su último año tiene que escoger, de un conjunto de opciones que se le presentan, los futuros estudios a que va a dedicar, posiblemente, el resto de su vida productiva.

En situaciones de este tipo, quiérase o no, la toma de decisión esta siempre asociada a un determinado riesgo, es decir, se puede tomar una decisión desacertada. En estos casos puede resultar que no se tenga una clara noción del riesgo a que nos exponemos o sencillamente, aún y cuando se conozca sobre la existencia de tal riesgo, seamos incapaces de medirlo.

El profesional médico también constantemente se enfrenta a la toma de decisiones. Diariamente, cuando menos, tiene que discriminar entre alternativas de acción para decidir el tratamiento a seguir con un paciente determinado. En este caso el médico está perfectamente consciente del riesgo que enfrenta si toma la decisión equivocada, por lo que le es indispensable tener elementos de base científica que le ayuden a escoger, ante la incertidumbre, la mejor alternativa. A pesar de que esos elementos están, en general, bien establecidos para la práctica clínica, las Ciencias Médicas, como cualquier otra, no se salva del escrutinio científico constante y por ende la investigación científica juega un papel fundamental en su desarrollo.

Es de conocimiento casi universal que el adecuado desarrollo de la actividad investigativa se comienza a garantizar cuando se hace un buen planteamiento del problema a solucionar. En muchas ocasiones esa acción lleva al investigador de forma natural a construir una hipótesis científica y en el proceso de verificarla, tiene que aplicar técnicas que le proporcionen información acerca de la magnitud del riesgo a que se expone cuando toma la decisión final de aceptar o rechazar su hipótesis. Parece entonces lógico que, como primer paso, la hipótesis científica tenga su expresión en términos de una o varias hipótesis contrastables.

La Inferencia Estadística como cuerpo de conocimientos contiene procedimientos útiles a estos fines, que surgen a partir de la construcción de una hipótesis estadística.

Definición 1. Una **hipótesis estadística** es una afirmación sobre el comportamiento de una variable aleatoria, susceptible de ser verificada.

La filosofía detrás del proceso de verificación de la hipótesis estadística en cierto sentido se puede equiparar con la que rige el método de demostración por reducción al absurdo; es decir, afirmar algo y para demostrar su veracidad suponer primero que no se cumple, y entonces, por medio de un proceso lógico llegar a una contradicción con alguna propiedad ya establecida , y de este modo, arribar a la conclusión de que lo afirmado inicialmente es cierto.

Planteando el problema en términos estadísticos, supóngase que se quiere examinar la validez o no, de una hipótesis referida a un parámetro de la población, digamos que se quiere probar que la media poblacional μ no es igual a un valor determinado μ_0; lo expresado se acostumbra a representar por:

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

donde el símbolo $\mathbf{H_0}$ designa lo que se conoce como **hipótesis nula** y $\mathbf{H_1}$ a la **hipótesis alternativa**.

Similarmente se podría querer contrastar la pareja de hipótesis siguientes:

H_0: $\pi = \pi_0$

H_1: $\pi > \pi_0$,

donde π_0 es un valor particular de la proporción poblacional π.

Lo expuesto no agota el universo de posibilidades, sin embargo, lo que si puede afirmarse incuestionablemente , es que, para expresar un problema en términos de una prueba de hipótesis, se necesitan, una hipótesis nula y otra alternativa, que describan adecuadamente la situación contenida en el marco del problema.

Definición 2. La **hipótesis estadística** siempre estará integrada por la hipótesis nula y la hipótesis alternativa (es decir, ambas forman una unidad). En la **hipótesis nula** se afirma o considera la no ocurrencia del resultado esperado, mientras que la **hipótesis alternativa** contradice, en algún sentido, la hipótesis nula. De acuerdo a esto, rechazar H_0 equivale a probar el resultado esperado.

Ejemplo 13.- Supongamos que el 85 % de los casos de una infección X, curan mediante el empleo de un antibiótico A, conocido por la practica medica. Supóngase ahora que, en un instituto de investigaciones farmacológicas han desarrollado un nuevo antibiótico B, ideado para el tratamiento de esta infección y quieren probar que es más efectivo que el tradicional antibiótico A, para ello aplican el antibiótico B a un número determinado de enfermos seleccionados aleatoriamente. Para corroborar la efectividad del nuevo antibiótico se puede plantear la hipótesis estadística siguiente:

H_0: $\pi = \pi_0 = 0.85$

H_1: $\pi > \pi_0 = 0.85$,

o lo que es lo mismo:

H_0: $\pi = 0.85$ vs H_1: $\pi > 0.85$

Se debe notar que lo que desean probar los investigadores, el predominio del tratamiento B sobre el A, se ve reflejado en la hipótesis alternativa, donde debe observarse que el símbolo de proporción, π, ahí usado hace referencia a, **la proporción (poblacional) de casos que se curan por uso del antibiótico B**, aun cuando este símbolo no lo exprese claramente. Por ello, otra posible alternativa de expresar la hipótesis estadística, consiste en incorporar algo, al parámetro estadístico que permite expresar la hipótesis, pudiendo ser esto, un subíndice u otra cosa adecuada al caso, de modo que quede claro el significado del resultado final.

En relación con este ultimo ejemplo la hipótesis estadística pudiera expresarse por medio de:

H_0: $\pi_B = 0.85$ vs H_1: $\pi_B > 0.85$, si para todos es transparente que el valor 0.85 es un cifra asociada al antibiótico A, ó bien mediante,

H_0: $\pi_B = \pi_A = 0.85$ vs H_1: $\pi_B > \pi_A = 0.85$, si se desea que esto quede claramente expresado.

Puede usarse la notación $\pi(B)$, $\pi(A)$ en sustitución de π_B, π_A si alguien lo prefiere. La forma notacional particular que se adopte importa, pero no determina, lo que interesa es que si esta se usa, ella incorpore lo que en esencia permite distinguir o diferenciar las cosas en si.

Si en la muestra seleccionada se obtiene que un número determinado, digamos por ahora k de ellos, reaccionan favorablemente, es decir se curan, entonces rechazamos H_0 y declaramos el antibiótico B como mejor, si por el contrario el número de pacientes curados es menor que k, no rechazamos H_0 y entonces se declara que ambos antibióticos son, al menos, equivalentes.

Ejemplo 14.- Se ha determinado que, en general los niveles de colesterol en sangre se distribuyen normal con media 249 mg/100ml y desviación estándar 50 mg/100ml. Un grupo de investigadores de un Instituto de Enfermedades Cardiovasculares realizo un estudio en una muestra aleatoria de 150 vegetarianos, determinándose el nivel promedio muestral del colesterol en dicho conjunto de personas.

La hipótesis científica a ser probada estadísticamente es si individuos cuya ingesta está compuesta solo por vegetales, tienen los mismos niveles de colesterol que la población general. La hipótesis nula sería:

H_0: $\mu = \mu_0 = 240$ (o bien $\mu_{Veg.} = \mu_0 = 240$, donde $\mu_0 = 240$ representa lo general conocido)

El sentido de nula, viene dado, por que no se espera encontrar ninguna variación en el colesterol medio de los vegetarianos, salvo la debida al azar. La hipótesis alternativa sería:

H_1: $\mu \neq \mu_0 = 240$ (o bien $\mu_{Veg.} \neq \mu_0 = 240$)

En este caso, los investigadores, sólo están interesados en probar que existen diferencias entre el promedio de colesterol de vegetarianos y el de la población total. Ellos estarán de acuerdo, en rechazar H_0, tanto si el nivel medio de colesterol en vegetarianos es menor que un valor k_1, como si es mayor que un valor k_2.

Ahora, se impone, encontrar alguna forma de poder arribar a una conclusión con respecto a, rechazar o no H_0, incluyendo, al igual que lo hicimos en el caso de los intervalos de confianza, la información contenida en una única muestra.

Es tradicional plantear el problema en la forma siguiente:

Notemos que, en cada uno de los ejemplos, debemos rechazar o no H_0, basándonos en la información contenida en la muestra. En general, si el valor del estimador, **digamos $\bar{x}$, $\hat{p}$ o cualquier otro**, es tal que hay que rechazar la hipótesis nula, se dice que se ha obtenido un resultado significativo con un nivel de probabilidad dado.

Es posible que a causa de, la aleatoriedad de las observaciones muestrales, la estimación obtenida se desvíe tanto de lo esperado, que se tome, la decisión de rechazar H_0, siendo sin embargo H_0 cierta o verdadera. Es lógico o conveniente por tanto, que la probabilidad de que esto suceda, sea pequeña. Dentro de esta metodología, esta probabilidad, recibe el nombre de **nivel de significación**.

Definición 3. El **nivel de significación** de una prueba de hipótesis es el valor máximo de probabilidad que se está dispuesto a aceptar, para que ocurra el suceso de rechazar la hipótesis nula asumiendo que ésta es verdadera.

El nivel de significación, es un valor arbitrario, en el sentido de que es seleccionado a priori por el investigador de acuerdo a su experiencia y deseo. Siendo una probabilidad, puede asignársele cualquier valor entre 0 y 1, pero como es importante usar una cifra pequeña, los valores que con más frecuencia se utilizan son 0.05 y 0.01 o inclusive mas pequeños, aunque poco frecuentes en la practica usual . Se acostumbra a denotar este valor por la letra griega □.

Es conveniente notar que el uso del término, **significación**, es debido a que la diferencia entre, el valor hipotético (también llamado, teórico) y el hallado en la muestra (conocido como, practico), se considera lo, **suficientemente grande**, como para que no sea solamente atribuible al azar; es decir, que el concepto se refiere al estado de ser, **estadísticamente significativo**, y no es utilizado en el sentido funcional habitual de la palabra.

Con esto queremos decir, que la significación estadística, es realmente importante cuando está precedida de un planteamiento correcto del problema de investigación, y su resultado es compatible con el sustrato biomédico que la origina.

Una vez seleccionado el nivel de significación, se tiene una idea mucho más clara de cuando vamos a rechazar la hipótesis nula. Así, se rechaza H_0, cuando la disparidad entre el conjunto de datos observados y la hipótesis nula, asumiendo, que ésta es verdadera, se obtiene, con una probabilidad menor o igual al nivel de significación □.

Se espera por tanto, que a lo mas el □·□□□□% de las infinitas muestras posibles a seleccionar con el objetivo de verificar la hipótesis nula, nos conduzcan a rechazarla cuando ella sea verdadera; lo que nos ofrece cierta seguridad con respecto a lo acertado de la decisión que estamos tomando.

Definición 4. El conjunto de valores muestrales que conducen a rechazar H_0 se denomina o conoce como, **región crítica o de rechazo,** de la prueba de hipótesis.

Cuando, **H_0 es realmente verdadera y se toma la decisión de rechazarla**, se produce desde el punto de vista probabilístico, la ocurrencia de un evento o suceso, habitualmente denominado como, **error de tipo I**, **con probabilidad de ocurrencia** □, es decir, **el nivel de significación de la prueba de hipótesis**.

Definición 5. Se llama **error de tipo I** a rechazar H_0, cuando lo que se formula en H_0 es cierto.

Ejemplo 15.- Supongamos ahora que, X es una variable aleatoria, que tiene como ley de probabilidad, una distribución normal con media μ y varianza σ^2 (en símbolos $X \sim N(\mu, \sigma^2)$)

Tenemos interés en probar la hipótesis: H_0: $\mu = \mu_0$ Vs. H_1: $\mu \neq \mu_0$

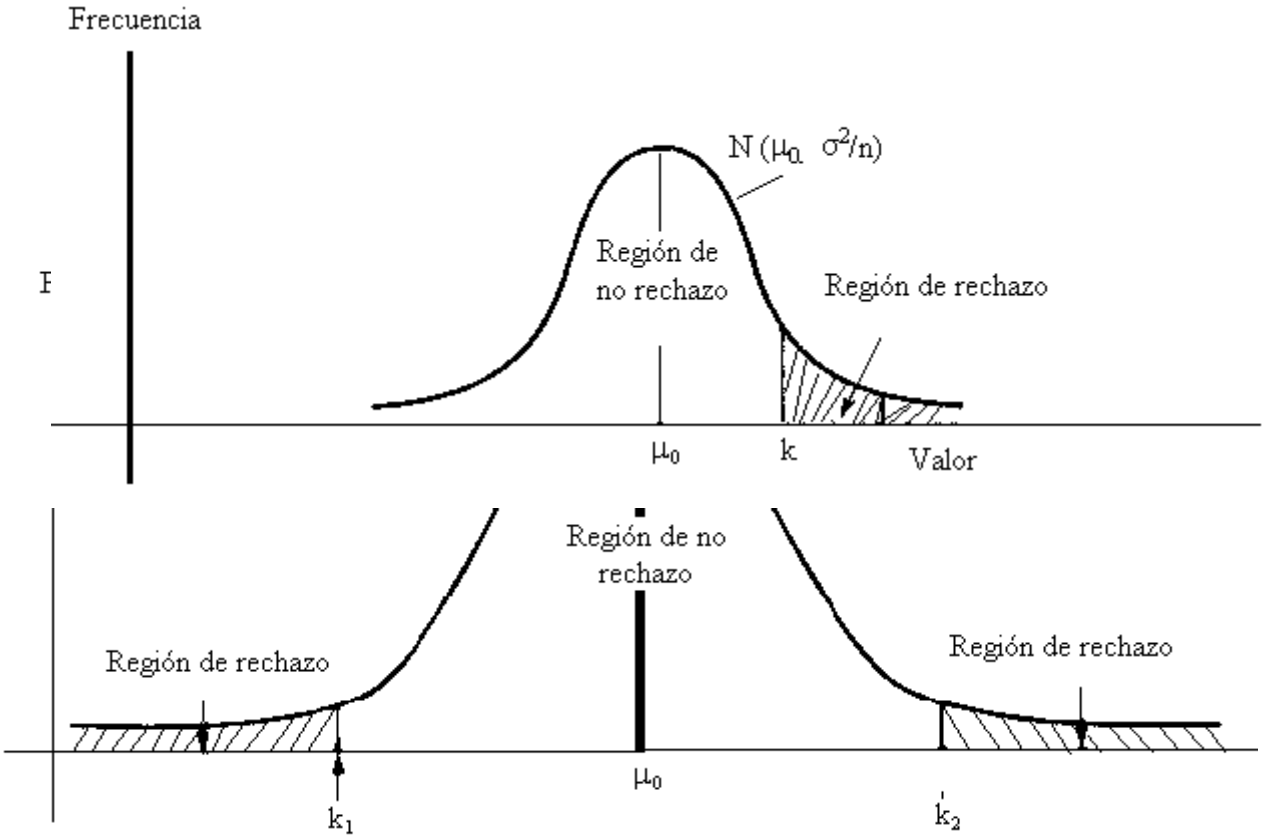
y se ha escogido □ = 0.05, que expresado en porcentajes, quiere decir, un nivel de significación del 5 %, entonces, suponiendo que se cumple la hipótesis nula, se rechazará ésta, si la media muestral $\bar{x}$ está entre los valores extremos con probabilidad 0.05, en cualquiera de las dos direcciones de la media hipotética μ_0. Si $\bar{x}$ no se ubica en esa región, entonces no se rechaza H_0.

En el gráfico 1 se representa la región crítica suponiendo que la hipótesis nula es cierta.

Ya que, la distribución normal es simétrica, parece bastante lógico distribuir a partes iguales la probabilidad de rechazar H_0, cuando esta es verdadera, entre las dos colas de la curva, y entonces rechazaremos H_0 sólo si $\bar{x} < k_1$ o $\bar{x} > k_2$. La región crítica o de rechazo, entonces estará constituida por el conjunto de las todas

muestras de un tamaño, n, dado tales que el valor observado de, $\bar{x}$, cumpla con una de las dos siguientes condiciones: $\bar{x} < k_1$ o $\bar{x} > k_2$.

Si, por otra parte, $\bar{x}$ no se ubica en alguna de estas zonas, lo que sucede con probabilidad $1 - \alpha$, entonces en este caso se decide no rechazar H_0, y en consecuencia a esa región central bajo la curva normal se le llama, **región de no rechazo**. En muchos textos de estadística se le denomina región de aceptación,



nosotros preferimos utilizar una terminología menos categórica puesto que la decisión siempre se toma ante la incertidumbre.

Figura 1. Región crítica para la prueba bilateral de una media poblacional, suponiendo que H_0 es cierta.

Las pruebas de hipótesis de este tipo, cuya región critica se representa del modo grafico ya indicado, es decir, por medio de los dos extremos o colas de la curva normal, reciben el nombre de **pruebas bilaterales** o **de dos colas**.

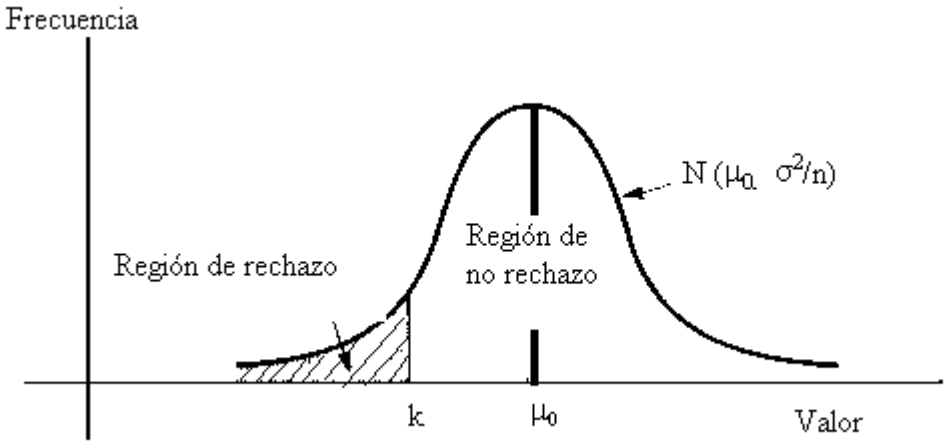
Si en lugar de la hipótesis que se consideró en el ejemplo 15, hubiésemos considerado una hipótesis del tipo:

$H_0: \mu = \mu_0$ Vs. $H_1: \mu > \mu_0$.

Entonces a esta corresponderá, la **región de rechazo {X : $\bar{x}$ > k}** y la de **no rechazo {X: $\bar{x} \leq$ k}**, como puede verse en la figura siguiente.

Figura 2. Región crítica para la prueba unilateral derecha de una media poblacional, suponiendo que H_0 es cierta.

Si en vez de la hipótesis anterior, hubiésemos tenido que considerar, por ser necesario para la solución de un



problema practico una opción de hipótesis del tipo: H_0:

$\mu = \mu_0$ Vs. H_1: $\mu < \mu_0$

A esta corresponde, la **región de rechazo {X: $\bar{x}$ < k}** y la de **no rechazo {X: $\bar{x} \geq$ k}**, tal como se muestra en la figura que a continuación se expone:

Figura 3. Región crítica para la prueba unilateral izquierda de una media poblacional, suponiendo que H_0 es cierta.

Las dos últimos tipos de pruebas de hipótesis que se han explicado, en los que por la forma de expresarse su región critica, es decir, por estar formada por valores situados en uno **sólo de los extremos** o **colas de la curva normal**, se denominan como **pruebas de hipótesis unilaterales o de una cola**.

Cuando se introdujo el concepto, nivel de significación de una prueba de hipótesis, y se explico lo que esto quiere decir, se dijo, que este valor esta asociado a la ocurrencia del llamado error tipo I o suceso que consiste en, rechazar H_0 cuando H_0 es verdadera, hecho que es factible que ocurra cuando el valor observado de $\bar{x}$ en una muestra, este alejado de μ_0 en cierto sentido prefijado por la región critica de que se trate. Sin embargo, es posible que la estimación de μ o valor observado de $\bar{x}$**,** esté tan próximo al valor hipotético μ_0, que se decida no rechazar H_0 cuando ésta es en efecto falsa. Si esto sucede estamos cometiendo un nuevo error, que se conoce como error de tipo II. La probabilidad de cometer un error de tipo II es designada habitualmente por la letra griega □.

Definición 6. Se conoce como **error de tipo II**, al suceso o evento que consiste en, **no rechazar H_0 cuando lo que se expresa en ella es falso**.

Por último, si la estimación se desvía tanto del valor hipotético que se decide rechazar H_0 cuando es en efecto falsa, se habrá tomado una decisión correcta.

Resumiendo las situaciones anteriores respecto a las decisiones sobre H_0 a partir de los valores observados en la muestra aleatoria tenemos que:

	Decisión sobre H_0	
Si H_0 es:	No rechazar	Rechazar
Verdadera	Acción correcta	**Error de tipo I**
Falsa	**Error de tipo II**	Acción correcta

Tratemos ahora de resumir los resultados hasta aquí obtenidos. Como se ha visto a través de la exposición de los temas llevados a cabo, todo problema de prueba de hipótesis consiste en lo siguiente:

1. Identificar una variable aleatoria X que tiene una **distribución conocida**, es decir, que pertenece a una clase determinada, por ejemplo a las del tipo normal, y con relación a la cual se quiere tomar una decisión respecto al valor de un parámetro desconocido, pero asociado a ella, digamos μ, σ^2 , ...,etc
2. Se plantea una hipótesis nula, donde se asume un valor para el parámetro; y una hipótesis alternativa donde se contradice lo expresado en la hipótesis nula.
3. Se escoge el nivel de significación □, que es la probabilidad de rechazar la hipótesis nula siendo esta cierta.
4. Se selecciona una muestra de tamaño n para estimar el parámetro desconocido y poder posteriormente decidir si se rechaza o no H_0.
5. Se define la región crítica para la prueba de hipótesis de interés.
6. Se toma la decisión de rechazar H_0, con un nivel de significación □, si el valor estimado del parámetro está en la región crítica y de no rechazar H_0 si este valor no está en la región crítica.

De ahora en adelante nos dedicaremos a encontrar las distintas regiones críticas para cada una de las pruebas de hipótesis que se planteen.

11.6.- PRUEBAS DE HIPOTESIS RELACIONADAS CON LA MEDIA.

Caso I. Decisión sobre μ con $\square^2$ conocida.

a. **Prueba de hipótesis unilateral**.

Ejemplo 16. Se conoce que la estatura en niños varones de 5 años tiene la ley de distribución normal con media 109.0cm y varianza 25.9cm^2. Un investigador presume que esta dimensión antropométrica está afectada en niños diabéticos de la misma edad. Para tratar de corroborar su suposición selecciona una muestra aleatoria de 81 niños diabéticos y obtiene que su talla media es de 107.5 cm. ¿Puede dicho investigador, basado en este resultado, afirmar que lo supuesto por el es acertado?

En este caso las hipótesis nula y alternativa serían: H_0: $\mu = 109$ y H_1: $\mu < 109$

De planteamiento anterior tenemos que $\bar{x} = 107.5$. Fijemos ahora $\square$ en 0.05, entonces de lo que se trata es de encontrar un k tal que si $\bar{x} < k$ podamos rechazar H_0 con probabilidad $\square$, cuando H_0 es verdadera. Es decir que k debe cumplir : $Pr(\bar{x} < k) = \square$, suponiendo que H_0 es verdadera. (1)

Bajo H_0, $\bar{x} \sim N(109, 25.9/\sqrt{81})$, luego podemos poner: $\frac{(k-\mu_0)}{\sigma/\sqrt{n}} = z_{\square}$ al hacer uso una vez más, de la estandarización de la variable aleatoria normal.

Multiplicando ambos lados de la igualdad por $\square / \sqrt{n}$ tendremos que:

$$k - \mu_0 = z_{\square}\, \square / \sqrt{n}$$

En consecuencia: $k = \mu_0 + z_{\square}\, \square / \sqrt{n}$ (2)

Luego si $\bar{x} < \mu_0 + z_{\square}\, \square / \sqrt{n}$, se rechazara H_0 con un nivel de significación $\square$, pero si $\bar{x} \geq \mu_0 + z_{\square}\, \square / \sqrt{n}$ no podemos rechazar la hipótesis nula.

Debemos calcular ahora el valor de k haciendo uso de (2).

Para ello se debe emplear la tabla de la curva normal estándar que se incluye en este libro, que es la correspondiente al área acumulada debajo de la curva normal estándar entre 0 y un valor positivo z (esto supone que en los cálculos se emplee la mitad del área bajo la curva, es decir 0.5, así como la simetría), por lo que $z_{\square\square} = z_{0.05} = -z_{0.5-0.05} = -z_{0.45} = -1.64$.

Sustituyendo en (2) tenemos que :

$k = 109.0 - 1.64 \cdot \sqrt{25.9/81} = 109.0 - 8.35 / 9 = 108.1$.

Pero $\bar{x} = 107.5 < 108.1$, lo que nos conduce a rechazar H_0 con un nivel de significación de 0.05.

El investigador concluye que la diferencia entre la talla promedio de los niños diabéticos estudiados y la correspondiente a la población, de su misma edad y sexo, es **estadísticamente significativa al nivel 0.05.**

Lo que aparece en negrita quiere decir que el investigador espera que la diferencia sea producto del azar solamente con una probabilidad del 5%; es decir, si se repiten investigaciones reproduciendo la misma metodología de trabajo empleada por él, se espera que aproximadamente a lo mas en el 5% de los casos se rechace H_0 suponiendo que ésta sea cierta.

La frase **suponiendo H_0 verdadera**, o más brevemente **bajo H_0**, tiene una importancia cardinal pues la región crítica se basa precisamente en ese supuesto y se trata de que sea poco probable cometer el error de rechazar H_0 siendo esta cierta o verdadera (error de tipo I). Esta es la razón por la que se prioriza mantener una cota fija para □ de a lo más 0.05.

El estudiante debe notar la sutileza del problema. No podemos afirmar que rechazamos la hipótesis nula porque no se cumple, diremos que se rechaza porque el nivel de significación □, que es la probabilidad de cometer un error de tipo I, es pequeño.

El valor 0.05 que se le asigna al nivel de significación puede deber su uso universal al hecho de que en la distribución normal, alrededor de la media más/menos dos desviaciones estándar está concentrada el 95% de la información; se espera entonces que los datos que se ubican en las llamadas colas de la distribución ocurran menos frecuentemente que el resto. Para el otro nivel de significación que también se utiliza mucho, 0.01, la frecuencia con que aparecen datos en los extremos es aún menor. Por esta razón en la literatura se habla de significativo si □ = 0.05 y de muy significativo si □ = 0.01.

La región crítica de la prueba considerada puede definirse también a partir del estadígrafo:

$$z = (\bar{x} - \mu_0) / \square / \sqrt{n}$$

que es el valor observado o practico del percentil de la distribución normal estándar por medio de los datos de la muestra, bajo la suposición de que se cumple H_0. Se quiere llamar la atención hacia el hecho de que en la mayoría de los libros de texto donde se tratan pruebas de hipótesis se utilizan estadígrafos para construir la región crítica pues casi siempre, bajo H_0 , tienen una ley de distribución conocida.

Entonces para proceder por este modo, se compara z con $z_{\square}$, que es el percentil teórico fijado a priori, es decir, con el conocimiento del valor de α para la misma distribución. Se rechaza H_0 si $z < z_{\square}$, y no se rechaza si $z \geq z_{\square}$.

Calculemos ahora el valor de z para el ejemplo anterior. Se tiene que,

$$z = (107.5 - 109.0) / 5.09/9 = -2.65.$$

Ciertamente este valor es menor que –1.64 por lo que se rechaza H_0 como era de esperar, pues las regiones críticas definidas de una u otra forma son equivalentes.

Por último se puede decidir rechazar o no H_0 comparando el nivel de significación real obtenido a partir de los datos, con el nivel de significación prefijado.

Por ejemplo, si □ = 0.05, bastaría con calcular la probabilidad p (el área bajo la curva) asociada al valor del estadígrafo z, por medio la tabla de la normal estándar y compararla con 0.05.

La región crítica o de rechazo sería, {x: p< □}.

La región de no rechazo, {x: p ≥ □}

Esta manera de definir la región crítica comparando p con □ para tomar una decisión con respecto a H_0 es muy usada, pues los paquetes de programas estadísticos comúnmente ofrecen el resultado del nivel de significación real u observado, denominándolo indistintamente como: significación (de una cola), significación (de dos colas) o p; entonces el investigador puede decidir de forma muy rápida si sus datos corroboran o no su suposición inicial.

En el caso del ejemplo que venimos desarrollando habría que hallar en la tabla de la distribución normal la probabilidad p, asociada al numero, –2.65, que es el valor de z observado. Puede comprobarse que ésta es aproximadamente 0.004, menor que 0.05, por lo que se rechaza H_0.

Definición 7. Sea X una variable aleatoria normal con varianza σ^2 conocida, suponiendo H_0 verdadera, la región crítica de la prueba de hipótesis

H_0: $\mu = \mu_0$ Vs. H_1: $\mu < \mu_0$

para un nivel de significación □, se puede expresar por medio:

1. $\{x: \bar{x} < \mu_0 + z_{□}\ □ / \sqrt{n}\ )$
2. $\{x: z < z_{□}\}$.
3. $\{x: p < □\}$.

Corresponde ahora encontrar la región crítica para la prueba de hipótesis unilateral derecha.

Ejemplo 17. Supongamos que se conoce que el promedio del nivel de colesterol en niños menores de 10 años de la provincia de Matanzas es de 175 μmol/mL y que la desviación estándar es 50 μmol/mL. Se extrae una muestra aleatoria de 10 niños obesos y se obtiene $\bar{x}$ = 200 μmol/mL. Se desea contrastar la hipótesis de que el colesterol medio en niños obesos es mayor que el de los niños de la provincia de Matanzas.

La hipótesis a contrastar sería:

H_0: $\mu = 175$ Vs H_1: $\mu > 175$

Tomemos un □ fijo. En este caso rechazar H_0 equivale a encontrar un k tal que: $Pr(\bar{x} > k) = □$, bajo H_0 (3)

Se cumple que $1 - Pr(\bar{x} \leq k) = Pr(\bar{x} > k)$, por lo que se puede poner

$1 - Pr(\bar{x} \leq k) = □$.

Sumando la probabilidad y restando □□en ambos lados de esta última igualdad tendremos que:

$1 - □□ = Pr(\bar{x} \leq k)$, bajo H_0 (4)

Es decir que (3) y (4) son condiciones equivalentes. Al mismo tiempo (4) es semejante a (1), lo que implica que k puede expresarse como en (2) pero sustituyendo $z_{□□}$por $z_{1-□}$; por consiguiente: $k = \mu_0 + z_{1-□}\ □ / \sqrt{n}$.

En este caso se rechazará H_0, cuando es verdadera, si $\bar{x}$ se mantiene por encima de k con probabilidad □; o lo que es igual, rechazo H_0 siendo cierta, si la probabilidad de que $\bar{x}$ sea mayor que k es baja.

Definición 8. Sea X una variable aleatoria normal con varianza conocida □□, suponiendo H_0 verdadera, la región crítica de la prueba de hipótesis

H_0: $\mu = \mu_0$ Vs. H_1: $\mu > \mu_0$

para un nivel de significación □, toma cualquiera de las formas siguientes:

1. $\{x: \bar{x} > \mu_0 + z_{1-□}\ □ / \sqrt{n}\ \}$.
2. $\{x: z > z_{1-□}\}$.

3. $\{x: p < \square\}$.

Se debe notar que la región 3 es la misma que cuando la prueba de hipótesis era unilateral izquierda; lo que es totalmente lógico puesto que si $\square\square$es el nivel de significación de la prueba, valores muestrales que produzcan cifras de la media muestral o del estadígrafo z que nos conduzcan a rechazar la hipótesis nula tienen, a lo más, una probabilidad de ocurrencia $\square$. En este caso a la hora de calcular p, como la región crítica se encuentra a la derecha del valor hipotético, tendremos que hallar el área bajo la curva que tiene como límite inferior el estadígrafo z .

Vamos a solucionar el ejemplo 17 evaluando los datos para las tres regiones críticas posibles.

Región crítica 1.Como ya sabemos en la tabla que se proporciona en este texto, en lugar de $z_{1-\square\square}$hay que buscar $z_{0.5-\square} = z_{0.45} = 1.64$. Entonces :

$k = 175 + 1.64 \cdot 50 / \sqrt{10} = 175 + 1.64 \cdot 50 / 3.16 = 200.93.$

Ya que $\bar{x} = 200 < 200.93$, no se rechaza H_0 y se concluye que la diferencia hallada entre el nivel medio de colesterol del grupo de niños obesos y el de la población no es estadísticamente significativo al nivel de 0.05.

Región crítica 2. $z = 200-175/ 50 / \sqrt{10} = 1.58$.

Como $z_{0.45} = 1.64$, no se rechaza H_0 con un nivel de significación del 5%.

Región crítica 3. En la tabla de la distribución normal estándar se puede ver que la p que corresponde a valores iguales o mayores que 1.58 es 0.057, mayor que 0.05 que es el nivel de significación fijado, por lo que de nuevo no se rechaza H_0.

b. Prueba de hipótesis bilateral.

En la mayoría de las ocasiones el interés del investigador es el de detectar diferencia significativa en cualquiera de las dos direcciones posibles, tanto relacionada con valores extremadamente bajos como con valores extremadamente altos. Cuando este es el caso, la prueba de hipótesis consiste en contrastar : $H_0: \mu = \mu_0$ Vs $H_1: \mu \neq \mu_0$

Para construir la región crítica tendremos que hallar k_1 y k_2 tales que

$$Pr(\bar{x} < k_1 \text{ ó } \bar{x} > k_2) = \square \text{ bajo } H_0 \qquad (5)$$

Por definición el conjunto que aparece en (5) entre paréntesis es la unión de las dos regiones sombreadas del gráfico 1, que no se interceptan en ningún punto; luego (5) equivale a:

$$Pr(\bar{x} < k_1) + Pr(\bar{x} > k_2) = \square\square \text{bajo } H_0 \qquad (6)$$

Debido a la propiedad de simetría de la distribución normal, es plausible igualar cada una de las probabilidades en (6) a $\square/2$ y por similitud con las pruebas unilaterales podremos hacer la siguiente definición.

Definición 9. Sea X una variable aleatoria normal con varianza conocida $\square^2$. Suponiendo H_0 verdadera, la región crítica de la prueba de hipótesis

$H_0: \mu = \mu_0$ Vs. $H_1: \mu \neq \mu_0$

para un nivel de significación $\square$, toma una cualquiera de las formas siguientes:

1. $\{x: \bar{x} < \mu_0 - z_{1-\square/2}\ \square / \sqrt{n}$, para $\bar{x}$ a la izquierda de μ_0 ó

$\bar{x} > \mu_0 + z_{1-\square/2}\ \square / \sqrt{n}$, para $\bar{x}$ a la derecha de $\mu_0\}$

2. $\{x: z < -z_{1-\square/2}$, para $z < 0$ ó $z > z_{1-\square/2}$, para $z > 0\}$.

Cuando la prueba es bilateral el cálculo del nivel de significación observado es algo mas trabajoso, por lo que no introduciremos su cómputo en este texto.

Ejemplo 18. El nivel promedio de protrombina en la población total es de 20 mg/100ml de plasma y su desviación estándar es de 4 mg/100ml. Una muestra de 40 pacientes portadores de un déficit de vitamina K presenta un nivel promedio de protrombina de 18.5 mg/100m.l. A partir de esos resultados ¿será posible concluir que el nivel promedio en esos pacientes es similar al de la población normal? Use $\square = 0.05$.

Para responder la interrogante vamos a contrastar la hipótesis nula H_0: $\mu = 20$ Vs. la alternativa H_1: $\mu \neq 20$.

Como $\square = 0.05$, $\square/2$ es 0.025. Utilizando la tabla de la normal estándar puede comprobarse que $z_{0.475} = 1.96$.

Ya que $\bar{x} = 18.5 < \mu_0 = 20.0$ tenemos que evaluar $\bar{x} < \mu_0 - z_{1-\square/2}\ \square / \sqrt{n}$.

Ahora bien $z_{0.475}\ \square / \sqrt{40} = 1.24$ y $20.0 - 1.24 = 18.76$.

Como $18.5 < 18.76$, se rechaza H_0 , concluyendo que el nivel promedio de protrombina en los pacientes portadores de un déficit en vitamina K no es similar al de la población normal, para un nivel de significación de 0.05.

Siendo los casos de pruebas bilaterales los que se aplican con más frecuencia y considerando que la construcción de la región crítica es en esencia igual para éstas que para las unilaterales, de aquí en adelante sólo desarrollaremos, con cierto detalle, las primeras.

11.7.- CASO II. DECISIÓN SOBRE μ CON σ^2 DESCONOCIDA.

Hasta ahora hemos asumido que la varianza poblacional es conocida, hecho que en la práctica ocurre con poca frecuencia. Lo más común es que haya que tomar una decisión acerca de la media de la población desconociendo el valor de su varianza.

Consideremos de nuevo el ejemplo 18 pero suponiendo ahora que se estimó la desviación estándar con los datos de la muestra y se obtuvo s = 4.6.

Para verificar la hipótesis nula se calculó $k_1 = \mu_0 - z_{1-\alpha/2}\ \sigma / \sqrt{n}$ y se comparó con $\bar{x}$; pero ahora desconocemos σ por lo que en la expresión de k_1 tenemos que sustituirla por s, su estimación a partir de la información de la muestra. Entonces $z_{1-\alpha/2}$ ya no puede ser interpretado como el percentil $1-\alpha/2$ de la distribución normal (recuerde que en el epígrafe correspondiente se llego a esta conclusión ya que $(\bar{x} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$).

Puede demostrarse que en este caso el estadígrafo $t = (\bar{x} - \mu_0)\sqrt{n}/s$ tiene la distribución t de Student con n–1 grados de libertad, por lo que se utiliza esta distribución en la solución de la región crítica .

Definición 10. Sea X una variable aleatoria normal con varianza desconocida σ^2 que se estima, como es usual, a través de s^2. Suponiendo H_0 verdadera, la región crítica de la prueba de hipótesis:

$H_0: \mu = \mu_0$ Vs. $H_1: \mu \neq \mu_0$

para un nivel de significación α, toma cualquiera de las formas siguientes:

1. $\{x: \bar{x} < \mu_0 - t_{n-1,1-\alpha/2}\, s/\sqrt{n}$ para $\bar{x}$ a la izquierda de μ_0 ó

   $\bar{x} > \mu_0 + t_{n-1,1-\alpha/2}\, s/\sqrt{n}$ para $\bar{x}$ a la derecha de $\mu_0\}$

2. $\{x: t < -t_{n-1,1-\alpha/2}$ para $t < 0$ ó $t > t_{n-1,1-\alpha/2}$ para $t > 0\}$

donde:

t: percentil observado de la t– Student con n–1 grados de libertad y

$t_{n-1,1-\alpha/2}$:percentil $1-\alpha/2$ de la distribución t de Student con n–1 grados de libertad.

En caso de σ desconocida pero n suficientemente grande, digamos n > 30, los valores de la distribución t de Student son muy similares a los de la normal, por lo que se puede trabajar indistintamente con una u otra distribución.

Para la prueba de hipótesis unilateral

$H_0: \mu = \mu_0$, σ desconocida

$H_1: \mu < \mu_0$

la región crítica para el nivel de significación α será:

1. $\{x: \bar{x} < \mu_0 - t_{n-1,1-\alpha}\, s/\sqrt{n}\}$ ó
2. $\{x: t < -t_{n-1,1-\alpha}\}$.

En el caso de la prueba de hipótesis:

H_0: $\mu = \mu_0$ Vs. H_1: $\mu > \mu_0$, con □ desconocida

la región crítica para el nivel de significación □ puede construirse segun:

1. $\{x: \bar{x} > \mu_0 + t_{n-1,1-\square}\, s / \sqrt{n}\}$ ó
2. $\{x: t > t_{n-1,1-\square}\}$.

11.8- CASO III. DECISIÓN ACERCA DE LA DIFERENCIA DE DOS MEDIAS.

Hasta aquí hemos resuelto problemas relacionados con la toma de decisión, rechazar o no rechazar la hipótesis nula, basados en la información contenida en una única muestra cuyos resultados se querían comparar con un valor conocido del parámetro; sin embargo es muy común en la investigación biomédica, enfrentarse a situaciones donde es necesario evaluar el efecto de un tratamiento sobre un grupo de sujetos o comparar los resultados de dos ó más grupos con respecto, por ejemplo, a la introducción de nuevos medicamentos para tratar una enfermedad o un síntoma.

Las situaciones que se presentan a continuación se refieren pruebas de hipótesis para contrastar diferencia entre medias procedentes de dos poblaciones.

A. **Problema de las muestras apareadas**.

Ejemplo 19. Supongamos que se quiere comprobar la consistencia de dos técnicos en antropometría al medir el peso corporal de un conjunto de individuos. Se diseña un experimento donde se escoge un grupo de niños de primaria y cada uno de los técnicos mide, alternativamente, a todos los niños. Estos resultados aparecen en la tabla 1. donde para cada niño se tienen dos valores de peso, es decir contamos con dos conjuntos de valores de peso pero cada par está vinculado a un mismo niño. A este tipo de diseño se le denomina de muestra apareada.

En lugar de tener un conjunto de niños medidos por dos técnicos diferentes podría tratarse de

- un conjunto de pacientes con hipercolesterolemía a los que se les determinó el nivel de colesterol en sangre antes y después de un tratamiento con PPG (ateromixol).
- dos grupos de individuos diferentes pero seleccionados de forma tal que se esté seguro de que, de existir, las diferencias entre ellos se deban al factor en estudio **por ejemplo estudiar el efecto de dos analgésicos en gemelos.**

en todos estos casos estamos ante un diseño de **muestra apareada** y el interés radica en saber si los resultados obtenidos en las dos muestras son o no iguales.

De acuerdo al procedimiento que se ha venido aplicando lo primero que tenemos que hacer es plantear las posibles hipótesis, para lo que necesitamos hacer algunas suposiciones.

Tabla 1. Valores de peso (en kg) correspondiente a niños medidos por dos técnicos diferentes.

Niño no.	***PESO***		$(d_i = x_i - y_i)$	$(d_i^2 = [x_i - y_i]^2)$
	Técnico 1 (x_i)	Técnico 2 (y_i)		
1	21.0	20.8	0.2	0.04
2	24.2	23.8	0.4	0.16

3	25.8	25.6	0.2	0.04
4	30.4	30.2	0.2	0.04
5	27.0	27.4	– 0.4	0.16
6	27.2	27.4	– 0.2	0.04
7	28.4	29.4	– 1.0	1.00
8	24.4	24.6	– 0.2	0.04
9	31.4	31.2	0.2	0.04
10	21.2	21.2	0	0
11	24.6	24.2	0.4	0.16

Vamos a asumir que el peso obtenido por el técnico 1 es una variable aleatoria, X, con distribución $N(\mu_x, \square^2)$ y que el peso obtenido por el técnico 2 también es una variable aleatoria, Y, con distribución $N(\mu_y, \square^2)$; es decir, se supone que ambas provienen de poblaciones con medias desconocidas pero varianza común. Es razonable pensar que si los técnicos son consistentes entre sí, el promedio de las diferencias individuales entre una y otra medición deba ser cero. Si denotamos por d esas diferencias la hipótesis a plantear sería:

$H_0: \mu_d = 0$ Vs. $H_1: \mu_d \neq 0$ (7)

Donde, rechazar H_0 equivaldría entonces a decir que los técnicos no son consistentes, mientras que no rechazarla nos conduciría a pensar que ambos miden el peso, en promedio, de forma similar.

Es un resultado conocido que $d_i = x_i - y_i$ se distribuye normal con media y varianza que podemos denotar por μ_d y , respectivamente; entonces $\bar{d}$, o media muestral de las diferencias entre las mediciones que los técnicos hacen, tendrá a su vez una σ_d^2 distribución normal con parámetros μ_d y σ_d^2/n con lo que resulta que (7) puede interpretarse como una prueba de hipótesis bilateral para una sola muestra, en particular una muestra constituida por las diferencias entre pares de medidas referidas a cada individuo de un conjunto determinado.

Ya hemos visto que es poco frecuente conocer de antemano el valor de $\square^2$. En el caso de la hipótesis (7) tendríamos que tener información acerca de σ_d^2, la variabilidad en la población de diferencias individuales. Asumir que se conoce dicha variabilidad es poco práctico, por lo que al derivar la región crítica de la prueba de hipótesis vamos a utilizar desde un inicio el estimador de σ_d, que tiene la expresión:

$$s_d = \sqrt{\frac{\sum_i d_i^2 - \left(\sum_i d_i\right)^2 / n}{n - 1}} \qquad (8)$$

donde n es el número de pares.

Para la prueba (7) tenemos que el estadígrafo $t = (\bar{d} - 0)\sqrt{n} / s_d$ sigue una distribución t–Student con n–1 grados de libertad (hemos escrito el cero para que se vea la similitud entre esta expresión y la de la prueba bilateral con $\square^2$ desconocida), en consecuencia podemos hacer el siguiente planteamiento.

Definición 11. Sean d_i las diferencias observadas en una muestra apareada de tamaño n, donde $d_i \sim N(\mu_d, \sigma_d^2)$. Suponiendo H_0 verdadera, la región crítica de la prueba de hipótesis: H_0: $\mu_d = 0$ Vs. H_1: $\mu_d \neq 0$,

para un nivel de significación □, está constituida por los valores de d_i tales que:

$$(t < t_{n-1,1-\square/2} \text{ para } \bar{d} < 0 \text{ ó } t > -t_{n-1,1-\square/2} \text{ para } \bar{d} > 0)$$

donde:

t: percentil observado de la distribución t de Student con n–1 grados de libertad,

$t_{n-1,1-\square/2}$: percentil 1–□/2 de la distribución t de Student con n–1 grados de libertad y

s_d: se calcula a través de la fórmula (8).

Ejemplo 19 (continuación)

Considerar para esto un nivel □ de 0.05. y calculemos t. Aunque la fórmula (8) parece muy complicada es sencilla de solucionar siempre que se arregle la información como aparece en la tabla no. 1. A partir de ésta tenemos que:

$$\sum_{i=1}^{11} d_i^2 = 1.72 \qquad \left(\sum_{i=1}^{11} d_i\right)^2 = 0.04 \qquad \bar{d} = -0.02$$

luego $s_d = \sqrt{(1.72 - 0.04/11)/10} = 0.41$ y en consecuencia,

$t = -0.02\ \sqrt{11}/0.41 = -0.16$

En la tabla de la distribución t student, el percentil que corresponde a 10 grados de libertad y a 1–□/2 = 0.975 es $t_{10,0.975} = 2.228$. Como t < 0 debemos compararlo con $-t_{10,0.975}$ de lo que se deduce, que no se rechace H_0, y por tanto podemos pueda aceptarse que las medidas tomadas por ambos técnicos son similares.

Resolviendo este ejemplo a través de un paquete de programas estadísticos, digamos el SPSS, se puede comprobar que el nivel de significación observado p, en el caso de dos colas, es de 0.887; un valor mucho mayor que el 0.05 fijado de antemano lo que conduce a rechazar la hipótesis nula, como era de esperar.

B. **Problema de muestras independientes con varianzas iguales y conocidas**.

Muchas veces es de interés comparar la media de variables provenientes de dos poblaciones que tienen una característica de base diferente.

Por ejemplo el peso al nacer es una variable muy estudiada por la importancia que tiene, entre otros aspectos, en el crecimiento y desarrollo futuro del recién nacido, el objetivo de un estudio podría ser analizar las diferencias entre los promedios de talla de dos grupos de niños de 7 años, donde uno de los grupos está formado por niños clasificados como bajo peso al nacer y el otro por los que no. En este caso queremos hacer comparaciones entre las medias de dos muestras independientes.

Ejemplo 20. Un grupo de 3699 niños de ambos sexos de 7 años de edad fue clasificado de acuerdo a si su peso al nacer estuvo por debajo de 2500 gramos (bajo peso al nacer), o no. Los datos correspondientes a la estatura promedio de cada grupo se dan en la tabla 2 que se muestra a continuacion.

Tabla 2. Talla media(cm) y varianza(cm^2) en niños de ambos

sexos según peso al nacer.

Bajo peso al nacer	Tamaño del grupo	ESTATURA	
		Media	Varianza
SI	273	119.8	41.22
NO	3426	121.9	33.52

Vamos a denotar con X e Y las variables aleatorias, estatura a los 7 años en niños que tuvieron y no tuvieron bajo peso al nacer, respectivamente.

Supongamos que $X \sim N(\mu_x, \square^2)$ y $Y \sim N(\mu_y, \square^2)$, (debe notarse que en ambos modelos de distribución hemos incorporado la suposición de iguales varianzas)

Sería de interés evaluar la diferencia entre $\bar{x}$ y $\bar{y}$ para saber si es significativa, lo que estaría indicando que los dos grupos de niños difieren significativa-mente en la estatura alcanzada a los 7 años.

En términos de una prueba de hipótesis el problema puede quedar planteado como sigue: H_0: $\mu_x = \mu_y$ Vs. H_1: $\mu_x \neq \mu_y$,

O también del siguiente modo: H_0: $\mu_x - \mu_y = 0$ Vs. H_1: $\mu_x - \mu_y \neq 0$

Ahora sólo tendríamos que definir la región crítica.

Tanto $\bar{x}$ como $\bar{y}$ se distribuyen según la ley normal, por lo que se puede demostrar que bajo H_0 su diferencia, que será una estimación de la diferencia $\mu_x - \mu_y$, se distribuye $N[0, \square^2(1/n + 1/m)]$, donde n y m son los tamaños de cada grupo.

De acuerdo al proceder usual de estandarización, $z = (\bar{x} - \bar{y})/[\square \sqrt{1/n + 1/m}]$, cuando σ es conocida, este valor puede compararse entonces con el percentil hallado en la tabla normal estándar partir del conocimiento del nivel de significación , para tomar la decisión de rechazar o no la hipótesis nula.

Definición 12. Sean X e Y variables aleatorias tales que $X \sim N(\mu_x, \square^2)$ y

$Y \sim N(\mu_y, \square^2)$, con $\square^2$ conocida . Suponiendo que se cumple H_0 la región crítica de la prueba de hipótesis:

H_0: $\mu_x - \mu_y = 0$ Vs. H_1: $\mu_x - \mu_y \neq 0$

para un nivel de significación prefijado $\square$ toma cualquiera de las formas siguientes:

1. $\{x,y : \bar{x} - \bar{y} < -z_{1-\square/2} \square \sqrt{1/n + 1/m}$ para $\bar{x} - \bar{y} < 0$ ó

$\bar{x} - \bar{y} > z_{1-\square/2} \square \sqrt{1/n + 1/m}$ para $\bar{x} - \bar{y} > 0\}$.

2. $\{x,y: z < -z_{1-\square/2}$ para $z < 0$ ó $z > z_{1-\square/2}$ para $z > 0\}$.

Ejemplo 20 (continuación). Supongamos que el valor de la desviación estándar común es de 6 cm. Vamos a calcular la región crítica (1).

$\bar{x} - \bar{y}$ = 119.8–121.9 = –2.1, como esta diferencia es negativa utilizaremos la primera desigualdad de la región crítica escrita en la forma (1).

$-z_{1-\alpha/2} \sqrt{1/n + 1/m} = -1.96*6*\sqrt{1/273 + 1/3426} = -0.74$, mayor que –2.1;

lo que indica que se debe rechazar H_0 con una significación del 5% y por ende admitir la posibilidad de que las estaturas procedan de poblaciones distintas.

C. **Problema de muestras independientes con σ^2 desconocida pero iguales para ambas poblaciones**. En esta nueva situación tenemos los mismos supuestos que en el inciso (B) pero con la varianza es desconocida; siendo éste, el estado actual se procede, como habitualmente, a estimar σ^2 utilizando la información de cada una de las dos muestras disponibles.

Podemos denotar por s_x^2 y s_y^2 a la varianza muestral de X e Y respectivamente; a partir de estas estimaciones se calcula una varianza combinada o dependiente de las varianzas de las dos muestras que tiene la forma:

$$s_c^2 = [(n-1)s_x^2 + (m-1)\ s_y^2] / n+m-2. \qquad (9)$$

Entonces se puede demostrar, suponiendo H_0 cierta que,

$$t = (\bar{x} - \bar{y}) / s_c \sqrt{1/n + 1/m} \qquad (10)$$

tendrá una distribución t–Student con n+m–2 grados de libertad.

Ahora tenemos la información necesaria para poder definir la región crítica de esta prueba.

Definición 13. Sean X e Y variables aleatorias tales que $X \sim N(\mu_x, \sigma^2)$ y

$Y \sim N(\mu_y, \sigma^2)$, donde σ^2 se asume igual para ambas poblaciones pero desconocida . Suponiendo que se cumple H_0 la región crítica de la prueba de hipótesis:

H_0: $\mu_x - \mu_y = 0$ Vs. H_1: $\mu_x - \mu_y \neq 0$

para un nivel de significación prefijado α toma cualquiera de las formas siguientes:

1.{x,y: $\bar{x} - \bar{y} < -t_{n+m-2,1-\alpha/2}\ s_c \sqrt{1/n + 1/m}$ para $\bar{x} - \bar{y} < 0$ ó

$\bar{x} - \bar{y} > t_{n+m-2,1-\alpha/2}\ s_c \sqrt{1/n + 1/m}$ para $\bar{x} - \bar{y} > 0$}.

2.{x,y: $t < -t_{n+m-2,1-\alpha/2}$ para $t < 0$ ó $t > t_{n+m-2,1-\alpha/2}$ para $t > 0$}.

Donde s_c y t se dan en (9) y (10) respectivamente y $t_{n+m-2,1-\alpha/2}$ es el percentil

$1-\alpha/2$ de la t–Student con n+m–2 grados de libertad.

Ejemplo 20(continuación). Vamos ahora a suponer σ^2 desconocida e igual en ambas poblaciones. Con los datos de la tabla 2 podemos hallar la desviación estándar combinada.

$s_c^2 = [(n-1)s_x^2 + (m-1)\ s_y^2] / n+m-2 = (272*41.22 + 3425*33.52)/3697$

= 34.09

luego $s_c = \sqrt{34.09} = 5.84$ y $t_{n+m-2,1-\alpha/2} = t_{3697,0.975} = 1.96$ (note que: como n+m es muy grande el valor de la t coincide con el de la distribución normal)

Hallemos ahora la región crítica basada en t.

$t = (\bar{x} - \bar{y}) / s_c \sqrt{1/n + 1/m} = -2.1 / 5.84*0.06 = -5.72$

Como $t < 0$, usamos la primera desigualdad de (1) como región crítica, al ser

$-5.72 < -1.96$ rechazamos H_0 con un nivel de significación del 0.05.

11.9- PRUEBAS DE HIPOTESIS CON PROPORCIONES

Al igual que en el caso de la estimación puntual y la estimación por intervalo de confianza, es posible construir pruebas de hipótesis para verificar suposiciones acerca del comportamiento de variables en la población cuando éstas no están medidas en escala contínua, por lo que no pueden ser apropiadamente resumidas a través de la media aritmética.

Caso I. Decisión sobre π con n suficientemente grande.

Ejemplo 21.Vamos a suponer que se conoce bien la frecuencia con que se presenta una enfermedad determinada en la población bajo condiciones normales, sin ningún tipo de acción sobre ella. Un grupo de investigadores ha desarrollado un tratamiento preventivo novedoso para dicha enfermedad y se aplica en una muestra aleatoria de la población base para comprobar su efectividad.

Si denominamos la proporción conocida de enfermos como π_0, las hipótesis estadísticas para la inferencia sobre π bajo las condiciones del tratamiento preventivo aplicado serían:

H_0: $\pi = \pi_0$ Vs. la alternativa H_1: $\pi \neq \pi 0$. (11)

Con este conjunto de hipótesis podremos probar si el tratamiento nuevo ejerce alguna influencia sobre la frecuencia de presentación de la enfermedad.

Siguiendo similar procedimiento al aplicado en el caso de la decisión sobre μ con σ conocida; fundamentaremos nuestra decisión de rechazar o no H_0 partiendo de la extracción de una muestra aleatoria de tamaño n de la población base y observando el número k de individuos que, a pesar de habérsele aplicado tratamiento preventivo, se enferman; la decisión depende entonces de la relación $k/n = \hat{p}$, es decir de la proporción muestral.

En general lo que queremos es rechazar la hipótesis nula con una seguridad aceptable de no cometer un error, el de rechazar H_0 cuando lo que ahí se afirma es verdad (error de tipo I). Se necesita encontrar entonces, los valores de $\hat{p}$ que nos conducen a rechazar H_0 con probabilidad a lo más α, cuando H_0 es cierta. Como la prueba (11) es bilateral o de dos colas, esos valores pueden encontrarse tanto hacia el extremo izquierdo de π_0 como hacia el derecho. Entonces, similarmente al caso de μ, tenemos que hallar valores c_1 y c_2 tales que: $Pr(\hat{p} < c_1 \text{ ó } \hat{p} > c_2) = \alpha$.

La variable aleatoria $\hat{p}$ tiene una distribución binomial y ya sabemos que bajo la suposición de que n, el tamaño muestral, es lo suficientemente grande cualquier variable binomial puede ser tratada como una variable aleatoria normal.

En este caso bajo H_0, $\hat{p} \sim N[\pi_0, \pi_0(1 - \pi_0)/n]$.

Estandarizando se concluye que el estadigrafo:

$z = (\hat{p} - \pi_0) / \sqrt{\pi_0(1-\pi_0)/n} \sim N(0,1)$.

Con esta información podremos hacer la siguiente definición:

Definición 14. Para la prueba de hipótesis: H_0: $\pi = \pi_0$ Vs. H_1: $\pi \neq \pi_0$,

suponiendo que se cumple H_0, y n es suficientemente grande, para un nivel de significación □, la región crítica puede tomar cualquiera de las formas siguientes:

1.{k: $\hat{p} < \pi_0 - z_{1-□/2} \sqrt{\pi_0(1-\pi_0)/n}$ para $\hat{p} < \pi_0$ ó

$\hat{p} > \pi_0 + z_{1-□/2} \sqrt{\pi_0(1-\pi_0)/n}$ para $\hat{p} > \pi_0$}.

2.{k: $z < -z_{1-□/2}$ para $z < 0$ ó $z > z_{1-□/2}$ para $z > 0$}.

Para la prueba de hipótesis unilateral: H_0: $\pi = \pi_0$ Vs. H_1: $\pi < \pi_0$,

la región crítica puede asumir cualquiera de los tres formas siguientes:

1.{k: $\hat{p} < \pi_0 + z_{□} \sqrt{\pi_0(1-\pi_0)/n}$ }

2.{k: $z < z_{□}$}

3.{k: $p < □$}

y para la prueba de hipótesis, también unilateral: H_0: $\pi = \pi_0$ Vs. H_1: $\pi > \pi_0$,

la región crítica se puede construir de las siguientes maneras:

1.{k: $\hat{p} > \pi_0 + z_{1-□} \sqrt{\pi_0(1-\pi_0)/n}$ }

2.{k: $z > z_{1-□}$}

3.{k: $p < □$}

Ejemplo 21 (continuación). Supongamos que la proporción de enfermos antes del tratamiento preventivo era de 0.3 y que de 100 individuos observados posteriormente sólo 20 enfermaron. ¿Se puede afirmar que es un éxito la innovación del tratamiento preventivo con un nivel de significación del 5%?

Se desea en este problema realizar el contraste de la hipótesis:

H_0: $\pi = 0.3$ vs H_1: $\pi \neq 0.3$.

Primero hay que estimar la proporción de enfermos en la muestra de 100 individuos. Se tiene que,

$\hat{p} = 20/100 = 0.2$ y como $\alpha = 0.05$ entonces $z_{1-□/2} = z_{0.975} = 1.96$

Como $\hat{p}$ está a la izquierda de π_0 tenemos que utilizar:

$\hat{p} < \pi_0 - z_{1-\alpha/2} \sqrt{\pi_0(1-\pi_0)/n}$ (desigualdad izquierda de (1) para una prueba bilateral). Sustituyendo se tiene, $0.2 < 0.3 - 1.96 \sqrt{0.3*0.7/100} = 0.21$.

Como se cumple, $\hat{p}$ está en la región crítica, luego se rechaza H_0. Se puede entonces concluir con un nivel de significación del 5%, que el tratamiento preventivo fue un éxito.

Caso II. Decisión sobre la diferencia de dos proporciones poblacionales.

En este acápite nos dedicaremos a definir la región crítica adecuada para la prueba de hipótesis relacionada con la comparación de proporciones en dos poblaciones; problema que se presenta con bastante frecuencia en la investigación biomédica.

Ejemplo 22. En un municipio de Ciudad de la Habana se decide analizar la relación entre enfermedad diarreica aguda (EDA) en niños menores de 5 años y las condiciones sanitarias del hogar.

Se hizo un índice de condiciones sanitarias que incluyó el suministro de agua potable, apreciación de limpieza del hogar y apariencia personal del niño y de la madre, que se resumió en dos categorías: bien (B) y regular o mal (R/M) y se previo la creación dos grupos de niños, como se muestra a continuacion ;

Grupo 1: comprende los niños menores de 5 años evaluados con B

Grupo 2: abarca los niños del mismo rango de edad evaluados con R/M.

En cada uno de estos grupos se recogió información con respecto a si el niño tuvo algún episodio de EDA en el último año.

Si llamamos π_1 a la proporción poblacional de niños del grupo 1 con EDA y π_2 a la misma proporción pero del grupo 2, podría plantearse la hipótesis:

H_0: $\pi_1 = \pi_2 = \pi$ Vs. H_1: $\pi_1 \neq \pi_2 = \pi$, o en forma equivalente:

H_0: $\pi_1 - \pi_2 = 0$ Vs. H_1: $\pi_1 - \pi_2 \neq 0$

Donde, si se rechaza H_0 se puede hablar, con un nivel de significación α, de que la proporción de enfermos proviene de poblaciones diferentes.

Supongamos que se extraen dos muestras de tamaño n_1 y n_2, ambas bastante grandes. Sean k_1 y k_2 el número de niños con EDA en cada grupo; entonces los estimadores $\hat{p}_1 = k_1/n_1$ de π_1 y $\hat{p}_2 = k_2/n_2$ de π_2 tienen por el teorema del limite central distribuciones que se aproximan por $N[\pi_1, \pi_1(1-\pi_1)/n_1]$ y $N[\pi_2, \pi_2(1-\pi_2)/n_2]$ respectivamente.

Sustentado en lo anterior y bajo H_0: $\pi_1 = \pi_2 = \pi$, se concluye que:

$$\hat{p}_1 - \hat{p}_2 \sim N[0, \pi(1-\pi)(1/n_1 + 1/n_2)].$$

Como π es desconocido se usa la información de ambas muestras para calcular un estimador adecuado para π, que podemos llamar $\hat{p}$, y a partir de ese valor calcular la varianza combinada.

Se puede demostrar que bajo H_0, el estimador de π viene dado por

$$\hat{p} = (n_1 \hat{p}_1 + n_2 \hat{p}_2) / (n_1 + n_2)$$

Si en esta fórmula sustituimos las expresiones de $\hat{p}_1$ y $\hat{p}_2$ en términos de k_1, k2, n_1 y n_2 tendremos que: $\hat{p}$ = $(n_1 k_1/n_1 + n_2 k_2/n_2) / (n_1 + n_2)$, donde simplificando, se llega a: $\hat{p} = (k_1 + k_2) / (n_1 + n_2)$, significando $k_1 + k_2$ el número total de enfermos.

Es decir que el estimador de π bajo H_0 viene dado por el cociente entre el total de enfermos y el total de niños en ambos grupos; un resultado muy comprensible y que además, facilita el cálculo de $\hat{p}$.

La varianza combinada se estima de la siguiente forma:

$$var_c = \hat{p}\ \hat{q}\ (1/ n_1 + 1 / n_2), \text{ donde } \hat{q} = 1 - \hat{p}$$

Siguiendo el procedimiento habitual para una prueba de hipótesis bilateral, se trata de buscar aquelos valores c_1 y c_2 tales que:

$Pr (\hat{p}_1 - \hat{p}_2 < c_1$ ó $\hat{p}_1 - \hat{p}_2 > c_2) = \square$ o similarmente

$Pr (\hat{p}_1 - \hat{p}_2 < c_1) = Pr (\hat{p}_1 - \hat{p}_2 > c_2) = \square/2$

Como $z = (\hat{p}_1 - \hat{p}_2) / \sqrt{var_c} \sim N(0,1)$, se tiene que $c_1 = -z_{1-\square/2} \sqrt{var_c}$ y $c_2 = z_{1-\square/2} \sqrt{var_c}$, donde $z_{1-\square/2}$ es el percentil $1-\square/2$ de la distribución normal estándar; podemos hacer entonces la siguiente definición:

Definición 15. Para la prueba de hipótesis:

$$H_0: \pi_1 - \pi_2 = 0 \text{ Vs. } H_1: \pi_1 - \pi_2 \neq 0,$$

Bajo la consideración de que: se cumple H_0, y que n_1 y n_2 son suficientemente grandes. Para un nivel de significación $\square$ dado$\square$, la región crítica puede tomar cualquiera de las formas siguientes:

1. $\{k_1, k_2: \hat{p}_1 - \hat{p}_2 < -z_{1-\square/2} \sqrt{var_c}$ para $\hat{p}_1 - \hat{p}_2 < 0$ ó

$\hat{p}_1 - \hat{p}_2 > z_{1-\square/2} \sqrt{var_c}$ para $\hat{p}_1 - \hat{p}_2 > 0\}$

2. $\{k_1, k_2: z < -z_{1-\square/2}$ para $z < 0$ ó $z > z_{1-\square/2}$ para $z > 0\}$.

Ejemplo 22 (continuación). En la tabla 3 que se muestra, se resumen los datos correspondientes a 630 niños menores de 5 años seleccionados de cada uno de los grupos considerados, clasificados de acuerdo a si fueron o no afectados por EDA en el último año.

Tabla 3. Proporción de niños menores de 5 años enfermos de EDA en cada grupo.

EDA	Índice de condiciones sanitarias		Total
	B	R/M	
SI	92	107	199
NO	223	208	431
Total	315	315	630

La hipótesis a contrastar será:

H_0: $\pi_1 - \pi_2 = 0$ Vs. H_1: $\pi_1 - \pi_2 \neq 0$

Si ponemos □ = 0.05, entonces $z_{1-0.05/2} = z_{0.975} = 1.96$.

Con las cifras de la tabla 3 se puede proceder a calcular $\hat{p}_1$, $\hat{p}_2$ y $\hat{p}$:

Se tiene que:

$\hat{p}_1$(proporción de niños menores de 5 años con índice B) = 92/315 = 0.292.

$\hat{p}_2$(proporción de niños menores de 5 años con índice R/M) = 107/315.

= 0.340.

$\hat{p}$ (proporción total de niños con EDA) = 199 / 630 = 0.316.

De este ultimo valor se deduce que $\hat{q} = 1 - \hat{p} = 1 - 0.316 = 0.684$.

Calculemos ahora $var_c = \hat{p}\,\hat{q}\,(1/n_1 + 1/n_2)$.

Sustituyendo valores, $var_c = 0.316*0.684*(1/315 + 1/315) = 0.0014$.

Valor este cuya raíz cuadrada es 0.037.

Calculemos ahora $\hat{p}_1 - \hat{p}_2$ para ver por medio de su signo que desigualdad de la región critica de (2) empleamos para la toma de decisión.

Tenemos que $\hat{p}_1 - \hat{p}_2 = 0.292 - 0.34 = -0.048$, luego debemos emplear la desigualdad de la izquierda, pero antes debemos hallar el valor del percentil z correspondiente al nivel de significación 0.05.

Este es $z_{1-□/2} = z_{0.975} = 1.96$, luego $-z_{1-□/2}\sqrt{var_c} = -1.96 \cdot 0.037 = -0.073$.

Como –0.048 no es menor que – 0.073, no se puede rechazar H_0 con el nivel de significación previsto; por consiguiente la diferencia de casi el 5% de enfermos hallada entre ambos grupos no es suficiente para avalar que los niños procedan de poblaciones diferentes.

Otra forma de proceder consiste en calcular z = –0.048/0.037 = –1.3 y siendo z negativo compararlo con $-z_{0.975} = -1.96$, luego como –1.3 es mayor que

–1.96, no se puede rechazar H_0, y se concluye lo ya expresado con anterioridad.

11.10.- EJERCICIOS RESUELTOS.

Parte A: Estimación.

1. Asumiendo que la desviación estándar es conocida y con valor de 0.1, calcule un intervalo de confianza del 95% para la verdadera media de la fracción de eyección del ventrículo izquierdo (variable FEVI), considerando para ello una muestra de 27 pacientes con cardiomiopatía dilatada aguda. Los datos se muestran a continuación:

0.19, 0.24, 0.24, 0.18, 0.17, 0.22, 0.40, 0.23, 0.40

0.14, 0.23, 0.14, 0.20, 0.30, 0.20, 0.07, 0.30, 0.12

0.19, 0.13, 0.24, 0.17, 0.32, 0.24, 0.32, 0.19, 0.28

Solución:

Como □□es conocida, para calcular el intervalo se puede aplicar, la expresión $\bar{x} \pm z_{1-\alpha/2} \cdot \sigma/\sqrt{n}$, donde primero hay hallar el estimador puntual de $\bar{x}$.

= (0.19 + 0.24 + ... + 0.18 + 0.28)/ 27 = 6.05/27 = 0.224

$\bar{x} = \sum x_i / n$ Como datos se tiene que, □□= 0.1, n = 27, □□= 1 – 0.95 = 0.05, luego □/2 = 0.025, y en consecuencia, $z_{1-\alpha/2} = z_{0.975}$, pero como la tabla con la que se trabaja en el libro es la correspondiente a los percentiles de la distribución normal para el área bajo la curva entre 0 y z ello equivale a buscar $z_{0.475}$ =1.96, luego sustituyendo en la formula se tiene que:

LI = $0.224-1.96 \cdot 0.1/\sqrt{27}$ LS = $0.224+1.96 \cdot 0.1/\sqrt{27}$

LI = 0.224-0.038 LS = 0.224+0.038

LI = 0.186 LS = 0.262

Respuesta. El intervalo de confianza del 95% para la media verdadera de FEVI es (0.186,0.262), valor que se espera este contenido en el intervalo con un 95% de confiabilidad.

2.Responda el ejercicio anterior pero sin asumir que la desviación estándar es conocida.

Solución:

En este caso ya no se puede aplicar la formula anterior, sino que la que se aplica es $\bar{x} \pm t_{n-1,1-(\alpha/2)}\, s/\sqrt{n}$.

Luego hay que estimar s^2 a partir de los datos de la tabla, para lo cual se tiene que:

+ $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{1.522 - 27 * 0.224^2}{26} = 0.006$ Donde, $\sum (x_i)^2 = 0.19^2 + 0.24^2 + ... + 0.18^2 + 0.28^2 = 1.522$

Por lo que s = 0.08.

Como n<30 y □□es desconocida, para calcular los límites del intervalo de confianza se desglosa la expresión general en dos, que identificaremos por LI y LS respectivamente, a saber:

$$LI = \bar{x} - t_{n-1,1-(\alpha/2)} \cdot s/\sqrt{n} \text{ y } LS = \bar{x} + t_{n-1,1-(\alpha/2)} \cdot s/\sqrt{n}$$

Donde, sustituyendo se tiene que:

$LI = 0.224 - t_{26,0.975} \cdot 0.08/\sqrt{27}$ $\quad$ $LS = 0.224 + t_{26,0.975} \cdot 0.08/\sqrt{27}$

$LI = 0.224 - 2.056 \cdot 0.08/5.2$ $\quad$ $LS = 0.224 + 2.056 \cdot 0.08/5.2$

$LI = 0.192$ $\quad$ $LS = 0.256$

Respuesta. El intervalo de confianza del 95% para la media verdadera de FEVI es (0.192,0.256).

3. En un estudio, 89 de 283 mujeres con infertilidad tuvárica primaria, reportaron haber usado algún dispositivo intrauterino (DIU); mientras que 640 de 3833 mujeres fértiles también reportaron su uso.

a. Calcule los estimadores puntuales del porcentaje de mujeres que, en uno y otro grupo, usaron DIU.

b. Construya los intervalos de confianza del 95% para las estimaciones puntuales halladas en a.

c. ¿Podrían calcularse intervalos del 90% de confianza? Explique su respuesta.

Solución:

a. Tenemos dos grupos: mujeres fértiles (A) y mujeres infértiles (B). Las estimaciones puntuales del porcentaje son:

$$\hat{P}_A = \frac{89 \cdot 100}{283} = 31.45\% \quad \text{y} \quad \hat{P}_B = \frac{640 \cdot 100}{3833} = 16.70\%$$

b Hay que verificar si, para cada grupo, $n \cdot \hat{p} \cdot \hat{q} > 5$

Grupo A: $283*0.3145*0.6855 \approx 61$

Grupo B: $3833*0.167*0.833 \approx 533$

Es decir que para ambos grupos es válido utilizar la expresión de calculo para el intervalo dada por: $\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{p} \cdot (1-\hat{p})/n}$, solo que, los valores a ser sustituidos se interpretaran como porcientos.

Intervalo de confianza para el grupo A:

$LI = 31.45 - 1.96 \cdot \sqrt{31.45 \cdot 68.55/283}$ $\quad$ $LS = 31.45 + 1.96 \cdot \sqrt{31.45 \cdot 68.55/283}$

$LI = 31.45 - 5.41$ $\quad$ $LS = 31.45 + 5.41$

$LI = 26.04\%$ $\quad$ $LS = 36.86\%$

Intervalo de confianza para el grupo B:

$LI = 16.70 - 1.18$ $\quad$ $LS = 16.70 + 1.18$

$LI = 15.52\%$ $\quad$ $LS = 17.88\%$

Respuesta: El intervalo de confianza del 95% para el porcentaje de mujeres que usaron DIU entre las mujeres infértiles es (26.04%,36.86%); es decir, se espera, con una probabilidad de 0.95, que el verdadero valor del porcentaje sea una cifra comprendida entre esos límites.

El intervalo de confianza del 95% entre las mujeres fértiles va desde 15.52% hasta el 17.88%.

c. Si, siempre es posible encontrar un intervalo de confianza para cualquier nivel deseado; sólo hay que buscar el percentil apropiado. En este caso, como queremos 1– □□= 0.90, □□= 0.1 y □/2 = 0.05; hay que buscar en la tabla de la distribución normal el percentil $z_{0.95}$; que es 1.64.

4. Se realizó un estudio de casos y controles (en este tipo de estudios se seleccionan los individuos que tienen una enfermedad, o una condición dada (casos) para compararlos con otro grupo de individuos donde la enfermedad o condición no está presente (controles)), acerca de la efectividad de la prueba de Papanicolau en la prevención del cáncer cervical. Se encontró que el 28.1% de los 153 casos con cáncer cervical y el 7.2% de los 153 controles sanos nunca se habían realizado dicha prueba antes del momento en que se hizo el diagnóstico de los casos.

a. Calcule un intervalo del 95% de confianza para el porcentaje de casos con cáncer cervical que nunca se hicieron la citada prueba antes del diagnóstico.

b. Construya el intervalo de confianza, también del 95%, para el porcentaje de los controles que nunca se hicieron la prueba.

c. De acuerdo a los resultados anteriores, ¿piensa Ud. que la prueba es útil en la prevención del cáncer cervical?

Solución:

a. Se tiene que $\hat{P}$ = 28.1%, por lo que $\hat{Q}$ = 71.9% con n = 153.

Se cumple que $n \cdot \hat{p} \cdot \hat{q} > 5$; luego puede aplicarse la aproximación a la normal.

LI = $28.1 - 1.96 \cdot \sqrt{28.1 \cdot 71.9 / 153}$ LS = $28.1 + 1.96 \cdot \sqrt{28.1 \cdot 71.9 / 153}$

LI = 28.1 – 7.1 LS = 28.1 + 7.1

LI = 21.0% LS = 35.2%

El intervalo de confianza del 95% para el porcentaje de casos que nunca se hicieron la prueba va desde el 21.0 hasta el 35.2%.

b. En este caso también puede aplicarse la aproximación a la normal y los límites inferior y superior del intervalo son:

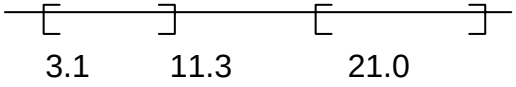
LI = $7.2 - 1.96 \cdot \sqrt{92.8 \cdot 7.2 / 153}$ LS = $7.2 + 1.96 \cdot \sqrt{92.8 \cdot 7.2 / 153}$

LI = 7.2 – 4.1 LS = 7.2 + 4.1

LI = 3.1% LS = 11.3%

El intervalo de confianza del 95% para el porcentaje de controles que nunca se hicieron la prueba va del 3.1% al 11.3%.

c. Para mejor comprensión vamos a representar los intervalos obtenidos en los incisos a y b en una línea recta:



Como se ve ambos intervalos están bien separados, el límite inferior del intervalo para los casos con cáncer cervical ubicado bien hacia la derecha del límite superior del intervalo correspondiente a los controles sanos en el momento del diagnóstico de los casos. Mientras que a lo más el 11.3% de los controles nunca se habían hecho el Pap., al menos el 21.0% de las enfermas nunca se lo habían hecho; parece entonces razonable pensar que la prueba es útil en la prevención del cáncer cervical.

5. A continuación se presentan los resultados acerca de los cambios en la creatinina sérica experimentados por un grupo de pacientes de diálisis en estadio final de insuficiencia renal, evaluados al inicio del comienzo de las diálisis y 18 meses después:

	Inicio (n=102)		Final (n=69)	
	Media	DE	Media	DE
Creatinina sérica (mmol/L)	0.97	0.22	1.00	0.19

Construya el intervalo de confianza del 95% para la media de creatinina sérica al inicio de las diálisis.

Solución:

Se desconoce la desviación estándar poblacional, pero n = 102; para el intervalo de confianza del 95% habrá que usar la fórmula $\bar{x} \pm 1.96 \cdot s/\sqrt{n}$.

$LI = 0.97 - 1.96 \cdot 0.22/\sqrt{102}$ $\quad$ $LS = 0.97 + 1.96 \cdot 0.22/\sqrt{102}$

$LI = 0.97 - 0.04$ $\quad$ $LS = 0.97 + 0.04$

$LI = 0.93$ $\quad$ $LS = 1.01$

Respuesta: El intervalo de confianza para la media de creatinina sérica al inicio de las diálisis, en pacientes con insuficiencia renal, es (0.93 mmol/L, 1.01 mmol/L).

6.Se realiza el pesquizaje de pacientes hipertensos entre los residentes en el área perteneciente a un Consultorio del Médico de Familia y se les indica captopril. A la semana siguiente son visitados de nuevo y se les toma la presión. Los datos correspondientes a la presión en el momento del pesquizaje y una semana, después se dan en la tabla 2.

Tabla 2. Presión arterial sistólica de hipertensos, tomada en el momento del pesquizaje y una semana después de tomar captopril.

Paciente	Sistólica		Paciente	Sistólica	
Número	Inicio	Semana	Número	Inicio	Semana
1	200	188	6	225	222

Paciente Número	Sistólica Inicio	Semana	Paciente Número	Sistólica Inicio	Semana
2	194	212	7	203	190
3	236	186	8	180	154
4	163	150	9	177	180
5	240	200	10	240	225

a. Halle las estimaciones puntuales de la media y la desviación estándar de las diferencias entre la presión arterial sistólica inicial y la correspondiente a una semana después.

b. Asuma que dicha diferencia se distribuye normal con media desconocida μ y desviación estándar conocida 20 mmHg. Halle un intervalo de confianza del 95% para μ.

Solución:

a. Llamemos d_i , donde i = 1,....,10, a las diferencias entre la presión arterial antes y después del captropil. Entonces:

d_1=12 d_3=50 d_5=40 d_7=13 d_9=-3

d_2=-18 d_4=13 d_6=3 d_8=26 d_{10}=15

de aquí que $\sum d_i = 151$ y $\sum d_i^2 = 5825$

La estimación puntual para la media de las diferencias será: $\bar{d} = 15.1$ y para la desviación estándar

$$s = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{5825 - 10(15.1)^2}{9}} = 19.85$$

b. Tenemos que n es 10 y □□es conocida, luego el intervalo del 95% vendrá dado por $\bar{d} \pm 1.96\square/\sqrt{n}$, por tanto los limites inferior y superior serán

LI =15.1–1.96·20/$\sqrt{10}$ =2.7 LS =15.1+1.96·20/$\sqrt{10}$ =27.5

Es decir que el intervalo de confianza del 95% para la media de la diferencias entre las presiones arteriales sistólicas antes y después del captropil va desde 2.7 mmHg hasta 27.5 mmHg.

7. Suponga que deseamos estimar la concentración en orina de una dosis específica de ampicillina después de cierto período. Se reclutan 25 voluntarios y se encuentra que tienen una concentración media de 7.0μg/mL con una desviación estándar de 2.0μg/mL. Asuma que la población base de concentraciones tiene una distribución normal.

a. Encuentre un intervalo de confianza del 99% para la concentración media en la población.

b. ¿Qué tamaño de muestra se necesitará para garantizar un intervalo de longitud igual a 0.5μg/mL, si se asume que la desviación estándar poblacional es también 2.0μg/mL?

Solución:

a. n = 25, $\bar{x} = 7.0$ y s = 2.0

Como la media y la desviación de la población son desconocidos y n<30 hay que usar:

$LI = \bar{x} - t_{n-1,1-(\alpha/2)} \cdot (s/\sqrt{n})$ $LS = \bar{x} + t_{n-1,1-(\alpha/2)} \cdot (s/\sqrt{n})$

α = 0.01 por lo que $1-\alpha/2 = 0.995$

$t_{24,0.995} = 2.797$, luego

$LI = 7.0 - 2.797 \cdot 2/\sqrt{25}$ $LS = 7.0 + 2.797 \cdot 2/\sqrt{25}$

LI = 5.9 μg/mL LS = 8.1 μg/mL

b. Tenemos que usar $n = (z_{1-\alpha/2} \cdot \sigma / d)^2$, donde d (precisión), es la mitad de la longitud del intervalo de confianza. Si deseamos que la longitud total sea 0.5, necesitaremos d = 0.5/2 = 0.25μg/mL. Puede verificarse que $z_{0.995} = 2.58$; sustituyendo encontramos que:

$$n = [2.58 \cdot 2/0.25]^2 = 426$$

Respuesta: La longitud deseada se puede garantizar con una muestra de tamaño 426.

Parte B: Prueba de hipótesis.

8. Un fabricante de una medicina alega que la misma es efectiva en un 90% para aliviar la alergia por un período de 8 horas. En una muestra de 200 personas que tenían alergia, la medicina proporcionó mejoría a 160. Diga si lo alegado por el fabricante es cierto. Use α = 0.01.

Solución:

Sea π la proporción poblacional de personas que mejoran su alergia usando la medicina. Estamos interesados en comprobar si la proporción de personas aliviadas por la medicina es menor que lo que se alega, puesto que si fuese igual o mayor el resultado respaldaría lo que afirma el productor. Entonces escogemos la hipótesis:

H_0: $\pi = 0.9$ vs H_1: $\pi < 0.9$

La región crítica para esta prueba es $\{k: z < z_{\alpha}\}$, con

$z = (\hat{p} - \pi_0) / \sqrt{\pi_0(1-\pi_0)/n}$.

Bajo H_0, $\pi_0 = 0.9$, luego $\sqrt{\pi_0(1-\pi_0)/n} = \sqrt{0.9*0.1/200} = 0.021$

Hallemos el valor de $\hat{p}$, la estimación de π: $\hat{p} = 160/200 = 0.8$.

Sustituyendo, z = (0.8 – 0.9) / 0.021 = – 0.1/0.021 = – 4.71.

Como α = 0.01, entonces $z_{0.01} = -z_{0.5-0.01} = -z_{0.49} = -2.33$

Como – 4.71 < –2.33 llegamos a la conclusión, de que lo alegado por el fabricante no se cumple usando un nivel de significación del 1%,.

9. Dos grupos de 100 personas cada uno, padecen una enfermedad. Al grupo 1, además del tratamiento habitual, se le aplica un suero y el grupo 2 (grupo control) permanece con el tratamiento habitual. Al final del ensayo se encuentra que 75 pacientes del grupo 1 y 65 del grupo 2 se recuperaron. ¿Cree usted que existe una diferencia significativa entre los que se recuperan en uno y otro grupo? Responda la pregunta para:

a. Nivel de significación de 0.05.
b. Nivel de significación de 0.01.
c. Comente los resultados alcanzados en los incisos a y b.

Solución:

Si se considera que π_1 represente la proporción poblacional de personas que se recuperan cuando se usa el suero y π_2 lo mismo pero sin usar el suero. Entonces podemos plantear la hipótesis: H_0: $\pi_1 = \pi_2$ vs H_1: $\pi_1 \neq \pi_2$

Calculemos la diferencia entre las proporciones estimadas de pacientes que se recuperaron en cada grupo.

$\hat{p}_1 - \hat{p}_2 = 75/100 - 65/100 = 0.1$

Como esta diferencia es positiva tenemos que utilizar, la desigualdad:

$$\hat{p}_1 - \hat{p}_2 > z_{1-\alpha/2} \sqrt{var_c} \qquad (1)$$

como región crítica, donde var_c es la varianza combinada hallada a partir de la información de ambas muestras.

$\hat{p} = (75 + 65) / (100 + 100) = 140/200 = 0.7$ $var_c = 0.7 \cdot 0.3 \cdot 2 /200 = 0.0021$

Con $\alpha = 0.05$, se tiene que :

$z_{1-\alpha/2} = z_{0.5-0.05/2} = z_{0.475} = 1.96$, entonces el producto en la desigualdad (1) es igual a : $1.96 \cdot \sqrt{0.0021} = 0.09$ y como $0.1 > 0.09$ se rechaza la hipótesis nula y se habla de una diferencia significativa entre las personas que se recuperan a favor del grupo donde se introdujo el suero.

Con $\alpha = 0.01$, entonces:

$z_{1-\alpha/2} = z_{0.5-0.01/2} = z_{0.495} = 2.57$, luego el producto en la desigualdad (1) es igual a : $2.57 \cdot \sqrt{0.0021} = 0.12$ y como $0.1 < 0.12$ no se rechaza la hipótesis nula y no se puede hablar de una diferencia significativa entre las personas que se recuperan en uno y otro grupo.

Como el cambio en el nivel de significación produjo tomar decisiones distintas con respecto a la hipótesis nula, por ello es muy recomendable, en casos como estos donde el valor crítico observado está muy cerca del límite de la región crítica, calcular p. Puede comprobarse en la tabla de la distribución normal estándar que el área acumulada hasta $z = 0.1/\sqrt{0.0021} = 2.18$ es 0.0146; es decir que $0.05 > p > 0.01$. Este es un ejemplo de la importancia que tiene reportar el nivel de significación observado.

10. Supongamos que en un determinado país el peso neonatal de los varones tiene una distribución normal con una media de 3.3 Kg. y una desviación estándar de 0.5 Kg.. Supongamos, además, que en una muestra aleatoria de 100 varones recién nacidos, todos procedentes de un determinado subgrupo étnico, el peso medio fue de 3.2 Kg.. Se desea determinar si el peso medio neonatal de este subgrupo étnico difiere del peso medio neonatal del país.

a. Plantee la prueba de hipótesis adecuada a este objetivo.
b. En base a la prueba del inciso a tome la decisión que corresponda al utilizar un nivel de significación del 5%.

Solución.

a. Sea μ el peso neonatal promedio en todo el país en su conjunto, de acuerdo con lo que se desea determinar la prueba de hipótesis será:

   H_0: $\mu = 3.3$ Vs. H_1: $\mu \neq 3.3$, con σ conocida e igual a 0.5

b. Para la prueba del inciso (a), el estadígrafo adecuado es $z = \dfrac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$

Sustituyendo los valores en la expresión de z, se tiene que:

$z = (3.2 - 3.3)/0.5/\sqrt{100} = -0.1/0.05 = -2.$

Como $z < 0$ hay que comparar con $-z_{0.5-0.05/2} = -z_{0.475} = -1.96$ y puesto que $-2 < -1.96$, entonces se rechaza la hipótesis nula y se admite que el peso medio neonatal del subgrupo étnico difiere del de la población en su conjunto para un nivel de significación del 5%.

11. Una empresa farmacéutica anuncia un medicamento cuya acción se mantiene sin vencimiento durante un tiempo promedio de 20 meses. Un laboratorio de experimentación lo probó con una muestra de tamaño 5 y obtuvo los resultados siguientes:

Tiempo promedio de vencimiento en meses: 19, 18, 22, 20 y 17.

¿Podría decir el laboratorio, con un nivel de significación del 5%, que el anuncio de la empresa no es correcto?

Solución.

Tenemos que asumir que el tiempo de vencimiento es una variable aleatoria con distribución normal; entonces, para poder responder la pregunta acerca de la afirmación que hace la empresa se puede plantear la hipótesis:

$H_0: \mu = 20$ vs. $H_1: \mu < 20$, con □□desconocida

puesto que, si el tiempo de vencimiento fuese mayor de 20 meses también sería favorable a lo que afirma la empresa farmacéutica.

Como se desconoce el valor de □□en la población base y, además, la muestra es bastante pequeña, para probar la hipótesis se deberá usar el estadígrafo

$$t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} \qquad (1)$$

Calculemos, $\bar{x}$ y s.

$\bar{x} = (19+18+22+20+17)/5 = 19.2$

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{1858 - 5 \cdot 19.2^2}{4} = 3.7 \text{ y } s = 1.92$$

Sustituyendo en (1) se tiene que $t = -0.93$

Tenemos que hallar en la tabla de la distribución t de Student el valor del percentil 1–□□, es decir $t_{n-1,1-\square\square} = t_{4,0.95} = 2.132$. En el caso de la prueba que estamos considerando la región crítica es: $\{x: t < -t_{n-1,1-\square}\}$ y como –0.93 no está en la región crítica no se puede rechazar la hipótesis nula, y el laboratorio no puede decir que lo que afirma la empresa no sea correcto.

12. Consideremos una muestra aleatoria de 257 individuos que fueron ingresados en un hospital de determinada región. La causa de hospitalización de 23 de ellos fue trastorno locomotor. Si la tasa de enfermedades locomotoras en hospitales semejantes de la región es de 5%, ¿puede considerarse que esta muestra sigue el patrón de enfermedad locomotora característico de la región? Utilice un nivel de significación de 0.05.

Solución.

Se π la proporción poblacional de individuos que ingresan con trastorno locomotor. La prueba de hipótesis adecuada sería:

$$H_0: \pi = 5\% \text{ Vs. } H_1: \pi \neq 5\%$$

Por la información que se posee $\hat{p} = 23/257 = 0.089$; $\hat{q} = 0.911$ y el producto $n \cdot \hat{p} \cdot \hat{q} = 20.8 > 5$ por lo que se puede usar el estadígrafo

$z = (\hat{p} - \pi_0)/ \sqrt{\pi_0(1-\pi_0)/n} = (8.9 - 5)/ \sqrt{5 \cdot 95/257} = 3.9/1.36 = 2.87$, siendo z positivo hay que compararlo con $z_{0.5-0.05/2} = z_{0.475} = 1.96$ luego se rechaza la hipótesis nula con un nivel de significación de 0.05 y no se puede considerar que la tasa de trastornos locomotores para ese hospital en específico siga el patrón de la región.

13. En un hospital clínico quirúrgico se hizo un estudio en las salas del servicio de cirugía general acerca de la estancia postoperatoria. Para ello se estudiaron las historias clínicas de 666 pacientes que habían pasado por el servicio durante un período de 1 año, y se recogió información sobre un conjunto de variables, entre ellas, edad, sexo y número de días que el paciente permaneció hospitalizado después de la operación. Los resultados de las estimaciones puntuales de los días de estancia media posoperatoria y su desviación estándar según edad y sexo se muestran en la tabla 1 siguiente:

Tabla 1. Media (en días) y desviación estándar de la estancia posoperatoria en un servicio de cirugía general.

Variable	Media	Desviación estándar	Número de Pacientes
Edad			
Hasta 45 años	5.0	6.7	249
46 años y más	4.5	3.5	417
Sexo			
Masculino	5.2	7.7	337
Femenino	4.6	4.0	329

¿Cree usted que halla razón para afirmar que la estadía posoperatoria se comporta diferente en cada grupo de edad? Utilice un □ □de 0.05.

Solución.

Supongamos que el tiempo de estadía en cada grupo es una variable aleatoria con distribución normal. Designemos por A y B respectivamente, a los pacientes de cada grupo de edad; si la diferencia entre los promedios de estadía en los grupos A y B resulta significativa y bajo la suposición de que los grupos son homogéneos para cualquier otra característica se puede hablar de un comportamiento diferenciado de la estadía por grupo de edad por lo que la hipótesis a verificar sería:

$H_0: \mu_A = \mu_B$

$H_1: \mu_A \neq \mu_B$, con □ desconocida pero igual en ambas poblaciones.

La diferencia entre las medias muestrales es 0.5 días, por lo que la comparación se realiza con $t_{n+m-2,1-\alpha/2} s_c \sqrt{1/n + 1/m}$. Como α es 0.05 el percentil de la t de Student que debemos buscar es el 0.975 con n+m-2 = 664 grados de libertad. Puede verificarse que ese valor es 1.96 (note que para un número tan alto de grados de libertad se puede trabajar indistintamente con los percentiles de la distribución de Student o con los de la normal).

Calculemos ahora s_c

$$s_c = \sqrt{\langle (n-1)s_A^2 + (m-1)s_B^2 \rangle / \langle n+m-2 \rangle}$$
$$= \sqrt{248 \cdot 6.7^2 + 416 \cdot 3.5^2 / 664} = 4.94$$

Entonces, $1.96 \cdot 4.94 \cdot \sqrt{1/249 + 1/417} = 0.78$ y como la diferencia entre las medias muestrales entre los grupos A y B está en la región de no rechazo, 0.5 < 0.78, no existe razón para afirmar que la estadía posoperatoria tenga un comportamiento diferente según los grupos de edad del paciente intervenido quirúrgicamente.

13. Como parte de un programa de instrucción de hábitos alimentarios 10 hombres entre 25 y 34 años adoptaron una dieta vegetariana por un mes. Durante la dieta el promedio diario de ingesta de ácido linoleico fue de 13g. Si la ingesta promedio entre hombres de la misma edad en la población general es de 15g con desviación estándar de 4g, entonces, con un nivel de significación de 0.05, pruebe la hipótesis de que la ingesta promedio de ácido linoleico en el grupo de 10 hombres es más baja que en la población. Calcule, además, el valor p para la prueba de hipótesis.

Solución:

La hipótesis nula sería H_0: $\mu = 15$ versus la hipótesis alternativa H_1: $\mu < 15$ con desviación estándar conocida igual a 4g.

Para $\alpha = 0.05$, $z_{0.05} = -z_{0.45} = -1.64$, luego $z = (13-15)\sqrt{10}/4 = -1.58$

La región crítica de la prueba es $\{x: z < z_\alpha\}$ y como –1.58 no cae en esta región no se puede decir, con una significación de 0.05, que la ingesta promedio de ácido linoleico sea más baja en este grupo de hombres que en la población general.

14. Asuma que la distribución de pesos al nacer en la población general es normal con media 7.5 lbs y desviación estándar de 1.25 lbs. Se desea probar una droga que, al ser administrada a las madres en el período prenatal, reduce el número de niños con bajo peso al nacer. Se seleccionan 20 madres, se les administra la droga y se obtiene que el peso medio de los 20 recién nacidos es de 7.6 lbs. Pruebe, para un nivel de significación de 5%, la hipótesis de que el peso medio en este grupo de niños es más alto que en la población general.

Solución:

Se pide probar la hipótesis:

H_0: $\mu = 7.5$ Vs. H_1: $\mu > 7.5$, con desviación estándar conocida e igual a 1.25

Para $\alpha = 0.05$, el percentil correspondiente de la normal estándar es 1.64; luego $z = (7.6-7.5)\sqrt{20}/1.25 = 0.36$

La región crítica de la prueba es $\{x: z > z_{1-\alpha}\}$ y como 0.36 no es mayor que 1.64 no se rechaza la hipótesis nula; es decir el peso promedio del grupo de 20 niños no es mayor que el de la población.

15. El nivel de creatinina sérica en sangre se considera como un buen indicador de la presencia o ausencia de afección del riñón. Las personas sanas tienen generalmente concentraciones bajas de creatinina sérica, mientras que las personas enfermas tienen concentraciones altas. Suponga que se desea investigar la relación entre abuso de analgésico y alteración de la función renal. En particular, suponga que se investigan 15 trabajadores de una fábrica, que son conocidos por el abuso que hacen del uso de analgésicos, y se les mide los niveles de creatinina. Estos son: 0.9, 1.1, 1.6, 2.0, 0.8, 0.7, 1.4, 1.2, 1.5, 0.8, 1.0, 1.1, 1.4, 2.2 y 1.4.
Si se asume que los niveles de creatinina para las personas sanas se distribuye normal con media 1.0 y desviación estándar 0.4

a. ¿Podemos hacer algún comentario acerca de los niveles de creatinina entre los que abusan de analgésicos y la población sana a través de alguna prueba estadística?
b. Responda la pregunta anterior si asume que la desviación estándar de la población no es conocida.

Solución.

a. Podemos utilizar la prueba de hipótesis H_0: $\mu = \mu_0$ Vs. H_1: $\mu > \mu_0$ pues esperamos que esos 15 trabajadores al ser clasificados como consumidores en exceso de píldoras analgésicas, tengan niveles mayores de creatinina que la población sana. Con desviación estándar conocida y bajo H_0, el estadígrafo $z = \frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$ tiene una distribución normal con parámetros 0 y 1; por tanto si se calcula z y lo comparamos con $z_{1-\square}$ se sabrá si podemos rechazar la hipótesis nula o no.

b. Podemos usar la misma hipótesis pero como se desconoce el valor de la desviación estándar y $n < 30$ hay que utilizar el estadígrafo $t= \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$ y hacer la comparación con el percentil 1-□ □de la t de Student con n-1 grados de libertad para tomar la decisión apropiada. Si $t > t_{n-1,1-\square}$ □se rechaza H_0 y si $t < t_{n-1,1-\square}$, no se rechaza H_0.

16. Es usual en estudios epidemiológicos donde se incluye el peso corporal que éste se toma directamente; sin embargo, si las personas son entrevistadas en el hogar, se les pide que digan su peso. Suponga que se lleva a cabo un estudio en 10 personas para probar la comparabilidad de los dos métodos. Los datos recogidos se ofrecen en la tabla 2.

Tabla 2. Peso reportado por el entrevistado y peso tomado directamente en lbs.

Sujeto N_0	1	2	3	4	5	6	7	8	9	10

Peso **Reportado: PR**	120	120	135	118	120	190	124	175	133	125
Peso Directo: PD	125	118	139	120	125	198	128	176	131	125
Diferencia PR – PD = d	–5	+2	–4	–2	–5	+2	+4	+1	–2	0
$(PR - PD)^2 = d^2$	25	4	16	4	25	4	16	1	4	0

a. ¿Qué tipo de prueba debe usarse, unilateral o bilateral?
b. ¿Específicamente, cuál sería la prueba de hipótesis que se usaría?
c. Realice la prueba con □□= 0.05.

Solución.

a. Ya que se carece de información sobre la existencia de alguna tendencia, entre los individuos de la población, a reportar valores más altos o más bajos que su peso real, se debe usar una prueba de hipótesis bilateral.
b. Se debe utilizar una prueba para muestra apareada pues para cada uno de los 10 individuos tenemos la medida correspondiente a su peso corporal por dos métodos diferentes; entonces asumiendo que tanto el peso reportado como el tomado directamente, son variables con distribución normal se puede plantear:
   H_0: $\mu_d = 0$ Vs. H_1: $\mu_d \neq 0$
   Donde μ_d representa la media de las diferencias entre el peso recogido de una y otra forma en la población general de diferencias. Si no se rechaza H_0 podemos admitir que los métodos son comparables.
c. Vamos a calcular $t = \bar{d}\sqrt{n}/s_d$ que es el estadígrafo apropiado a la prueba propuesta en el inciso b.
   De la tabla 2 se tiene que:
   $\bar{d} = -25/10 = -2.5$
   $$s_d^2 = \frac{\sum d_i^2 - n\bar{d}^2}{n-1} = \frac{159 - 10*(-2.5)^2}{9} = 10.69$$ y de aquí que $s_d = 3.27$
   por consiguiente $t = (-2.5) \cdot 3.16/3.27 = -2.42$

En la tabla de la distribución t de Student puede comprobarse que la cifra que corresponde al percentil 0.975 con 9 grados de libertad es 2.262. Como $t < 0$ la región de rechazo será $t < -t_{9,0.975}$ y como efectivamente, –2.42 pertenece a la región crítica se rechaza la hipótesis nula y no se admite que los métodos son comparables.

11.11- EJERCICIOS PROPUESTOS.

Parte A: Estimación.

1. Supongamos que la estadía en días de un grupo de 25 pacientes en una sala de Geriatría, durante un período determinado, se comportó como sigue:

5,10,6,11,5,14,30,11,17,3,9,3,8,8,5,5,7,4,3,7,9,11,11,9,4

Asumiendo que la desviación estándar poblacional es de 6 días, compute un intervalo de confianza del 95 % para la estadía media de hospitalización en la sala de Geriatría.

2. Construya el intervalo de confianza del 95 % para la estadía media de hospitalización (ver ejercicio anterior) pero asumiendo que se desconoce □.

a. De también el intervalo para el 99 % de confianza.
b. Haga un comentario acerca de la relación entre ambos intervalos (Sugerencia: utilice el concepto de precisión).

3. Construya el intervalo de confianza del 95 % para la media del la creatinina sérica en los pacientes a que se refiere el ejercicio 5, evaluados a los 18 meses de iniciada las sesiones de diálisis.

4. Se realizó un estudio en un cierto hospital, con el objetivo de relacionar la reactividad a la tuberculina con la actividad laboral, encontrándose que 93 de las 221 enfermeras del hospital reaccionaron positivamente a la tuberculina. De el intervalo de confianza del 95 % para la proporción de positivos a la prueba entre las enfermeras.

5. Se les tomó la presión arterial sistólica a 200 personas con glaucoma y se encontró que la media era de 140 mmHg y el desvío estándar de 25 mmHg. Dé un intervalo de confianza del 95 % para la media de la presión arterial sistólica de personas con glaucoma.

6. Suponga que se lleva a cabo un ensayo clínico para probar la eficacia de una nueva droga, la espectinomicina, en el tratamiento de la gonorrea en mujeres. Se le aplica una dosis de 4 g diarios de la droga a 46 pacientes y al examinarlas una semana después, 6 de las mismas tienen aún gonorrea.

a) ¿Cuál es la estimación puntual para π, probabilidad de fracaso del tratamiento con la nueva droga?
b) ¿Cuál es el intervalo de confianza del 95 % para π?

7. Se está realizando un estudio en un grupo de cerdos de experimentación. Se inoculan con una dosis fija de una toxina que provoca aumento de tamaño del hígado y se encuentra que de los 40 cerdos de experimentación, 15 tienen el hígado aumentado de tamaño.

a ¿Cómo se calcula el estimador puntual de la probabilidad de que un cerdo tenga el hígado aumentado de tamaño?

b. ¿Cuál es el intervalo del 95 % de confianza para dicha probabilidad, asumiendo que se cumple la aproximación a la distribución normal?

8. Compute un intervalo de confianza del 99 % para la incidencia de muerte cardiovascular en una muestra de 750 hombres, seguidos durante 6 años, si al final de ese tiempo 64 murieron producto de algún accidente cardiovascular.

9. A un grupo de 50 hombres menores de 55 años, con una historia previa de infarto del miocardio, se les indicó, como parte de un programa experimental, una dieta a base de vegetales. Después de 5 años, 2 hombres del grupo murieron. ¿Cuál es el estimador puntual de la tasa de mortalidad en este grupo de hombres?

10. A continuación se presenta la información correspondiente a la media y desviación estándar de la talla (en cm) en niñas de 7 años, según si fueron o no bajo peso al nacer (peso hasta los 2500g).

Grupo	*Media*	*DE*	*n*
Bajo peso	120.0	5.11	148
Normo peso	121.6	5.56	1576

a. Proporcione intervalos de confianza del 95 % para cada uno de los grupos.

b. Basado en los resultados del inciso a, de su opinión acerca de la importancia de la diferencia entre la talla media de ambos grupos.

11. Al medir el tiempo de reacción, un psicólogo estima que la desviación estándar es de 0.05 seg. ¿De qué magnitud debe ser la muestra de mediciones a fin de que él confíe que:

a. el 95 % y

b. el 99 % del error de su estimación no exceda los 0.01seg?

12.Supongamos una muestra aleatoria de 100 trabajadores de una fábrica en los que el valor medio de la concentración de plomo en sangre es 90μg/l. Si la concentración tiene una distribución normal con una desviación estándar de 10, halle los límites de confianza del 95 % para la concentración media de plomo de los trabajadores de la fábrica.

Parte B: Pruebas de hipótesis.

12. En una muestra de 63 pacientes con lupus eritomatoso se encontró que el promedio de Hb sérica fue de 10.5 mg/100ml de suero y la suma de cuadrados de las desviaciones alrededor de la media fue de 182.39746. Considerando que según las normas clínicas los valores promedio de Hb sérica se distribuyen normal con media de 12.5 mg/100ml; ¿es posible concluir, con una significación del 5 %, que los niveles de Hb sérica en pacientes lúpicos son diferentes a los de la población normal?

1. Consideremos que a cada uno de 10 estudiantes, del género masculino, del primer año de Medicina se les tomó el ritmo cardíaco en latidos por minuto. Se encontró que tenían una media de 68.7 l/min. y una desviación estándar de 8.67 l/min. Se conoce que el ritmo cardíaco promedio en hombres jóvenes es de 72 l/min. ¿Cree usted que la información obtenida corrobora el conocimiento anterior? Utilice un nivel de significación de 0.05.

13. Las tasas de participación en un programa de despistaje de cáncer fueron, para una región durante numerosos años, del 20 % como promedio. Un nuevo programa de salud pública fue evaluado en una muestra de 200 habitantes de esta región y de ellos 80 habían participado del programa de detección de cáncer. ¿Es la tasa de participación hallada en la muestra significativamente diferente de la proporción referida a la población?

14. Suponga que se lleva a cabo un ensayo clínico en 164 mujeres, para probar la eficacia de una nueva droga en el tratamiento de la gonorrea. Con ese fin se forman dos grupos:

Grupo 1: integrado por 90 pacientes, recibe la nueva droga.

Grupo 2: integrado por el resto de las pacientes, recibe penicilina que es la droga de elección.

Una semana después se examina a todas las mujeres de nuevo y se determina que el 13 % del primer grupo padece aún de gonorrea, mientras que la tasa de fracaso del grupo 2 es del 10 %. ¿Qué se puede decir acerca de la eficacia de la nueva droga?

15. Supongamos que se quiere probar la hipótesis de que las madres con bajo nivel socioeconómico tienen niños, cuyo peso al nacer, es más bajo que el de la población general. Para verificar dicha hipótesis se obtiene la lista de los pesos al nacer de 100 niños a término nacidos en una maternidad enclavada en un área

de bajo nivel socioeconómico. Se encuentra que tienen un peso medio de 3267 gm. y una desviación estándar de 682 gm. Si se conoce que el promedio de peso de los recién nacidos en el país es 3409 gm. y la desviación estándar es de 710 gm, ¿puede afirmarse, con un nivel de significación del 1 %, que el peso medio del hospital es más bajo que el promedio nacional?

16. Construya las hipótesis adecuadas en el caso del ejercicio resuelto 6, en relación con el género. ¿Hay diferencia significativa en la estadía posoperatoria en este caso?

17. Diga si se rechaza o no la hipótesis nula en los casos que se presentan en los incisos a y b del ejercicio resuelto 9. Use □□= 0.05.

18. Se evaluó la dieta de 51 niños del género masculino, cuyas familias estaban por debajo del nivel de pobreza y se encontró que la media del consumo diario de hierro era de 12.5 mg, y su desviación estándar 4.75 mg. Suponga que la media del consumo diario de hierro en una población de niños de la misma edad, provenientes de todos los estratos sociales, es de 14.44 mg. Se desea probar si la media del consumo diario de hierro entre los niños de bajo nivel de ingresos monetarios es diferente a la de la población general. Plantee las hipótesis que pueden ser utilizadas y lleve a cabo su evaluación utilizando el estadígrafo correspondiente y □□= 0.05.

19. Un investigador desea estudiar si la presión arterial de la persona puede ser afectada por la posición de reposo que esta adopte en el momento en que se le tome. Con este propósito decide seleccionar 10 pacientes y tomarles la presión en dos posiciones diferentes: sentado y acostado. La información se presenta en la tabla 1.

Tabla 1. Efecto de la posición en los niveles de presión arterial sistólica.

Paciente	Sentado	Acostado	Paciente	Sentado	Acostado
1	142	154	6	100	100
2	100	106	7	108	120
3	112	110	8	94	90
4	92	100	9	104	104
5	104	112	10	98	114

Usando un nivel de significación del 5 % pruebe la hipótesis de que la posición afecta el nivel de la presión arterial.

Capitulo 12. Métodos estadísticos para el análisis de datos cualitativos: Análisis de tablas de contingencia.

12.1 Introducción.

El análisis tradicional de tablas de contingencia ha estado dado por: la prueba χ^2 de Independencia y la de Homogeneidad. Ambas tienen sus especificidades. Sin embargo, en la mayoría de los textos se estudian juntas, o incluso como una única prueba.

En este capítulo abordaremos, fundamentalmente las pruebas de hipótesis referidas y otras más específicas; que te introducirán en el estudio de un grupo de pruebas que han sido consignadas bajo el nombre de métodos no paramétricos[5].

12.2 Conceptos básicos.

Una tabla de contingencia (TC), es un arreglo compuesto de filas y columnas, donde se clasifica información relativa a una muestra, respecto de 1, 2 o más variables aleatorias, mediante el uso de escalas apropiadas. La forma más conocida de tratamiento estadístico de éstas son, la prueba de independencia y la de homogeneidad.

Si bien estas pruebas tienen muchos aspectos comunes, ellas difieren en cuanto a:

- Las particularidades de las TC que sirven de soporte a cada prueba.
- La hipótesis que se somete a prueba.
- Y, obviamente, la conclusión a que se arriban con cada una de ellas.

Con relación a las particularidades de las TC, se pueden hacer los siguientes señalamientos:

A. Se supone en primer termino que se selecciona una muestra aleatoria de tamaño n de una población dada. Posteriormente los elementos de esta muestra pueden ser clasificados de manera, simple, mediante el empleo de una escala asociada a una variable (conjunto de categorías o clases exhaustiva y excluyente) previamente seleccionada, o en forma múltiple (conjunta), si se utilizan 2 o mas escalas correspondientes a variables dadas, esta ultima alternativa origina las llamadas TC múltiples (bivariadas o dobles, trivariadas, etc) .

Para ilustrar lo expresado considérense las situaciones siguientes:

1. Se tiene una muestra aleatoria de 500 personas, estas pueden ser clasificadas respecto a:

- Peso y estatura, mediante escalas apropiadas previamente seleccionadas para ellas, este caso da origen a una TC bivariada que tendrá tantas casillas o celdas de clasificación, como indique el producto (multiplicación) de la cantidades de categorías de ambas escalas.
- Si tienen el colesterol normal o no y si se consideran respecto al peso, por medio de criterios previamente establecidos, como personas de bajo peso, con peso normal, o por el contrario con sobrepeso. Como resultado, aquí se obtiene una tabla bivariada que posee 6 celdas o casillas de clasificación.
- Si las mismas consideran que sus hijos (caso de tenerlos) hicieron una selección de la carrera universitaria que van a estudiar correcta o no, o por el contrario están indecisos,. En este caso, se obtiene una TC que posee solamente 3 celdas o casillas, una por cada categoría.

---

[5] Una prueba no paramétrica es aquella cuyo modelo no especifica las condiciones de los parámetros de la población de la que se extrajo la muestra; no hipotetiza sobre los parámetros poblacionales, sino sobre alguna característica de la misma o sobre la distribución poblacional.

2 Considérese que un grupo de investigadores desea **estudiar la enfermedad de Alzheimer en asociación con el sexo**, y para ello diseñaron un estudio analítico de prevalencia en una muestra aleatoria de 779 adultos de Ciudad de la Habana, cuyos resultados se exponen mediante la clasificación doble siguiente:

Sexo	Enfermedad de Alzheimer		Total
	Presente	Ausente	
Masculino	7	287	294
Femenino	33	452	485
Total	40	739	779

Aquí se tiene una TC, de las llamadas 2 por 2, ya que ambas variables de clasificación son del tipo de las dicotómicas, es decir, de la clase de variables que solo tienen dos valores posibles.

Tomando como referencia el ejemplo de esta última tabla, se hace notar que, lo único que en ella resulta conocido de antemano, es el valor **779 (el tamaño de la muestra observada)**, ya que todos los demás valores que se referencian en dicha tabla, es decir, **los valores de las celdas o casillas (7, 287, 33 y 452) y los totales marginales por columnas y filas (40, 739, 294 y 485): resultan desconocidos e impredecibles antes de llevar a cabo la clasificación**, debido al carácter aleatorio de lo observado en dicha tabla con respecto a la muestra.

Esta **particularidad de la TC**, señalada en lo ya expuesto, es lo que la hace apropiada para su empleo por medio de la **PH (prueba de hipótesis) para la independencia.**

B. Para exponer el otro modo de obtener tablas de contingencia, se utilizara la misma tabla anterior, que a continuación reproducimos:

Sexo	Enfermedad de Alzheimer		Total
	Presente	Ausente	
Masculino	7	287	294
Femenino	33	452	485
Total	40	739	779

Pero supóngase ahora que los investigadores no desean estudiar lo expuesto en el ejemplo anterior, sino que, estos quieren analizar ahora, **si la presencia o no de la enfermedad difiere respecto al sexo**. Para ello deciden obtener dos muestras aleatorias una de cada sexo, de tamaños 294 y 485 en hombres y mujeres respectivamente. Luego de tener las muestras de cada sexo, estas se clasifican respecto a, si las personas padecen o no de Alzheimer, obteniéndose la tabla anterior.

¿En que difiere, la tabla presente de la anterior?. Bueno, los números dentro de la tabla (en las celdas) y en los márgenes son los mismos, en esto se esta claro que no difieren. En realidad se diferencian, en que ahora la tabla actual, **tiene todo un margen conocido de antemano**, esto es; **los números 294, 485 y 779 no son resultado de la clasificación**, los restantes si, **los valores de las casillas y los totales del otro margen.**

El que una TC posea un margen predeterminado es una característica distintiva de que respecto a dicho margen, es decir, en relación con cada una de las categorías o clases de la escala ahí representada, se ha llevado a cabo un muestreo, efectuándose luego una clasificación de los elementos de estas muestras por

medio de las categorías (de la escala) que presenta la otra variable, esta es una particularidad de las TC, que hace posible el empleo de las mismas para llevar a cabo la prueba de homogeneidad.

Como se ha podido observar para generar una de las dos TC se emplea una sola muestra, notándose en este caso que ambos márgenes son desconocidos, mientras que para construir la otra se emplean tantas muestras aleatorias independientes como categorías tenga la variable del margen, cuyos totales van a quedar predeterminados por los tamaños de muestras correspondientes.

A continuación se presenta una estructura de tabla de contingencia que ilustra lo expresado, pero de modo general:

Variable A (f muestras independientes)	Variable B				Total
	B_1	B_2	B_3	....B_c	
A_1	n_{11}	n_{12}	N_{13}	....n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	N_{23}	....n_{2c}	$n_{2.}$
A_3	n_{31}	n_{32}	N_{33}	....n_{3c}	$n_{3.}$
. . A_f	n_{f1}	n_{f2}	n_{f3}	....n_{fc}	$n_{f.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	....$n_{.c}$	n

Nótese que en los márgenes de esta tabla aparecen los totales por filas y columnas pero escritos en la forma, $n_{.j}$ y $n_{i.}$, significándose con estas expresiones los totales correspondientes a la columna j-esima y la fila i-esima de la misma. A los efectos del calculo simbólicos estos valores responden a las formulas siguientes:

$$n_{\cdot j} = n_{1j} + n_{2j} + n_{3j} + \cdots + n_{fj}, \text{ para } j \text{ entre } 1 \text{ y } c.$$

$$n_{i\cdot} = n_{i1} + n_{i2} + n_{i3} + \cdots + n_{ic}, \text{ para } i \text{ entre } 1 \text{ y } f.$$

12.3 Prueba de independencia.

La prueba de independencia o asociación, se utiliza, como ya se dijo, cuando se tiene una muestra de n individuos que se clasifican respecto a dos variables, preferentemente cualitativa (nominales dicotómicas o politómicas)[6] y se desea conocer a partir de datos muestrales, si existe asociación de estas a nivel poblacional. El uso y las hipótesis de esta prueba se resume en el siguiente recuadro:

Hipótesis:

H_0: Existe (poblacionalmente) independencia entre las variables estudiadas

H_1: No existe (poblacionalmente) independencia.

Estadígrafo de la Prueba:

[6] Es válido aclarar que si bien puede utilizarse la prueba para cualquier tipo de variable, ella es preferiblemente para variable cualitativas nominales, en tanto existen pruebas de mayor especificidad para variables medidas en una escala ordinal o de mayor nivel.

La construcción del estadígrafo de la prueba reposa en el principio de la comparación entre las frecuencias observadas en cada celda de la tabla de contingencia a partir de un estudio concreto y las esperadas, es decir, las frecuencias que deberían ocurrir en cada celda, si no existiera asociación entre las variables en estudio; de la forma en que aparecen en la siguiente expresión (fórmula 1):

$$\chi^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \qquad (1)$$

Donde: n_{ij}: frecuencia absoluta observada en la celda i, j

e_{ij}: frecuencia absoluta esperada en la celda i, j

calculándose e_{ij}, por medio de, $e_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$, es decir, mediante el producto de los totales marginales de las celdas i, j divididos por el tamaño de la muestra.

El estadígrafo χ^2 de Pearson sigue una distribución χ^2 con $(f-1)\,(c-1)$ grados de libertad.

La regla de decisión es: rechazar H_0 si $\chi^2_{obs} \geq \chi^2_{1-\alpha,(f-1)\cdot(c-1)}$,

donde los valores que aparecen en esta desigualdad, son:

- El de la izquierda, es el valor calculado mediante la expresión (1) a partir de los valores **muestrales observados** en la TC (de ahí el subíndice, **obs)** y,
- El de la derecha, es el valor teórico hallado a partir de la tabla de la distribución chi-cuadrado, siempre que se prefije el nivel de significación α de la PH.

12.4 Prueba de Homogeneidad

La Prueba de Homogeneidad se utiliza cuando se estudian o analizan muestras independientes de f subpoblaciones ($f \geq 2$), determinadas por los distintos niveles (f) de una escala respecto de una variable dada y se clasifican las observaciones de éstas, respecto a c categorías ($c \geq 2$) de otra escala perteneciente a otra variable con la finalidad de conocer si la distribución de la variable estudiada difiere de las "*f*" poblaciones subyacentes de las cuales se obtuvieron las muestras.

El uso de esta prueba tiene asociada la siguiente estructura:

Hipótesis:

H_0: No difiere la distribución de las variables estudiadas en las "*f*" poblaciones subyacentes de las cuales se obtuvieron las muestras.

H_1: Difiere la distribución de las variables estudiadas en las "*f*" poblaciones subyacentes de las cuales se obtuvieron las muestras.

Estadígrafo de prueba:

El estadígrafo de la prueba de hipótesis y la regla de decisión son justamente las mismas que para la prueba de independencia, de ahí que no la presentemos de nuevo.

A continuación se resuelve un ejemplo práctico mediante una de las pruebas estudiadas.

Ejemplo:

– un grupo de investigadores desea probar la hipótesis siguiente: "La aparición de la enfermedad de Alzheimer está asociada al sexo"; para ello diseñan un estudio analítico de prevalencia en una muestra de 779 adultos mayores de Ciudad de La Habana durante 1999. los resultados se exponen en la tabla siguiente:

Tabla de valores observados.

Sexo	Enfermedad de Alzheimer		Total
	Presente	Ausente	
Masculino	7	287	294
Femenino	33	452	485
Total	40	739	779

Hipótesis:

H_0: Existe poblacionalmente independencia entre el sexo y la enfermedad de Alzheimer.

H_1: No existe poblacionalmente independencia entre el sexo y la enfermedad de Alzheimer.

Estadígrafo de prueba:

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Para obtener el EP se necesita calcular los e_{ij}. La siguiente tabla presenta estos valores.

Tabla de valores esperados.

Sexo	Enfermedad de Alzheimer	
	Presente	Ausente
Masculino	15.1	278.9
Femenino	24.9	460.0

Sustituyendo en fórmula 1:

$$\chi^2 = \frac{(7-15.1)^2}{15.1} + \frac{(287-278.9)^2}{278.9} + \frac{(33-24.9)^2}{24.9} + \frac{(452-460)^2}{460}$$

$$\chi^2 = 4.35 + 0.24 + 2.63 + 0.14 \approx 7.35$$

Regla de decisión:

$7.35 > 3.84$[3] $\Rightarrow$ rechazo H_0

Respuesta:[7]

Hay evidencia suficiente para plantear asociación entre el sexo y la enfermedad de Alzheimer al 95% de confiabilidad.

*Caso particular 2x2

[7] **Valor correspondiente al percentil 95 de la distribución** χ^2

Si bien el EP descrito con anterioridad es válido para el análisis tradicional de cualquier tabla de contingencia, cuando se trate de una TC de 2x2 éste se resume en la siguiente expresión (fórmula 2):

$$\chi^2 = \frac{n_{..}\left(n_{11}n_{22} - n_{12}n_{21}\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

la regla de decisión para este caso es: rechazo H_0 si $\chi^2_{obs} \geq \chi^2_{teórico}[1-\alpha\,;\,1gl]$

Resolvamos el ejemplo anterior mediante la expresión planteada:

$$\chi^2 = \frac{779\left(7x452 - 287x33\right)^2}{294x485x40x739} \approx 7.35$$

Como puede verse los resultados son idénticos a los obtenidos con la fórmula 1.

*Corrección por continuidad de Yates.

Desde que en 1934, Yates propuso introducir al estadígrafo χ^2 de Pearson una corrección, por el hecho de utilizar una distribución contínua (Jí Cuadrado) para representar la distribución muestral de variable(s) discreta(s), han surgido opiniones a favor y en contra de este proceder, las cuales han trascendido hasta nuestros días.
La objeción más frecuentemente esgrimida es que la corrección "sobrecorrige", dificultando el rechazo de H_0 de forma excesiva. Con relación a este planteamiento investigaciones empíricas han demostrado que:

- ✓ En el caso de tablas de contingencia *f*xc, la corrección por continuidad de Yates sobrecorrige, por lo que no debe utilizarse nunca.
- ✓ En el caso de TC de 2x2, con independencia del método de muestreo que generó la tabla, la corrección actúa como si los cuatro márgenes estuvieran fijos, en cuyo caso las probabilidades exactas asociadas con las frecuencias observadas en cada celda puede, bajo la restricción anterior, derivarse de la distribución hipergeométrica, de ahí que esta sea recomendada siempre en este caso. Se ha comprobado que la probabilidad asociada a este EP es casi idéntica a la probabilidad exacta de Fisher (Prueba de Fisher–Irwin)

El estadígrafo corregido para el caso 2x2 es (Fómula 3):

$$\chi^2 = \frac{n_{..}\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - \frac{1}{2}n_{..}\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

El estadígrafo corregido para el caso *f*xc es (Fórmula 4[4]):

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{\left(\left|n_{ij} - e_{ij}\right| - 0.5\right)^2}{e_{ij}}$$ [8]

Sólo se realizará la sustracción planteada si la diferencia modular es mayor que 0.5.
Así, si utilizamos el estadígrafo corregido en el ejemplo expuesto con anterioridad:

$$\chi^2 = \frac{779\left(\left|7x452 - 287x33\right| - \frac{779}{2}\right)^2}{294x485x40x739} \approx 6.47$$

Aunque se mantiene la decisión de rechazo, note como el EP se redujo de 7.35 a 6.47.

* Errores más frecuentes en el uso de las Pruebas de Independencia y Homogeneidad.

Las dócimas de independencia u homogeneidad se usan con bastante frecuencia en la investigación en salud. Sin embargo, en muchas ocasiones éstas son inadecuadas. Lo cual conduce a errores en la toma de decisión que deviene de estos resultados.
Con relación a la utilización errónea de estas pruebas varios autores han brindado sus criterios. Así, López Pardo señala lo siguiente:

1. Uso irreflexivo de la prueba: antes de realizar cualquier prueba deben inspeccionarse los datos. Este procedimiento, puede en ocasiones, aportar la información necesaria. De hecho en muchas ocasiones, un análisis descriptivo de los datos puede evidenciar diferencias, o por el contrario, evidenciar que no hay diferencias, sin necesidad de acudir a las pruebas de significación.
2. Utilización de la prueba sin tener en cuenta el tipo de variable: en este aspecto, suelen cometerse errores en dos sentidos.
   - ✓ Clasificar una variable cuantitativa en su naturaleza con una escala de menor nivel: en estadística, salvo excepciones, debemos evitar perder información. De hecho, si clasificamos las variables con escalas de un nivel inferior estamos perdiendo información. Así, si se desea ver si una variable continua, digamos la edad, está asociada a una variable cualitativa nominal dicotómica, la prueba de comparación de medias, o su homóloga no paramétrica (t de Wilcoxon o U de Mann–Whitney), resultan más convenientes, y si ésta es politómica, el ANOVA de 1 vía o Kruskal Wallis son más adecuadas.
   - ✓ Utilizar la prueba cuando una de las variables es cuantitativa ordinal. En este caso, si la otra es dicotómica, la Prueba de Bartholomew es más específica, o el Ridit Analysis, cuando esta es politómica.
3. Uso del estadístico χ^2 como una medida de asociación: la prueba de independencia, es una prueba de significación de la asociación. Es decir, sólo nos dice si la asociación entre las variables investigadas puede atribuirse o no al azar. No nos permite conocer la magnitud de la asociación. Este último aspecto puede conocerse mediante el uso de la correlación entre variables, en el caso de Tablas

[8] El efecto sobrecorrector de la corrección por continuidad de Yates en este estadígrafo ha propiciado que el mismo se encuentre en desuso.

de Contingencias podría ser conveniente el cálculo de coeficientes de correlación como ϕ, Contingencia, V de Cramer, etc.

4. Utilización de la prueba cuando existen valores esperados muy pequeños.

Se ha demostrado que en Tablas de Contingencia (fxc), cuando más del 20% de las frecuencias esperadas es inferior a 5, o alguna de éstas es infererior a 1. No deberá realizarse la prueba. En esta caso, deberá colapsarse convenientemente la tabla, convertirla en (2x2) y aplicar la Prueba de Probabilidades Exactas de Fisher.

En el caso de Tablas de Contingencia (2x2), si la menor frecuencia esperada es inferior a 5 no deberá realizarse la prueba, al menos el estadístico χ^2 sin corrección. En este caso se empleará el estadístico χ^2 corregido, o en su defecto la Prueba de Probabilidades Exactas de Fisher.

5. Uso de la prueba cuando se dispone de valores promedios o porcentajes.

Las pruebas de independencia u homogeneidad se realizan a partir de las frecuencias observadas, no de medidas de resúmenes.

*Detección de fuentes de significación en tablas de contingencia: el análisis de los residuos.

Cuando se rechaza la hipótesis de nulidad en el ATC, no se obtiene información sobre las partes específicas de la TC causantes de la dependencia o la heterogeneidad. Sin embargo, en ocasiones se quieren hacer comparaciones entre celdas o de líneas de una misma tabla. Se han desarrollado varios métodos con este propósito, como la llamada χ^2 a posteriori que no estudiaremos en este texto por formar parte de un volumen posterior, y el análisis de los residuos para el caso de fxc, que estudiaremos a continuación.

El análisis de los residuos es una técnica desarrolla por Haberman (1973), mediante la cual se pueden identificar las celdas que hacen un aporte significativo al estadístico de prueba en una TC de fxc. Este análisis se realiza después de conocer que existe dependencia o heterogeneidad, por lo tanto tiene como supuesto la realización de una prueba a priori y el rechazo de la hipótesis de nulidad de independencia u homogeneidad.

Este análisis incluye: la obtención de los residuos estandarizados, la obtención de los residuos ajustados y la prueba de hipótesis respectiva.

– Obtención de los residuos estandarizados.

Un residuo estandarizado o tipificado (E_{ij}) es la diferencia entre los valores observados (n_{ij}) los esperados dividida por la raíz cuadrada del valor esperado (e_{ij}), lo cual se denota por la siguiente expresión:

$$E_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

– Obtención de los residuos ajustados.

Un residuo ajustado se obtiene mediante la división del residuo estandarizado por la raíz de su varianza (V_{ij}), lo cual se denota por la siguiente expresión:

$$d_{ij} = \frac{E_{ij}}{\sqrt{V_{ij}}}$$

donde :

$$V_{ij} = \left[1 - \frac{n_{i.}}{n_{..}}\right] \quad \left[1 - \frac{n_{.j}}{n_{..}}\right]$$

– Prueba de Hipótesis

H_0: $\mu_{d_{ij}} = 0$

H_1: $\mu_{d_{ij}} \neq 0$

Donde $\mu_{d_{ij}}$ es el valor medio de todos los d_{ij} a nivel poblacional.

EP

$d_{ij} \sim N(0,1)$

Regla de decisión: Rechazo H_0 si: $\left|d_{ij}\right| \geq Z_{1-\alpha/2}$

–Observaciones:

1. En caso de rechazo de H_0, se dice que el residuo ajustado es significativo, o mejor aún que su celda es fuente de significación en la tabla de contingencia analizada (su celda está fuertemente asociada o está contribuyendo de manera importante a la dependencia. O de otra forma, si se mide homogeneidad, digo que esta celda contribuye al comportamiento heterogéneo de las variables en estudio).
2. El signo del residuo indica cuál es mayor de los dos valores (el observado o el esperado).

A continuación ilustramos con un ejemplo:

La tabla siguiente muestra los resultados de un estudio hecho para comparar los efectos PAS y la estreptomicina en el tratamiento de la TB pulmonar.

Esputo	***Tratamiento***		***Estreptomicina y PAS***	***Total***
	PAS	***Estreptom.***		
Frotis +	56	46	37	139
Frotis – ***Cultivo +***	30	18	18	66
Frotis – ***Cultivo –***	13	20	35	68
Total	99	84	90	273

a) Determine si existe asociación significativa entre los resultados del esputo y el tratamiento impuesto.
b) En caso de rechazar H_0 (existe dependencia), determine cuáles casillas hacen un aporte importante al estadígrafo de prueba.

Respuesta/ (utilizamos el nivel de significación 5%, o sea, α=0.05)

$e_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}}$, entonces

$e_{11} = \frac{(139)(99)}{273} = 50.4$ $\quad e_{21} = \frac{(66)(99)}{273} = 23.9$ $\quad e_{31} = \frac{(68)(99)}{273} = 24.7$

$e_{12} = \frac{(139)(84)}{273} = 42.8$ $\quad e_{22} = \frac{(66)(84)}{273} = 20.3$ $\quad e_{32} = \frac{(68)(84)}{273} = 20.9$

$e_{13} = \frac{(139)(90)}{273} = 45.8$ $\quad e_{23} = \frac{(66)(90)}{273} = 21.8$ $\quad e_{33} = \frac{(68)(90)}{273} = 22.4$

H_0: no existe asociación significativa entre los resultados del esputo y el tratamiento impuesto.

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \qquad \chi^2_{\text{teórico}}=9.49$$

sustituyendo tenemos que:

$$\chi^2 = 17.7$$

Como $\chi^2_{obs} > \chi^2_{teórico}$ y además la probabilidad asociada al valor, $\chi^2 = 17.7$ está entre 0.01 y 0.00, por tanto es menor que α=0.05.

Entonces, por cualquiera de las dos vías rechazo la hipótesis nula y podemos decir que existe asociación significativa entre los resultados del esputo y el tratamiento impuesto.
Ahora, como ya sabemos que existe asociación, corresponde realizar una prueba a posteriori para conocer cuáles celdas son las fuentes de la dependencia.

Aplicando las fórmulas que ya conoces tenemos los siguientes resultados:

Residuo tipificado	*Varianza*	*Residuo Ajustado*
E_{11}= 0.78	*V_{11}=0.32*	*d_{11}=1.39*
E_{12}= 0.49	*V_{12}=0.35*	*d_{12}=0.83*
E_{13}= -1.29	*V_{13}=0.34*	*d_{13}=-2.22**
E_{21}= 1.24	*V_{21}=0.49*	*d_{21}=1.77*

E_{22}= 0.50	V_{22}=0.53	d_{22}=0.69
E_{23}= -0.8	V_{23}=0.52	d_{23}=-1.1
E_{31}= 2.46	V_{31}=0.48	d_{31}=3.56*
E_{32}= -0.2	V_{32}=0.52	d_{32}=-0.27
E_{33}=2.68	V_{33}=0.5	d_{33}=3.77*

** Estas son las celdas que contribuyen a la significación del estadígrafo.*

Ejercicios propuestos:

1. Se plantea la hipótesis que las embarazadas atendidas por médicos no especializados tienen mayor riesgo de presentar complicaciones durante el parto. Se realiza un estudio en 75 embarazadas, cuyos resultados se exponen a continuación.

Calificación del Médico	***Presencia de complicaciones en el parto***		***Total***
	Sí	***No***	
Especialista	7	23	30
No Especialista	18	27	45
Total	25	50	75

a) ¿Puede plantearse que en la población de la cual se extrajeron estas muestras, la calificación del médico se encuentra asociada a la presencia de complicaciones durante el parto?.

2. Un grupo de investigadores desea conocer si el grupo sanguíneo del receptor está asociado al rechazo del Trasplante Renal, para ello desarrollan una investigación, la cual arrojó los siguientes resultados.

Grupo Sanguíneo	***Rechazo del T. Renal***		***Total***
	Sí	***No***	
A	5	25	30
B	2	20	22
AB	30	5	35
O	1	30	31
Total	38	80	118

a) ¿Existe asociación entre las variables investigadas?
b) En caso de existir, encuentre las celdas que contribuyen a esta.

Bibliografía

1. Armitage P, Berry G. **Statistical Methods in Medical Research**. 3rd ed. Oxford: Blackwell Scientific Publications, 1994
2. Altman DG. **Practical statistics for medical research.** London: Chapman and Hall. 1992
3. Daniel WW. **Bioestadística**. Base para el análisis de las ciencias de la salud. 3ra edición. México. D. F: Limusa; 1997
4. Norman GR, Streiner DL. **Bioestadística**. España: Hartcourt Brace; 1998
5. Martínez Canalejo H, Santana Porben S. **Manual de procedimientos bioestadísticos**. Tomo I. ISCM de la Habana. 1989.
9. Armitage P, Berry G. **Estadística para la Investigación Biomédica**. Doyma, Barcelona, 1992.
10. Versión electrónica del manual de la Universidad de Málaga. **Bioestadística: métodos y aplicaciones**. Málaga; 1998.
11. Dawson-Saunders B, Trapp RG. **Bioestadística.** 2da Edición. El Manual Moderno. México; 1999.
12. Fliess J L. **Statistical Methods for Rates and Proportions**. Second Edition. New York. John Wiley and Sons; 1981
13. López Pardo C. **Curso "Métodos No Paramétricos en la Investigación Contemporánea"**. La Habana. Universidad de La Habana; 1990

*Prueba de probabilidades exactas de Fisher.

MINISTERIO DE SALUD PUBLICA.

FACULTAD DE SALUD PUBLICA
DIRECCION NACIONAL DE ESTADISTICAS
DPTO. DE BIOESTADÍSTICA

TEMAS DE ESTADÍSTICAS SANITARIAS

AUTORES:

DRA. MIRIAM GRAN ALVAREZ Msc.
DIRECCIÓN NACIONAL DE ESTADÍSTICAS. MINSAP

DRA. ILEANA CASTAÑEDA ABASCAL Msc.
FACULTAD DE SALUD PUBLICA. MINSAP

ESPECIALISTAS DE SEGUNDO GRADO EN BIOESTADISTICA.
PROFESORAS AUXILIARES. FACULTAD DE SALUD PUBLICA

COLABORARON EN LA ELABORACIÓN Y REVISIÓN:

DRA. MARÍA NIETO LLUIS. Msc.
DRA. MARÍA ESTHER ALVAREZ LAUZARIQUE. Msc.
ESPECIALISTAS DE SEGUNDO GRADO EN BIOESTADÍSTICA.
DIRECCIÓN NACIONAL DE ESTADÍSTICAS. MINSAP.

1- Estadísticas de Salud

1.1- Conceptos de Estadísticas de Salud.

Es la información numérica, imprescindible y cuantificable para conocer el Estado de Salud de la Población así como para planificar, organizar, evaluar y controlar programas y acciones de salud.

La información estadística de salud debe reflejar lo más fiel posible la realidad objetiva que mide.

Las estadísticas de salud tienen cada vez mayor alcance y complejidad en la medida que se amplia y desarrolla el que hacer en salud.

1.2- Importancia y uso de las Estadísticas de Salud.

Las estadísticas de salud tienen uso individual y estadístico. El uso individual se refiere al uso de los registros médicos de cada persona que accede a los servicios de salud donde quedan registrados ciertas características del individuo y la historia de su enfermedad, muerte, tratamientos u otros servicios recibidos. Los registros médicos deben poseer los atributos de confidencialidad y custodia lo cual se regula por leyes y reglamentaciones con amparo legal. El uso estadístico se refiere al manejo de datos agregados donde se suman los datos relativos a cada individuo en modelos que compilan la información individual o "de caso a caso" con las periodicidades establecidas para los diferentes niveles del sistema nacional de salud.

Las estadísticas de salud son imprescindibles para movilizar recursos humanos y materiales, interviene de manera determinante en el proceso de toma de decisiones en los diferentes niveles de mando, estratégico y operativo. Las estadísticas de salud se utilizan en todas las etapas de la administración o la gerencia del sector de la salud así como son de utilidad para otros sectores que se ocupan de las demás actividades socio económicas del país. Las estadísticas de salud, fundamentalmente las de mortalidad y natalidad son utilizadas sistemáticamente en los estudios demográficos y cálculos de población.

De tal manera, no resulta difícil entender el valor que en nuestros tiempos tiene el uso de las estadísticas de salud para desarrollar la gerencia sobre una base científica. Se utilizan en todas las etapas de la misma con especial interés en el proceso de toma de decisiones.

Las estadísticas sanitarias se utilizan además en:

- Investigaciones.
- Docencia de pregrado y posgrado.
- Gerencia, docencia e investigaciones relacionadas con otras ramas de la actividad socio económica del país.

1.3- Clasificación de las Estadísticas de Salud.

Estadísticas de Población. Información numérica acerca de la composición y principales características de las agrupaciones humanas. Son objeto de estudio de la Demografía.

Estadísticas Vitales. Información numérica cuantificable sobre hechos vitales.

Hecho Vital: Todo hecho relacionado con el comienzo y fin de la vida del individuo y con los cambios de su estado civil que puedan ocurrirle durante su existencia. Ej. : Nacimiento, Defunción, Defunción Fetal, Matrimonio, Divorcio, Adopciones, Legitimaciones, Reconocimientos y otros.

Estadísticas de Morbilidad. Información numérica sobre las enfermedades, principales padecimientos, discapacidad y secuelas de enfermedades o de hechos accidentales o intencionales (causas violentas de enfermedad) que se presentan en la población.

Estadísticas de Recursos. Información numérica sobre los recursos materiales y humanos disponibles y su aprovechamiento, que posee el Sistema Nacional de Salud para su gestión.

Estadísticas de Servicios. Información numérica acerca del volumen y calidad de los servicios de salud que se prestan.

Estadísticas de Vivienda. Información cuantificable relacionada con las viviendas en que el hombre habita y sus características.

Estadísticas de Saneamiento. Datos numéricos sobre las condiciones ambientales y sanitarias del entorno donde viven, trabajan, estudian o realizan otras actividades las comunidades humanas.

Estadísticas Económicas. Información numérica necesaria para el análisis de la actividad económica de salud. Ej. : Costos, Gastos, Inversiones, Exportaciones y otras.

2- Sistemas de Información Estadística.

2.1- Sistemas de Información.

La Teoría de Sistemas es aplicable al sector de la Salud ya que él mismo es un Sistema, por cierto, complejo, que interactúa con otros sistemas tan complejos como este.

El sistema de información de Salud debe estudiarse en su conjunto como una reunión de sub-sistemas en interacción interna y externa, de acuerdo a la estructura en la cual está conformado. Cada aspecto de la salud es un sub-sistema dentro de otro de mayor complejidad.

Uno de los sistemas de servicios del sector salud, por ejemplo, es el área de la Salud Materno-Infantil. Observamos que siempre habrá otros sistemas que interactúan de forma dinámica con este tanto internamente (Área de Asistencia Médica, Área de Medicamentos, Área de Epidemiología, otros), como externamente (Higiene Ambiental, Educación, y otros.) Así el área de Salud Materno-Infantil tiene a su vez diferentes sub-sistemas que la componen como vertientes fundamentales de trabajo: Atención gineco-obstétrica, atención durante el trabajo de parto y al parto, cuidados especiales perinatales, estado nutricional

de la gestante y el niño (a), atención ginecológica infanto juvenil, planificación familiar, atención durante el puerperio, cáncer ginecológico y otros. Cada una de estas vertientes tendrá uno o varios programas con un fin establecido a corto, mediano o largo plazo y sus objetivos definidos a alcanzar.

Para evaluar la marcha de cada programa es necesario todo un ***sistema de información*** que permita retroalimentar el programa y realizar los ajustes necesarios.

Para comprender qué es un sistema de información, tenemos que relacionarlo con el proceso de comunicación, entendiendo este como un proceso mediante el cual se trasmiten "mensajes" para generar información con vista a:

- Establecer la política nacional de salud.
- Trazar las estrategias y planes para ejecutar dicha política.
- Planificar las acciones de salud a ejecutar.
- Seguir o monitorear de cerca el desarrollo de los planes y programas.
- Controlar las actividades, tomando las decisiones que se requieran para adecuar dichos planes y programas.
- Proceso de toma de decisiones.

De esta forma, conoceremos los aspectos en que podemos influir o afectar intencionalmente la composición y operación de un sistema con el objetivo de ajustarlo, modificarlo o cambiarlo para optimizar el funcionamiento del mismo.

En la actualidad se diferencian nítidamente los conceptos de estadística e información. La estadística da prioridad a lo relacionado con la recolección y procesamiento de los datos. La información realza el uso de ellos.

Datos: Según el diccionario de la lengua española, es un antecedente para llegar al conocimiento exacto de un hecho. Es una magnitud o caracterización de algo. Son **estáticos**. No cambian una vez obtenidos. Cuando se les procesa y presenta en un contexto apropiado pueden generar entendimiento.

Información: Según el diccionario de la lengua española, es la reseña, representación o concepción derivada de la observación, lectura o instrucción. Es conocimiento en relación con un hecho, que surge de la confrontación de datos con los conocimientos que existen sobre el mismo. La información es **dinámica.**

El dato es un hecho aislado, un producto bruto que constituye la materia prima de la información. Para que se convierta en información ha de ser organizado, analizado y comunicado o emitido adecuadamente a fin de que pueda ser comparado y evaluado de acuerdo al conocimiento previamente adquirido de los hechos que estos representan.

Un sistema de información es un proceso que conlleva una actividad continua y adaptable a las circunstancias y necesidades, tratar de llevarlo a un esquema fijo, restringe una de sus propiedades principales: la de la adaptación rápida a situaciones de cambio.

Una definición aceptada de sistema de información es:

Sistema que se conforma con el conocimiento adquirido sobre un hecho y con elementos de análisis, evaluación, comparación y control, que permiten la toma de decisiones sobre un conjunto de actividades del sistema de servicio donde este opera y conduce al logro de la finalidad y objetivos del mismo.

Un sistema de información debe brindar información veraz, oportuna, relevante, exacta, útil y periódica. Para su diseño se utiliza la metodología y técnicas del análisis de sistema, para permitir el organizar la captación, emisión, procesamiento de los datos y tratamiento de la información, con vistas a lograr un sistema de información que permita evaluar el cumplimiento de los objetivos del sistema de servicio de que se trate, ya que es necesario prever qué indicadores se necesitan, como obtenerlos y por que vías y cómo habrán de llegar.

2.2- Sistemas de Información Estadísticos. (SIE)

Todo sistema de información ha de tener un sub-sistema que se encarga de la recolección, flujo y procesamiento de datos que permita generar información oportuna, confiable y necesaria. Este sub-sistema es llamado Sistema Estadístico.

SIE: Es el sub-sistema del sistema de información que se encarga de la recolección, procesamiento, flujo y presentación de datos a los diferentes niveles donde este opera.

Para que el Sistema Estadístico se convierta en Sistema de Información Estadístico[, debe además contener elementos de análisis, evaluación, comparación y control que permitan la toma de decisiones sobre el conjunto de actividades del programa.

Como todo sistema ha de tener los componentes esenciales de este: entradas, procesador o procesos, salidas, retroalimentación, estar bien delimitado y contar con definiciones claras y precisas. En todo sistema de información estadístico debe estar presente la estrecha relación entre productores o proveedores y usuarios a fin de mantener la coherencia entre las necesidades de información a satisfacer y su satisfacción real.

CLASIFICACIÓN DE LOS SIE

DE ACUERDO AL NIVEL JERÁRQUICO:

- GLOBAL
- RAMAL

DE ACUERDO AL ALCANCE:

- NACIONAL
- TERRITORIAL (PROVINCIA / LOCALIDAD)

DE ACUERDO A LOS MEDIOS TÉCNICOS EN QUE SE SUSTENTE:

- MANUAL
- AUTOMATIZADO
- MIXTO

DE ACUERDO A SU FINALIDAD O PROPÓSITO:

- RECURSOS
- SERVICIOS
- MORTALIDAD
- MORBILIDAD
- OTROS

DE ACUERDO A SU CONTENIDO:

- GENERALES
- ESPECIALES

Concluyendo: La información pasa por tres fases fundamentales del sistema, en el caso de la información de gestión. En su nacimiento, es un dato primario, algo que aún no está elaborado, que tiene un valor potencial, el cual debe convertirse en real mediante un trabajo posterior. El dato primario, junto con otros, recibe "tratamiento" por medio de operaciones tales como, clasificación, tabulación, cálculo, consolidación y otras operaciones. El producto terminado, constituye la última fase de la información, en esta, la información es ya un resultado listo para ser utilizado por los usuarios.

Los organismos internacionales especializados en la actividad de los SIE recomiendan una serie de principios para su buen funcionamiento. Estos son:

Principios Generales que sustentan un Sistema de Información Estadístico (SIE).
Recomendaciones.

- Definición del Sistema de Información Estadístico. Comprende la captación de los datos, validación y procesamiento, transmisión, preparación de informes, difusión y comunicación estadística.

- Función del SIE. Desarrollar las estadísticas en diferentes áreas de medición. Disponer de fuentes de información. (Banco de Datos / Bases de Datos amigables, interrelacionadas, que permitan la fácil carga y recuperación de la información)

- Carácter confidencial de los registros primarios. Legislación y regulaciones que protejan la información de carácter personal.

- Organismo (s) que se encargarán del sistema. Asignar responsabilidad y funciones a los organismos que se ocuparán del SIE.

- Coordinación entre los organismos que participan y son responsables del funcionamiento del SIE.

- Evaluación crítica periódica del SIE. Designar autoridad encargada de la evaluación crítica del SIE. Realizar estudios a fondo para la evaluación de todas y cada una de las etapas del SIE. (captación del dato, procesamiento, validación, salidas, comunicación, difusión, uso y satisfacción de necesidades de información)

- Uso del muestreo.

- Implementación de encuestas periódicas o eventuales. La realización de encuestas como un complemento para profundizar en ciertas variables y confrontar salidas de información del SIE.

- Interrelación con otros Sistemas.

Principios para el registro de los datos.
Recomendaciones.

- Definir y establecer los datos a registrar. Adopción o elaboración de definiciones.

- Establecer los procedimientos y regulaciones de registro. Incumben al declarante y al registrador.

- Obligatoriedad para la captación del dato. (Medidas educativas / Coercitivas.) Respaldo legislativo.

- Organización para la captación del dato. Estructura jerárquica, oficinas estadísticas geográficamente distribuidas. Tratar de captar el dato lo más cerca posible al lugar de acaecimiento

del evento a medir. Fácil accesibilidad para el declarante, horas adecuadas, cobertura satisfactoria para la captación del dato. Todo ello facilitará cumplir con los plazos de tiempo establecidos para el proceso estadístico, fundamentalmente con la captación de la información.

- Tecnología para la captación. En la medida que la tecnología sea más moderna y eficiente el proceso de la captación del dato será cuantitativamente y cualitativamente superior. Uso de la Informática.

- Gratuidad. La declaración del dato debe ser un acto gratuito.

- Plazos para la captación del dato. Su duración debe ser tal que garantice la oportunidad y exactitud del dato captado. Tener en cuenta procedimientos que permitan, en aras de la integridad, aceptar captaciones tardías.

- Confección de registros primarios y modelos sencillos con variables sencillas, fiables, factibles de obtener, de fácil comprensión para el declarante y el registrador.

Principios para la transmisión, consolidación y emisión de la información estadística.

Recomendaciones.

- Preparar informes y tabulaciones

- Trasmitir toda la información recogida en todas y cada una de las áreas de captación, sin omisión alguna.

- Uso de la Telemática en el proceso de transmisión de los datos, desde todos los niveles posibles.

- Flujo de información por niveles (Piramidal) Transmisión de información más o menos detallada según niveles del SIE. Más detallada en los niveles inferiores u operativos. Más general en la medida que asciende por los niveles del SIE.

- Regulaciones en forma, contenido y fecha de transmisión de la información así como de la recepción de los mismos.

- Establecimiento de métodos de evaluación permanentes de la veracidad de las tabulaciones.

- Emisión de información estadística organizada y/o analizada. Publicaciones periódicas o eventuales para los usuarios en general. A solicitud de usuarios eventuales e individuales. La emisión

de la información estadística varía en forma, contenido y periodicidad según el usuario (s) a que va dirigida. El tipo de usuario está dado por las necesidades de información a satisfacer en cada caso y momento. La información estadística se ofrecerá en soporte papel y/o electrónica.

3- Actividad Estadística en Cuba y Estructura Estadística de Salud.

La Oficina Nacional de Estadística (ONE) es el órgano central del estado encargado de dirigir, ejecutar y controlar la política nacional en relación con las estadísticas oficiales del país. Los diferentes ministerios y niveles nacionales del estado, cuentan con sus direcciones de estadísticas propias que satisfacen las demandas de información numérica de acuerdo a los requerimientos de cada uno y a su vez se rigen metodológicamente por la ONE.

Los SIE en Cuba son de tres tipos en dependencia de la institución a que se jerarquizan:

- SIEN: Sistema de Información Estadístico Nacional. Jerarquizado por la ONE independientemente de donde se recolecte, tabule y valide la información. Ej. : SIE de Nacimientos. Se recolecta y valida la información estadística en las unidades de salud y se tabula y emite por la ONE y sus dependencias. Recolectan y emiten información general y de amplio uso por los diferentes sectores de la actividad política y socio económica del país.

- SIEC: Sistemas de Información Estadísticos Complementarios. Uno para cada Organismo de la Administración del Estado Cubano. Ej. : SIEC de Salud, de Educación. Recolectan y emiten información más específica de cada sector. Son aprobados por la ONE.

- SIEL: Sistemas de Información Estadísticos Locales. Se diseñan para territorios o regiones de acuerdo a necesidades locales. Requieren aprobación de los niveles superiores.

El Sistema Nacional de Salud de acuerdo con la división política administrativa vigente del país posee cuatro niveles que desarrollan determinadas funciones en general y en particular en materia de estadísticas sanitarias de la manera siguiente:

Nivel	**Funciones**	**Dependencia Directiva.**
Nación	***Normativa Metodológica***	Dirección Nacional de Estadísticas y Registros Médicos. MINSAP
Provincia	**Dirección y Coordinación**	Departamentos Provinciales de Estadísticas y Registros Médicos.
Municipio	**Dirección y Coordinación**	Departamentos Municipales de Estadística y Registros Médicos.

Unidades de Salud	**Ejecutiva**	Departamentos de Estadística y Registros Médicos. Una en cada unidad de salud. Por cada 15 médicos de la familia hay un estadístico en el área de salud correspondiente.

2.4 Diseño de los SIE.

Los sistemas de información estadísticos se diseñan por medio de un minucioso trabajo científico multidisciplinario donde intervienen proveedores y usuarios.

El trabajo de diseño se realiza por etapas metodológicas que se suceden una tras otra por parte de un equipo de especialistas en estadísticas y del programa de salud que se trate, de tal manera que proveedores y usuarios de la información deben trabajar en estrecha relación a fin de garantizar el diseño más racional posible, la exitosa implantación del sistema, su actualización y perdurabilidad.

ETAPAS METODOLÓGICAS DEL DISEÑO DE UN SIE.

*** INVESTIGACIÓN PRELIMINAR: PRIMER ACERCAMIENTO AL PROBLEMA, DEFINICIONES. CONSTRUCCIÓN DEL MARCO TEÓRICO.**

*** ANTEPROYECTO: CREAR CONDICIONES DE INICIO DEL TRABAJO. ANÁLISIS GENERAL DEL SISTEMA. DISEÑO PRELIMINAR.**

*** *PROYECTO TÉCNICO*: *PRECISIÓN DE OBJETIVOS. SUBSISTEMAS MENORES. CONFECCIÓN DE PROGRAMAS AUTOMATIZADOS. CONFECCIÓN DE CODIFICADORES. PLAN DE IMPLANTACIÓN.***

*** *PROYECTO DE TRABAJO*: *PRECISIÓN Y ESPECIFICACIONES DE TAREAS ANTERIORES. MANUALES DE PROCEDIMIENTOS. CICLO DE PRUEBAS. AJUSTES. ELABORAR PLAN DE IMPLANTACIÓN.***

*** IMPLANTACIÓN: ASEGURAR CONDICIONES DE IMPLANTACIÓN. PLAN OPERATIVO DE IMPLANTACIÓN. ACEPTACIÓN DEL SISTEMA.**

*** MANTENIMIENTO Y DESARROLLO: SUPERVISIONES PERIÓDICAS. EVALUACIONES DE INTEGRIDAD Y CALIDAD. CAMBIOS NECESARIOS AL SISTEMA DESPUÉS DE UN TIEMPO DE EXPLOTACIÓN DETERMINADO**

ES IMPORTANTE MANTENER LA ESTRECHA RELACIÓN ENTRE PROVEEDORES Y USUARIOS YA QUE LOS SIE SURGEN A SOLICITUD DE LOS USUARIOS Y POR LA ACEPTACIÓN DE LOS PROVEEDORES DE LAS NECESIDADES DE INFORMACIÓN EXPRESADAS Y APROBADAS POR LOS NIVELES DE AUTORIDAD PERTINENTE.

2.5- Calidad de la información estadística.

La calidad de la información estadística expresa lo tanto que difiere su exactitud de lo conocido como error. La medición de la calidad de la información estadística se basa en dos conceptos básicos para el funcionamiento de un SIE:

- Cobertura: Posibilidad de recoger el dato establecido dondequiera que este se origine o lo que es igual, lo más próximo en tiempo y espacio a donde ocurra el evento que origina el dato primario a recolectar.

- Integridad: Proporción de datos registrados con relación al total de datos a registrar. Se refiere a que se registre, fluya y emita la totalidad de los eventos o datos que de ciertos hechos se originen y que así lo establezca el SIE en su concepción.

Sin una cobertura de estructuras para la recolección del dato y sin integridad de la recolección, obviamente la calidad de la información es baja.

La calidad de la información estadística en salud se mide a partir de la presencia de estos dos conceptos básicos y de la evaluación de ciertos atributos que en su conjunto expresan el nivel de la calidad de la información que se ofrece como resultado de un SIE.

ATRIBUTOS PARA EVALUAR LA CALIDAD DE LA INFORMACIÓN ESTADÍSTICA.

INTEGRIDAD: Completitud de la información.

VERACIDAD: Reducción de errores. La información estadística debe reflejar lo más aproximadamente posible la realidad objetiva.

CONFIABILIDAD: Se refiere a establecer márgenes de errores o intervalos de confianza de los datos estadísticos que se ofrecen.

OPORTUNIDAD: Que la información estadística esté " en tiempo " para los fines de la toma de decisiones.

ESTABILIDAD: La calidad de la información estadística requiere de estabilidad metodológica para asegurar su consistencia. Frecuentes cambios de procederes o definiciones desestabilizan un SIE y afectan la calidad de la información estadística.

SUFICIENCIA: Que la información estadística satisfaga las necesidades de información de los usuarios. El exceso de información implica gastos inútiles y poca utilización.

COMPATIBILIDAD: Definiciones y procederes compatibles dentro del SIE y entre los SIE. Sólo así se pueden hacer comparaciones en el tiempo y entre programas afines.

FLEXIBILIDAD: Se refiere a la información por niveles que ofrezca el SIE. A superior nivel del sistema de salud, información más general para el mando estratégico, fundamentalmente. En los niveles inferiores y más operativos el SIE debe ofrecer información detallada para la gerencia a ese nivel.

CLARIDAD: Expresión de la información clara y sencilla, explícita y bien definida.

EQUILIBRIO: La información que ofrece un SIE debe ser equilibrada de tal manera que exprese las principales actividades de todo el programa a que se refiere.

ECONOMÍA: Costos de implantación y funcionamiento de un SIE. Un SIE costoso difícilmente sea perdurable o sostenible por largo tiempo. Los SIE deben diseñarse con un pensamiento racional en términos de costos, deben ser económicos y eficientes.

PERTINENCIA: La información que ofrece un SIE debe ser pertinente a la situación de cada momento y debe prescindir de emitir información tangencial al problema de medición central. Ello conlleva a confusiones, distorsión de los resultados, elevación de los costos y baja eficiencia.

2.6 Comunicación Estadística.

Se entiende por Comunicación Estadística, el reflejo o proyección del proceso y resultados de la actividad estadística. Sin duda el cuidado con que se realice la misma favorecerá el mejor uso de los datos en sus diferentes funciones, fundamentalmente, en el proceso de toma de decisiones.

La comunicación estadística puede hacerse de diferentes formas que varían en función del contenido, la extensión y en aspectos formales propiamente dichos.

Antes de decidir la forma de la comunicación estadística se debe establecer lo siguiente:

- ¿A quién va dirigida?
- ¿Qué nos proponemos comunicar o para qué la necesitan los usuarios o receptores?
- ¿Cómo recibirán los resultados (vía, lugar, soporte, forma oral o escrita, etc.)
- ¿Cuándo o cada que tiempo se les comunicará la información?

- <u>A quién va dirigida</u>:

La información puede estar dirigida a dirigentes, investigadores, personal docente (profesores o alumnos), organismos internacionales, órganos de Gobierno, Partido o a un grupo heterogéneo de receptores para una misma información.

De acuerdo a quien va dirigida debe delimitarse:

- Nivel de Información que tienen sobre el tema.
- Qué necesita saber.
- Qué información previa necesitan para la total comprensión de lo que se va a comunicar.

Dependiendo de esto se identificará la profundidad, expresión en texto, gráfica o tabular, nivel de detalle, antecedentes del problema, referencias y otros aspectos que pueden o no resultar necesarios.

Por ejemplo: Un dirigente tendrá más interés en la fiabilidad de la información y en su aplicación en el proceso de toma de decisiones que realiza.

Los especialistas en el tema, requerirán más detalles así como los responsables de la actividad o programa en específico. El personal docente necesitará quizás una breve descripción del proceso estadístico que sustenta los resultados analizados y organizados así como la teoría que respalda la publicación a fin de instruir y poder evaluar a posteriori los conocimientos de los educandos. Si se trata de estadísticos requerirán saber además de los resultados, elementos técnicos del diseño del SIE que provee la información o características de los indicadores que se presentan para, de ser posible, repetir el proceso.

<u>Qué nos proponemos comunicar</u>:

Es necesario saber que se necesita trasmitir en dependencia de que se necesita conocer. Por tanto es útil decir o tener claro, por qué se da esa información, qué importancia reviste, qué se propuso estudiar y trasmitir por él o los autores, qué se encontró, qué conclusiones y qué recomendaciones se proponen.

Se debe siempre identificar las necesidades de información anteponiendo la relevancia y la selectividad a la cantidad, siempre que sea posible. Se debe además identificar la cultura organizacional en estadística y en el tema específico que se aborda.

<u>Cómo comunicar la información</u>:

Es necesario definir como trasmitir la información. La comunicación estadística se realiza de dos formas básicas, la oral y la escrita.

Una premisa para cualquiera de las dos formas es la utilización de un lenguaje claro y sencillo así como considerar todo aquello posible que facilite la comprensión, NO que la entorpezca.

Usar la expresión tabular cuando realmente aclare y haga relevante lo que es necesario, los gráficos no deben ser usados arbitrariamente ya que pueden confundir o distraer al usuario del propósito del trabajo. Los gráficos o imágenes solo deben usarse para destacar algún resultado importante o relevante.

En la comunicación oral, se debe hacer un uso pertinente de los medios audiovisuales, entre los que no se deben olvidar los rotafolios y pancartas. La presentación en display de una computadora sin más ayuda, debe reservarse para poca audiencia dada la estrecha visibilidad que ofrecen, de ser audiencias mayores a esta deben adjuntarse medios potenciadores de la imagen.

El interés que un auditorio o grupo de lectores puede tener sobre determinada comunicación estadística, varía de acuerdo a las características de la actividad profesional que los mismos realizan.

COMUNICACIÓN ESCRITA

Puede adoptar diferentes modalidades según esté dirigido a:

- ***Presentación Científica.***
- ***Aplicación a la práctica social.***
- ***Dirigentes.***
- ***Evento Científico.***
- ***Publicación eventual o periódica..***

Todas ellas tienen en común que deben decir:

- ***Para qué se realizó.***
- ***Importancia de los resultados que se muestran.***
- ***Propósito.***
- ***Como se hizo el estudio.***
- ***Qué resultados se encontraron.***
- ***Conclusiones.***
- ***Recomendaciones.***

COMUNICACIÓN ORAL

Debe adaptarse a las características del auditorio.

La extensión y el contenido dependerán de lo que sea necesario trasmitir en función de la mejor interpretación. Debe conocerse y adaptarse la comunicación al tiempo disponible.

La presentación debe ser sencilla, comprensible y ajustada al tiempo establecido.

La comunicación oral tiene la ventaja de que da oportunidad a preguntas y profundización en ciertos aspectos que no hayan quedado lo suficientemente claros.

Consejos para la presentación oral:

- ***Divida la exposición en Introducción, Desarrollo y Conclusiones / recomendaciones.***
- ***Apoye la presentación con medios audiovisuales que simplifiquen la comprensión NO que la entorpezcan.***
- ***Evite la improvisación.***
- ***Ajuste el tiempo de presentación.***
- ***Cerciórese de que sabe manejar los medios audiovisuales.***
- ***Sitúese en el lugar indicado. No dé la espalda al auditórium ni atraviese el haz de luz del retro o proyector. No gesticule de forma exagerada.***
- ***No lea, exponga con espontaneidad. Hable despacio y evite el uso de jergas y siglas.***

<u>Cuándo comunicar la información</u>:

Se refiere a en qué momento o cada que tiempo trasmitir la información. Se refiere a la periodicidad.

3- Clasificación Internacional de Enfermedades (CIE)

3.1- <u>Importancia del acto de clasificar</u>.

Es un acto necesario para el estudio de cualquier fenómeno, es la base de la generalización científica y desde el punto de vista metodológico, es esencial para unificar definiciones y sistemas de clasificación. Permite la comparación y el análisis de la información uniformemente clasificada.

3.2- Definición de Clasificación Internacional de Enfermedades. (CIE)

Una clasificación de enfermedades puede definirse como un sistema de categorías a la que se asignan entidades morbosas de conformidad con criterios establecidos. La clasificación puede girar en torno a muchos ejes posibles y la elección de uno en particular estará determinada por el uso que se hará de las estadísticas recopiladas. Una clasificación estadística de enfermedades debe abarcar toda la gama de estados morbosos dentro de un número manuable de categorías.

El hecho de ser Internacional, proviene de su aprobación y utilización por un elevado número de países del planeta.

3.3- Breve recuento histórico.

La Décima Revisión de la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud, vigente en Cuba desde el primero de Enero del año 2000, es la más reciente de una serie que se formalizó en 1893 como Clasificación de Bertillón o Lista Internacional de Causas de defunción.

En una reunión celebrada en Viena en 1891, el Instituto Internacional de Estadísticas, sucesor del Congreso Internacional de Estadística, encargó a un comité dirigido por Jacques Bertillón, jefe de los Servicios de Estadística de la ciudad de París, la preparación de una clasificación de causas de defunción. La clasificación de Causas de Defunción de Bertillón, recibió aprobación general y fue acogida por varios países y numerosas ciudades.

El Gobierno de Francia convocó en París en Agosto de 1900, la primera Conferencia Internacional para la revisión de la Lista de Bertillón o Lista Internacional de Causas de defunción, a donde asistieron delegados de 26 países y el 21 de Agosto de 1900 se adoptó una clasificación detallada de las causas de defunción, que comprendía 179 grupos y una clasificación abreviada que comprendía 35 grupos.

Se reconoció lo ventajoso de una revisión cada 10 años y se convocaron las mismas con esa o similar periodicidad. El fallecimiento de Bertillón en 1922, privó a la Conferencia Internacional de su verdadero líder, lo cual trajo consigo acciones para no abandonar esta actividad hasta que en 1946, en la Conferencia Internacional de la Salud, reunida en Nueva York, encargó a una comisión interina de la Organización Mundial de la Salud la responsabilidad de la sucesiva revisión de la Lista de Causas de Defunciones y el establecimiento de las Listas Internacionales de Causas de Morbilidad.

Para ello se nombró un Comité de Expertos integrado por especialistas de alta calificación y de diferentes disciplinas de varios países.

En 1948 la Primera Asamblea Mundial de la salud aprobó el informe de la Conferencia para la sexta revisión. La Conferencia para la sexta revisión decenal fue el comienzo de una nueva era en las estadísticas vitales y de salud internacionales.

Se han realizado revisiones de las listas o clasificación cada diez años aproximadamente, actualizándolas acorde al desarrollo de la humanidad en las ciencias médicas. Los cambios trascendentales se realizan cada 20 años a fin de no afectar grandemente la estabilidad metodológica y las comparaciones seculares. Hoy día se considera que revisiones decenales no resultan convenientes por el corto plazo de tiempo para una buena revisión e implantación.

La CIE X fue aprobada en 1989 por la Conferencia Internacional para la Décima Revisión de la Clasificación Internacional de Enfermedades y aprobada por la 43ª Asamblea Mundial de la Salud.

Actualmente existen nueve Centros Colaboradores de la OMS para la Clasificación Internacional de Enfermedades para asistir a los países en los problemas hallados en el desarrollo y uso de las clasificaciones relacionadas con la salud y en especial con el uso de la CIE. Además de estos nueve Centros Colaboradores de la OMS existen varios centros nacionales de referencia. Cuando los usuarios encuentran problemas deben consultar primero a estos centros.

Centros Colaboradores de la OMS:

- Centro Venezolano de la CIE. Para países de habla hispana.
- Instituto Australiano de Salud. Canberra, Australia. Para idioma inglés.
- Oficina de Censos y Encuestas de Población. Londres, Inglaterra. Para idioma inglés.
- Centro Nacional de Estadísticas de Salud. Hyattsville. Estados Unidos de América. Para idioma inglés.
- Hospital de la Unión de Colegios Médicos de Pekín. Pekín, China. Para idioma chino.
- INSERM. Francia. Para idioma francés.
- Departamento de Medicina Social. Uppsala, Suecia. Para los países nórdicos.
- Facultad de Salud Pública / Universidad de Sao Pablo. Sao Pablo, Brasil. Para el idioma portugués.
- Instituto N. A. Semasko. Moscú, Federación Rusa. Para idioma ruso.

3.4- Utilidad de la Clasificación Internacional de Enfermedades.

- Permite el conocimiento organizado de las causas de muerte, traumatismos, enfermedades y lesiones clasificadas de acuerdo a diferentes intereses.

- Logra uniformidad en la terminología y definiciones lo que permite realizar comparaciones entre países y regiones.

- Permite conocer las causas que conducen directamente a la muerte y las que desencadenan el proceso.

- Permite el conocimiento de las enfermedades y otras dolencias que padece la población a partir de terminologías y definiciones uniformes que facilitan su estudio y análisis comparativo.

- Contribuye a desarrollar una labor preventiva eficiente.

- Contribuye a elevar la calidad de la atención médica.

- Constituye una herramienta metodológica de valor para la investigación en mortalidad, morbilidad y otros problemas relacionados con la salud.

3.5- Aspectos básicos para el uso y aplicación de la Clasificación Internacional de Enfermedades.

- Llenado correcto del certificado médico de defunción y otros registros que se someten a codificación como la hoja de egreso hospitalario u hojas de consulta ambulatoria que contienen diagnósticos.

- Correcta codificación de las enfermedades y causas de muerte.

En Cuba, el certificado médico de defunción es llenado sólo por el médico que asiste la muerte. La parte de este modelo en que se consignan las causas de muerte exige un pensamiento científico por parte del médico certificante, de forma que se garantice se consignen las causas de muerte a través de un proceso lógico y se garantice la calidad de la información estadística de mortalidad.

El llenado correcto del certificado médico de defunción por el personal facultativo es de vital importancia para garantizar el éxito de la aplicación de la CIE.

Para las defunciones de 28 días y más las causas de muerte deben consignarse bajo el enfoque de Causa Básica.

Causa Básica de Defunción: En la Conferencia para la Sexta Revisión de la CIE se acordó que la causa de muerte para la tabulación primaria se denominara causa básica de la defunción. Esta es la enfermedad o lesión que inició la cadena de acontecimientos patológicos que condujeron directamente a la muerte, o las circunstancias del accidente o violencia que produjo la lesión fatal.

Desde el punto de vista de la prevención de la muerte, es necesario romper la cadena de sucesos o realizar la curación en algún momento de la evolución de la enfermedad. El objetivo más efectivo de los programas de salud es prevenir la causa que da origen a todos los demás trastornos o afecciones que conducen a la muerte. Esta decisión descansa evidentemente en un enfoque epidemiológico.

¿Cómo llenar correctamente un certificado médico de defunción en su variable "causas de muerte"?

En la Parte I del certificado, se anotan las enfermedades relacionadas con la cadena de acontecimientos que condujeron directamente a la muerte y en la Parte II se indican otras entidades morbosas que hubieran contribuido, pero que no están relacionadas directamente con la causa directa de muerte.

La afección registrada en la última línea de la Parte I es la causa básica y será la utilizada para las tabulaciones. Las causas consecuentes a esta se sitúan en las líneas superiores y en el orden que se presentaron en la historia de enfermedad y muerte de la persona.

Ej. : Mujer de 65 años.
Parte I

- Embolia pulmonar.
- Fractura patológica.
- Carcinoma secundario del fémur.
- Carcinoma de la mama.

Parte II:

Se anotaría cualquier estado patológico importante que contribuyó a la muerte, pero que no estuvo relacionado con la enfermedad que condujo directamente a ella.

La causa básica en este ejemplo es: Carcinoma de mama. Esta será la causa a tabular en las estadísticas rutinarias de mortalidad. El código alfa numérico que le corresponde a esta entidad según la CIE X es C50.9. Estos códigos conforman la base de datos automatizada de mortalidad la que se compone de un nomenclador que identifica cada categoría alfa numérica. Las CIE anteriores proveían sólo categorías numéricas. La CIE X provee categorías alfa numéricas (letras y números). Se utilizan las letras del abecedario a excepción de la U que provee 50 categorías para utilizarse en la investigación, en la asignación provisional de nuevas enfermedades de etiología incierta u otras necesidades locales.

El llenado de la variable causa de muerte en los certificados médicos de defunción para menores de 28 días en Cuba, se llena bajo el enfoque de "causa principal". En la primera línea, de arriba hacia abajo, se coloca la causa que a juicio del médico certificante fue la principal o de mayor importancia

para provocar la muerte. En la segunda línea, una o dos causas más (de existir) que consideren tuvieron también importancia. Se codifica en estos casos la causa principal.

El proceso de codificación la realiza el personal técnico y profesional entrenado que labora en las estructuras estadísticas de salud.

La codificación correcta se logra por la habilidad, pericia y conocimientos del personal en la aplicación de los procederes establecidos en la CIE.

La correcta codificación de las causas de muerte o enfermedad depende fundamentalmente de la adecuada aplicación de las reglas y procedimientos para el acto de la codificación, según la CIE X. Para esto es necesario saber aplicar las reglas de selección y modificación de causa básica de muerte, o de enfermedad si fuera el caso. Estas reglas tienen como propósito obtener la causa verdadera de muerte o enfermedad. Las reglas tratan de reducir al mínimo la arbitrariedad por parte de los codificadores, es decir, implican un algoritmo lógico y científico de selección de la causa a tabular respetando al máximo lo consignado por el medico certificante, que es el que registra el dato primario. El conocimiento de las reglas y su aplicación es objeto de adiestramiento permanente a los codificadores.

El acto de la codificación adecuada garantiza en gran medida la calidad de las estadísticas de mortalidad y de morbilidad.

3.6- Estructura de la CIE X.

La CIE X comprende tres volúmenes, el Volumen 1 que contiene la lista tabular de inclusiones y subcategorías de cuatro dígitos, el Volumen 2 que provee información a los usuarios de la CIE y el Volumen 3 que es el índice alfabético de la clasificación.

La mayor parte del Volumen I se dedica a la clasificación principal. También contiene otras listas especiales de tabulación, Definiciones y Reglamento de Nomenclatura.

En el volumen 2 se presentan las orientaciones y reglas de codificación de mortalidad y morbilidad.

El Volumen 3 contiene el índice alfabético para la lista tabular del Volumen 1. El Volumen 1 es la herramienta primordial para la codificación, sin embargo el índice alfabético del Volumen 3, es un complemento esencial de la lista tabular del Volumen 1, puesto que contiene un gran numero de términos diagnósticos que no aparecen en el Volumen 1 del que los médicos pueden hacer uso de ellos, por lo que el codificador debe usar ambos volúmenes conjuntamente. El Volumen 3 de índice alfabético de diagnósticos, tiene el propósito de incluir la gran mayoría de los términos diagnósticos que se usan en la actualidad, abarca

incluso términos imprecisos e indeseados, dado que esos términos todavía aparecen en los registros médicos y los codificadores necesitan una indicación para asignar un código de la clasificación, aun cuando lleve a una categoría residual o mal definida. Por tanto la presencia de estos términos en el Volumen 3, no debe interpretarse como una aprobación de su uso como terminología medica adecuada. Este índice se organiza por orden alfabético a fin de facilitar el trabajo de búsqueda de los términos diagnósticos por los codificadores.

La CIE X se divide en 21 capítulos. El primer carácter del código de la CIE es una letra, y cada letra se asocia a un capítulo en particular a excepción de la letra D que se asocia a dos capítulos: II y III y la letra H que se utiliza en el capítulo VII y VIII. Cuatro capítulos utilizan más de una letra en la primera posición de sus códigos: I, II, XIX y XX.

Capítulos de la CIE X:

I Ciertas Enfermedades Infecciosas y Parasitarias.

II Tumores (Neoplasias)

III Enfermedades de la sangre y de los órganos hematopoyéticos y ciertos trastornos que afectan la inmunidad.

IV Enfermedades endocrinas, nutricionales y metabólicas.

V Trastornos mentales y del comportamiento.

VI Enfermedades del sistema nervioso.

VII Enfermedades del ojo y sus anexos.

VIII Enfermedades del oído y de la apófisis mastoides.

IX Enfermedades del sistema circulatorio.

X Enfermedades del sistema respiratorio.

XI Enfermedades del sistema digestivo.

XII Enfermedades de la piel y del tejido subcutáneo.

XIII Enfermedades del sistema osteomuscular y del tejido conjuntivo.

XIV Enfermedades del sistema genitourinario.

XV Embarazo, parto y puerperio.

XVI Ciertas afecciones originadas en el período perinatal.

XVII Malformaciones congénitas, deformidades y anomalías cromosómicas.

XVIII Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte.

XIX Traumatismos, envenenamientos y algunas otras consecuencias de causas externas.

XX Causas externas de morbilidad y mortalidad.

XXI Factores que influyen en el estado de salud y contacto con los servicios de salud.

Cada capítulo se divide en grupos de categorías de 3 caracteres. Las categorías de 3 caracteres, corresponden a afecciones únicas, seleccionadas debido a su frecuencia, gravedad o vulnerabilidad a las acciones de salud.

La mayoría de las categorías de 3 caracteres están divididas por medio de un carácter después del punto decimal, lo que permite hasta 10 subcategorías. Cuando una categoría no está subdividida puede utilizarse la letra X para llenar la cuarta posición, de tal manera que los códigos tengan una longitud estándar cuando lo requieran los sistemas de procesamiento automatizados de datos.

Se provee las subdivisiones suplementarias para uso al nivel de quinto carácter o subsecuentes para fines específicos que prevé la CIE X.

Ejemplos:

1) Angina de pecho inestable.

Capítulo IX: Enfermedades del Sistema Circulatorio. (I00 – I99)
Grupo: Enfermedades isquemias del corazón: I20 – I25
Categoría: Angina de pecho: Código: I20
Subcategoría: Angina inestable: I20.0

2) Demencia en la enfermedad de Alzheimer de comienzo tardío.

Capítulo V: Trastornos mentales y del comportamiento. (F00 – F99)
Grupo: Trastornos mentales orgánicos, incluidos los trastornos sintomáticos (F00 – F09)
Categoría: Demencia en la enfermedad de Alzheimer F00.
Sub-categoría: Demencia en enfermedad de Alzheimer de comienzo tardío: F00.1

3) Infarto Agudo del Miocardio.

Capítulo: Enfermedades del sistema circulatorio (I00 – I99)
Grupo: Enfermedades isquémicas del corazón. (I20 – I25)
Categoría: Infarto agudo del miocardio: I21
No se subdivide en sub-categoría por lo que el código de a esta entidad es I21.X

Las estadísticas de morbilidad y mortalidad se expresan tabuladas y resumidas en listas abreviadas de mayor o menor extensión, que la CIE propone y que el país considere mas adecuada.

Para profundizar en la CIE X consultar: "Clasificación Estadística Internacional de Enfermedades y Problemas relacionados con la Salud" Volumen I, II y Publicación científica No. 554. OPS / OMS Washington, DC. EUA. 1995.

4. Estadísticas de Mortalidad

4.1 Estadísticas de Mortalidad. Conceptos.

La mortalidad es uno de los componentes que intervienen en la estructura por edad y sexo de la población junto a la fecundidad y las migraciones.

La muerte o defunción es uno de los hechos vitales que forman parte de las estadísticas vitales de mayor interés para la salud publica.

La mortalidad es la acción de la muerte sobre la población, según el concepto demográfico. Es importante para medir el Estado de Salud de la Población al constituir un efecto fenoménico de las variaciones que ocurren en esta.

Las Estadísticas de Mortalidad forman parte de las Estadísticas Vitales, de gran interés para el quehacer en salud.

Las estadísticas de mortalidad son aquellas que tienen como propósito conocer el numero de defunciones habidas en determinada colectividad humana durante un periodo de tiempo definido y su distribución de acuerdo a diferentes características de la población, entre estas características, las causas de muerte son de especial interés.

4.2 Utilidad de las estadísticas de mortalidad.

Son de gran utilidad para la planificación, ejecución y control de programas y acciones de salud.

Utilizadas sistemáticamente en los estudios y cálculos demográficos de población y en especial del indicador Esperanza de Vida.

Son muy usadas en investigaciones y para la docencia de pregrado y postgrado

4.3 Tendencia de la mortalidad

La tendencia de la mortalidad, desde hace alrededor de 200 años, ha sido descendente, en términos generales, debido al desarrollo científico de la humanidad, el control y prevención de muchas enfermedades principalmente transmisibles, así como el desarrollo socio económico de las sociedades. Las mejoras higiénicas y sanitarias y el desarrollo socio económico de las sociedades, han elevado las condiciones de vida de las personas y ello ha influido en el descenso de la mortalidad. Estas mejoras no han ocurrido de modo parejo para todos los países del mundo, incluso se aprecian diferencias entre regiones de un mismo país, lo que sitúa a países y regiones en condiciones de franca desventaja con relación a otros.

Factores que intervienen en las variaciones de la mortalidad.

* Factores Biológicos. Ej. : Edad y Sexo.

Sexo: A lo largo de toda la vida existe sobre mortalidad masculina Ocurren más nacimientos de niños que de niñas en una relación de 105 %, o sea 105 varones por cada 100 hembras. El efecto de la mortalidad hace que haya más mujeres que hombres en edades más avanzadas de la vida.

Los factores que favorecen la sobre mortalidad masculina pueden ser factores biológicos como la mayor reserva genética y la resistencia que poseen las mujeres a la enfermedad y muerte. Los factores sociales influyen también. La sociedad ha impuesto históricamente a cada sexo patrones de conducta y realización de actividades diferenciadas lo que ha provocado perfiles de mortalidad distintos para cada sexo, por ejemplo, los hombres desempeñan con frecuencia trabajos más riesgosos físicamente que las mujeres. Por otra parte es conocido que las mujeres muestran mayor responsabilidad por su salud. Las mujeres tienen una esperanza de vida al nacer superior a la de los hombres y alcanzan a vivir en general más años.

Edad: En países de bajo desarrollo socio económico la estructura de la mortalidad según edad es la siguiente: mortalidad elevada en las primeras edades de la vida lo que se corresponde con una pirámide poblacional de base ancha y vértice estrecho donde la mortalidad afecta a lo largo de toda la vida. Las causas que más afectan son las causas exógenas. En los países de mayor desarrollo socioeconómico la mortalidad en los menores de un año y en las edades más jóvenes, disminuye mientras se eleva la mortalidad en los grupos de edades más avanzadas. La pirámide de la población se presenta con base estrecha, cuerpo y vértice ancho, porque logran llegar mayor cantidad de efectivos poblacionales a edades avanzadas de la vida. Las causas

que más afectan son las endógenas, es decir, las caracterizadas por factores causales biológicos y degenerativos, por ejemplo: enfermedades cerebro vasculares y tumores malignos.

- Organización social.

La Organización Social establece el modo de producción de la sociedad, la distribución de los productos, los alimentos, la gratuidad o bajos precios de las medicinas y servicios médicos en general. Todo ello es diferente para cada país o región según la organización social que posean. Estos factores diferentes determinan y condicionan distintos perfiles de mortalidad.

- Medio ambiente

El Medio Ambiente incluye el clima, el aire, los procesos ecológicos y otros. Cuando estos componentes son desfavorables influyen desfavorablemente en la mortalidad.

- Factores causales. Factores exógenos y endógenos.

Los factores exógenos que causan muerte con frecuencia, son típicos de países subdesarrollados o con importante situación de pobreza y débiles sistemas de salud. Los endógenos tales como las enfermedades crónicas no transmisibles y accidentales son mas frecuentes en sociedades desarrolladas, con sistemas de salud desarrollados, con poblaciones donde sus individuos tienen elevada esperanza de vida, en las que están presentes además hábitos y estilos de vida inadecuados, estrés y otros.

4.5 Sistema de Información Estadístico de Defunciones y Defunciones Peri- natales en Cuba.

Es el conjunto de procedimientos encaminados a recolectar, procesar, validar y emitir datos relacionados con la mortalidad.

El Sistema de Información Estadístico de mortalidad como también suele llamarse, forma parte del Sistema de Información Estadístico Nacional (SIEN).

La captación del dato primario se realiza en todas y cada una de las unidades de salud a quienes se les ha atribuido la función de legalización del certificado médico de defunción, registro primario del sistema. La estructura estadística de salud es la responsable de procesar, validar y consolidar esta información fungiendo como una prolongación de las Oficinas del Registro del Estado Civil por resolución entre los respectivos Ministerios, de Justicia y de Salud Publica.

El personal autorizado como registrador de los datos de mortalidad en los certificados médicos de defunción es única y exclusivamente el personal medico.

En Cuba están vigentes tres certificados médicos de defunción que son los modelos de entrada al sistema. Estos son:

Certificado medico de defunción para las defunciones de 28 días y más de edad.

Certificado médico de defunción neonatal para defunciones de nacidos vivos hasta los 27 días de vida.

Certificado medico de defunción fetal para mortinatos o defunciones fetales de 20 semanas o más de gestación o de 500 gramos de peso o más.

Los certificados médicos de defunción constan de dos originales que son llenados por el médico ante el hecho vital: muerte. Un original es enviado para su archivo, custodia y uso individual a las Oficinas del Registro del Estado Civil, a excepción del fetal. El otro original fluye a través de las estructuras estadísticas de salud donde se revisan y tabulan.

En el nivel provincial se codifican las causas de muerte según lo establecido en la CIE X, se almacena la información en soporte electrónico, y se envía en ficheros por vía electrónica al nivel de nación. De esta manera se conforman las bases de datos de mortalidad en provincia y nación. En todos los niveles del sistema nacional de salud se valida la información y se solicitan los reparos de ser necesario de acuerdo a procederes y calendario establecido. Los reparos a los certificados médicos de defunción pueden ser confeccionados solo por el medico certificante y se solicitan, ante la sospecha de los revisores, de inconsistencias en las causas de muerte consignadas o de otro tipo, así como omisiones de variables que establece el modelo. Se envían reparos cuando el resultado de la necropsia indica causas diferentes a las consignadas inicialmente o cuando se reciben exámenes de laboratorio o bacteriológicos que indican etiologías diferentes o especificas de la causa de defunción.

Los Departamentos de Estadística o Registros Médicos de las unidades de salud, donde se registra y legaliza el certificado de defunción, reciben el documento, tramitan el permiso de enterramiento, revisan su calidad e integridad y asientan los datos en el Libro de Registro de Defunciones de la unidad. Envían los certificados recibidos al nivel inmediato superior correspondiente según subordinación municipal, provincial o de nación.

En el Departamento Municipal de Estadística se recibe y revisa el documento y se confeccionan tablas de salida. Los certificados son enviados al Departamento Provincial de Estadística.

En el Departamento Provincial de Estadística, se recibe y se revisa la integridad del documento. Se codifica la causa básica de defunción utilizando la CIE X. Se realiza cotejo con la Oficina Provincial de Estadística. Se almacena la información en soporte electrónico confeccionando la base de datos provincial. Se confeccionan tablas de salida para ese nivel y se envían los certificados y ficheros a la Dirección Nacional de Estadística del MINSAP.

En la Dirección Nacional de Estadística se recibe y revisa la integridad del documento y se revisa la codificación realizada en la provincia. Se entrega a la Oficina Nacional de Estadística la base de datos nacional y se entrega información a organismos internacionales según compromisos oficiales establecidos. Se confeccionan tablas de salida para el país.

El SIE de defunciones y defunciones perinatales es uno de los SIE más íntegros y que ofrece información más confiable y veraz en el país. Esta afirmación se basa en las evaluaciones nacionales sistemáticas que se realizan así como las evaluaciones realizadas por expertos de organismos internacionales.

4.6 Medidas de la mortalidad

Mortalidad absoluta. Numero absoluto de defunciones para un lugar y tiempo dado. Ej. : Defunciones ocurridas en Cuba en 1999: 79 486

Mortalidad Proporcional: Proporción de defunciones de acuerdo a determinadas carac-terísticas con relación al total de defunciones ocurridas en un lugar y periodo de tiempo dado.

La mortalidad proporcional suele calcularse por sexo, edad, causa de muerte, región y otras variables de interés.

$$\text{Mortalidad Proporcional (Lugar y tiempo)} = \frac{\text{Defunciones según determinadas características}}{\text{Total de defunciones.}} \times 100$$

La Mortalidad proporcional por causas es útil cuando se requiere conocer la importancia relativa de algunas causas de muerte. Por ejemplo para evaluar el desarrollo de la Salud Pública de un territorio se pudiera, entre otros indicadores, calcular la mortalidad proporcional por enfermedades infecciosas y si es elevado, se puede suponer mal estado de las condiciones higiénicas y sanitarias y por tanto la necesidad de diseñar un plan de acción para mejorarlas.

$$\text{Mortalidad Proporcional por edad} = \frac{\text{Defunciones a una edad dada o para un grupo de edad.}}{\text{Total de defunciones de todas las edades}} \times 100$$

$$\text{Mortalidad Proporcional por sexo} = \frac{\text{Defunciones de un sexo dado}}{\text{Total de defunciones (ambos sexos)}} \times 100$$

El indicador de mortalidad proporcional tiene como ventaja:

1- Simple cálculo

2- No requiere la población expuesta al riesgo

Desventajas:

Ignora completamente las estructuras poblacionales por lo que no es útil para la comparación, no expresa el riesgo de morir o lo que es igual la frecuencia en términos de probabilidad de morir por alguna causa o a una edad o sexo.

Tasas de mortalidad:

Tipos de tasas de mortalidad

Tasa Bruta, cruda o general de Mortalidad: relaciona el total de defunciones con la población de un área y tiempo dado. Mide el riesgo absoluto y debe ser utilizada con cuidado en las comparaciones entre regiones o países dado que el denominador incluye toda la población, la cual puede diferir en estructura de edad u otra variable de un lugar a otro. Cuando esto sucede es necesario estandarizar o ajustar la tasa bruta de mortalidad u observar las tasas de mortalidad especifica por edad o sexo o la característica en cuestión.

Tasas Especificas de Mortalidad. Relaciona el numero de defunciones por alguna característica de los fallecidos (sexo, edad) con la población total que posee esa característica. Es el denominador de la tasa la que la hace especifica o no. La tasa de mortalidad por Infarto agudo del miocardio es una tasa general o bruta, ahora bien, la tasa de mortalidad por Infarto agudo del miocardio en la población de 45 a 59 anos, es especifica ya que en el numerador y en el denominador se circunscribe el dato a ese grupo de edad.

Tasa de mortalidad infantil y sus componentes (TMI)

La mortalidad infantil se refiere a los fallecidos menores de un ano.

$$TMI = \frac{\text{Defunciones menores de un año, lugar y tiempo dado}}{\text{Nacidos Vivos, lugar y tiempo dado}} \times 1000$$

Los componentes de la mortalidad infantil son: Mortalidad neonatal precoz (menores de 7 días), tardía (de 7 a 27 días) y la mortalidad posneonatal (de 28 días a 11 meses). Para cada componente se calcula una tasa con igual denominador al de la tasa de mortalidad infantil con el numerador relativo a los fallecidos de la edad definidos para cada componente.

Como para el calculo de estas tres tasas se usa el total de nacidos vivos en el denominador, al sumarlas se obtiene la Tasa de Mortalidad Infantil.

Tasa de Mortalidad Perinatal 1: Esta tasa abarca el periodo fetal tardío de muerte y el neonatal precoz, o sea el periodo cercano al nacimiento.

$$\text{TM Perinatal (1)} = \frac{\text{Defunciones Fetales tardías (28 semanas y más ó 1000 gr. y más)} + \text{Def. Neonatales precoces (0 a 6 días)}}{\text{Nacidos Vivos} + \text{Defunciones Fetales Tardías}} \times 10^n$$

Tasa de mortalidad Perinatal II: Comprende un periodo mayor al incluir los fallecidos fetales intermedios y tardíos y los fallecidos neonatales tanto precoz como tardío.

$$\text{TM Perinatal (2)} = \frac{\text{Fetales Intermedias (20 semanas y más ó 500gr y más)} + \text{Defunciones Neonatales (0-27días)}}{\text{Nacidos Vivos} + \text{Defunciones Fetales Intermedias y tardías}} \times 10^n$$

Tasa de mortalidad fetal: Comprende los fallecidos fetales tardíos.

$$\text{Tasa Mortalidad Fetal} = \frac{\text{Defunciones Fetales Tardías (1000g y más ó 28 semanas y más)}}{\text{Defunciones Fetales Tardías} + \text{Nacidos Vivos}} \times 10^n$$

Tasa de Mortalidad Materna Directa: Es la relación de las defunciones maternas provocadas por causas directas (embarazo, parto y puerperio) con los nacidos vivos. Es el cociente de estos dos números y se multiplica usualmente por 10 000 o por 100 000.

Tasa de Mortalidad Materna Indirecta: Es la relación de las defunciones maternas por causas indirectas (causas que se desencadenan durante el embarazo o se agudizan tales como la diabetes, el asma), con los nacidos vivos.

Tasa de Mortalidad Materna Total: Es la relación del total de muertes maternas, directas o indirectas, con el total de nacidos vivos de un periodo y un lugar.

Tasa de Letalidad: Es la relación entre las defunciones que ocurren por una causa con el total de enfermos de esa causa. Expresa la severidad de la enfermedad o daño..

Existen otras medidas de la mortalidad como son los Años de Vida Potenciales Perdidos, la Tabla de Mortalidad y su indicador por excelencia, la Esperanza de vida. Estos son objeto de estudio de las ciencias demográficas.

Ejercicios Resueltos

A continuación aparece información sobre población y defunciones de Cuba durante 1 999. Calcule:

1) Mortalidad proporcional masculina.
2) Mortalidad proporcional femenina.
3) Tasa de mortalidad para el sexo masculino
4) Tasa de mortalidad para el sexo femenino
5) Tasa bruta de mortalidad
6) Tasa de mortalidad Infantil de Cuba en 1999.
7) Interprete los resultados
8) Diga los Sistemas de Información Estadístico de donde se puede obtener información para el cálculo de estos indicadores

Sexo	Población	Defunciones	Porcentaje	Tasas *
Masculino	5 579 257	43 794	55.1	7.1
Femenino	5 563 434	35 692	44.9	6.8
Total	11 142 691	79 486	100.0	6.4

*Tasas por 1 000 habitantes

Nacimientos Vivos Cuba 1999: 150 785
Defunciones infantiles de menos de un ano Cuba 1999: 966
Tasa de Mortalidad Infantil Cuba 1999: 6.4 por 1000 Nacidos Vivos.

Las interpretaciones de cada indicador son:

- Por cada 100 defunciones hubo 55.1 fallecidos del sexo masculino.
- Por cada 100 defunciones hubo 44.9 fallecidas del sexo femenino.
- El riesgo de morir en los hombres fue de 7.1 por cada 1 000 habitantes o por cada 1000 habitantes murieron 7.1 hombres.
- El riesgo de morir en las mujeres fue de 6.8 por cada 1 000 habitantes o por cada 1000 habitantes murieron 6.8 mujeres.
- El riesgo de morir en Cuba en 1999 fue de 6.4 por cada 1 000 habitantes o por cada 1000 habitantes murieron 6.4 personas.

- El riesgo de morir antes del primer ano de edad en Cuba para 1999 fue de 6.4 por cada 1000 Nacidos Vivos.

La respuesta del inciso 8 es:

Las defunciones se pueden conocer a través de las salidas de información del SIE de Defunciones y Defunciones Perinatales y la población a partir de las proyecciones de población. Los Nacidos Vivos a partir de las salidas del SIE de nacimientos.

Ejercicios propuestos

A continuación aparece información de Cuba de 1999. Utilizando la que convenga calcule los indicadores siguientes:

1) Porcentaje de defunciones para el grupo de edad de 1 a 4 años
2) Tasa de mortalidad para el grupo de edad de 5 a 14 años
3) Tasa bruta de mortalidad
4) Tasa de mortalidad infantil
5) Tasa de mortalidad materna
6) Tasa de mortalidad neonatal precoz
7) Tasa de mortalidad neonatal tardía
8) Tasa de mortalidad post-neonatal
9) Porcentaje de defunciones por tumores malignos
10) Tasa de mortalidad por tumores malignos

- Interprete los resultados en todos los ejercicios
- Diga los Sistemas de Información Estadístico de donde se pueden obtener los datos para el cálculo de esos indicadores

Datos

- Población total: 11 187 673
- Población masculina: 5 598 493
- Población femenina: 5 589 180
- Población del grupo de edad de 50 a 59 años: 544 772
- Población del grupo de edad de 5 a 14 años: 1 721 746
- Total de nacidos vivos: 150 785
- Defunciones por enfermedades del corazón: 11 580
- Defunciones por tumores malignos: 9 330
- Defunciones menores de 1 año: 966
- Defunciones de menores de 7 días: 431
- Defunciones de niños de 7 a 27 días: 155
- Defunciones de niños de 28 días a 11 meses: 380

- Total de defunciones: 79 486
- Defunciones de 1 a 4 años: 292
- Defunciones de 5 a 14 años: 525
- Mujeres fallecidas por causas del embarazo, parto y puerperio: 23 (dato ficticio)

5 Estadísticas de Natalidad

5.1 Generalidades.

Definición: Es aquella información numérica relacionada con los nacimientos que ocurren en cierta colectividad humana y su distribución de acuerdo a ciertas características del evento nacimiento per sé, así como características de los padres y el decursar del embarazo.

El nacimiento es otro hecho vital de gran interés para salud publica, por lo que defunciones y nacimientos son, de las estadísticas vitales, los que más se estudian en salud.

Otras definiciones básicas en las estadísticas de natalidad:

Nacido Vivo: Es el producto de la concepción que cualquiera que sea la duración del embarazo, sea expulsado o extraído completamente del seno materno, siempre que después de esa expulsión o extracción manifieste cualquier signo de vida.

Parto Institucional: Parto que ocurre en una institución del Sistema Nacional de Salud.

Tendencia de la Natalidad: La tendencia de la natalidad mundial es hacia la declinación. La declinación comenzó en Francia a principios del siglo XVIII y luego se produjo en otros países desarrollados. Hay grandes diferencias en los niveles de natalidad entre los países Por lo general hay una relación inversa entre el nivel socio económico de los países y su nivel de natalidad. En países de alto desarrollo hay en general baja natalidad y en los países en desarrollo alta natalidad.

Las causas que han determinado el descenso de la natalidad a nivel mundial son:

- Declinación real de la capacidad reproductiva.
- Factores culturales que rigen las costumbres matrimoniales. Incremento del divorcio y la soltería.
- Limitación voluntaria del tamaño de la familia, programas de planificación familiar y patrones culturales de fecundidad.
- Incremento de la Infertilidad dada entre otras, por hábitos y enfermedades del mundo moderno: alcoholismo, drogadicción, enfermedades de transmisión sexual y otros.
- Disminución de la mortalidad infantil y pre-escolar que otorga seguridad a la decisión de tener pocos hijos ya que con alta probabilidad sobrevivirán.

5.2 Sistema de Información Estadístico de Natalidad.

El sistema de información estadístico de nacimientos se incluye en el SIEN. El registro del dato primario, es el modelo de Inscripción de Nacimiento el cual es llenado por personal técnico de las estructuras estadísticas de los hospitales donde ocurren nacimientos o por el personal de las Oficinas del Registro del Estado Civil para los nacimientos extra institucionales que en nuestro país es solo el 0.1% del total de nacimientos. El 99.9% de los nacimientos en el país son institucionales, de ahí que los hospitales donde estos ocurren se conviertan en el lugar ideal para el registro del dato primario, esto se incorporó al quehacer estadístico de salud desde la década de los años 60 en virtud de una resolución entre los ministros de salud y de justicia.

El procedimiento de inscripción y registro es el siguiente:

Cada día el personal de estadística de los hospitales revisa el libro de partos para saber de los nacimientos vivos acontecidos durante el día anterior. Localiza la ubicación en sala del recién nacido y su mama y acude en un plazo de entre 24 y 48 horas posterior al nacimiento a inscribirlo. La declarante por excelencia es la madre quien debe mostrar su carne de identidad. Se confecciona un original y una copia por un técnico preparado para desempeñar esta función. El original pasa a las Oficinas del Registro del Estado Civil de la localidad o municipio donde reside la madre, donde se archiva y custodia para el uso individual que requiera. La copia fluye por las estructuras estadísticas de la Oficina Nacional de Estadística donde se valida, procesa y emite información anual de natalidad.

Actualmente se dispone de bases de datos automatizadas nacionales y provinciales confeccionadas a partir del modelo de inscripción de nacimientos. Esto facilita el almacenamiento y recuperación de la información para diversos usos.

Por el SIEC de salud fluye también información de nacimientos y otras variables como peso al nacer y otras con periodicidad más breve, tales como la semanal, mensual, trimestral, semestral y anual. Se recoge, en varios casos, por ocurrencia del hecho y además por lugar de residencia, así se satisface el interés gerencial y epidemiológico. La información del SIEC y del SIEN se coteja periódicamente a fin de ajustar errores.

5.3 Usos de la Estadística de Natalidad

Son de gran utilidad para la administración científica, estudios demográficos, investigaciones en el área de la salud reproductiva y otras. En la docencia y en otras actividades de diferentes ramas socio económicas del país.

5.4 Medidas de la natalidad y la fecundidad

Para la mejor comprensión de los indicadores es necesario considerar tres definiciones importantes:

Natalidad: Se refiere a la natalidad efectiva o real, es decir a la frecuencia de nacidos vivos que ocurren en el seno de una población.

Fecundidad: Es la capacidad real de reproducirse la población. Es una variable demográfica al igual que la mortalidad y las migraciones. Se basa en los nacimientos vivos acontecidos.

Fertilidad: Es la capacidad potencial de reproducirse la población. Esta relacionada con las características biológicas y físicas de los individuos.

Indicadores más utilizados:

Números absolutos, Proporción, Razón y Tasas.

Natalidad General: Numero total de nacidos vivos. Habitualmente se expresa en la tasa de natalidad.

$$\text{Tasa de Natalidad:} \quad \frac{\text{Nacidos Vivos}}{\text{Población total}} \times 1000$$

Es la probabilidad de que ocurran nacimientos vivos en una población. Expresa la reproducción de la población y por tanto su crecimiento. Requiere para su comparación entre regiones y países con diferentes estructuras por edad u otra, del ajuste o estandarización

Natalidad Proporcional: Proporción de nacidos vivos por una característica dada.

$$\text{Proporción de Nacidos Vivos} = \frac{\text{Nacidos Vivos de una categoría dada}}{\text{Total de Nacidos Vivos}} \times 100$$

Tasa General de Fecundidad:

$$\text{Tasa General de Fecundidad} = \frac{\text{Nacidos Vivos de una región y un periodo dado}}{\text{Población femenina de 15 a 49 años de una región y periodo dado}} \times 1\,000$$

Expresa la capacidad real de reproducirse de una población. Tiene en cuenta solo a la población femenina en su edad fértil. Se utilizan como grupos de edad fértil los de 15 a 49 y 12 a 49 anos. En Cuba suele usarse el grupo de 12 a 49. La especificidad de este grupo debe siempre aclararse en el indicador que se calcula y presenta. Aunque es una tasa específica debe ser usada con cautela para la comparación, ya que la estructura por edad de ese segmento poblacional puede diferir entre regiones y países así como que la

fecundidad no es igual a lo largo de todas las edades de la vida reproductiva en todos los lugares. Esto hace necesario en ocasiones la estandarización o el calculo de las tasas especificas por edad quinquenal.

La tasa de fecundidad general es generalmente es 4 o 5 veces más elevada que la tasa bruta de natalidad.

Tasas Específicas de Fecundidad por edad: Se refiere a la capacidad real de reproducción para grupos de edades específicos.

$$\text{Tasa Específica de Fecundidad por edad.} = \frac{\text{Nacidos Vivos de un grupo de edad dado de la madre región y periodo dado}}{\text{Población femenina de ese grupo de edad región y periodo dado}} \times 10^n$$

La representación gráfica a través del Polígono de Frecuencia de las tasas específicas de fecundidad, dibujan las Curvas de Fecundidad. Las curvas pueden ser Tempranas cuando la fecundidad más alta está en las edades entre 20 y 24 años, Tardía cuando la fecundidad más elevada está en el grupo de edad de 25 a 29 años y Dilatada cuando la fecundidad más elevada está entre 20 y 29 años, La fecundidad temprana es propia de regiones de condiciones socioeconómicas más desfavorables mientras que la tardía es característica de regiones de más desarrollo.

Tasa Global de Fecundidad: Expresa el número de hijos que en promedio tuviera cada mujer de una cohorte ficticia al terminar su vida fértil. Se debe suponer que la cohorte cumple con los supuestos siguientes:

- Que durante la vida fértil las mujeres tuvieran sus hijos de acuerdo a los niveles de la fecundidad de la población en estudio.
- Esta cohorte no está expuesta al riesgo de morir durante la etapa reproductiva.

Es por tanto una medida teórica en algunos aspectos pero puede ser calculada y ayuda a conocer los niveles de fecundidad de una población.

$$\text{Tasa Global de Fecundidad} = 5 \sum_{x=15}^{49} \text{tasas específicas por edad de fecundidad}$$

Tasa Bruta de Reproducción: La Tasa Bruta de Reproducción expresa la cantidad de hijas (hembras) que en promedio tuvieran las mujeres de una cohorte ficticia al terminar su edad fértil. Se debe suponer que la cohorte cumple con los supuestos siguientes:

- Que durante la vida fértil las mujeres tuvieran sus hijos de acuerdo a los niveles de la fecundidad de la población en estudio.

- Esta cohorte no está expuesta al riesgo de morir durante la etapa reproductiva.

$$\text{Tasa Bruta de Reproducción} = 5 \times K \times \sum_{x=15}^{49} \text{tasas específicas por edad de fecundidad}$$

Donde K = 0.4878. Se refiere a la proporción de nacimientos femeninos.

Tasa Neta de Reproducción: Es la tasa de fecundidad más refinada. Expresa la cantidad de hijas que en promedio tuvieran las mujeres de una cohorte ficticia al terminar su edad fértil, no tiene que cumplirse el segundo supuesto referente al riesgo de morir. Su cálculo es más complicado porque hay que utilizar una Tabla de Mortalidad, o sea incorpora el riesgo de morir de las mujeres a esas edades.

El valor de la Tasa Neta de Reproducción mide las condiciones de reemplazo de una población, si es 1 o más la población tiene garantizado el reemplazo poblacional porque cada mujer tiene una hija o más.

Existe relación entre la Tasa Bruta de Reproducción y la Tesa Neta de Reproducción porque las dos miden lo mismo, lo que en la primera no se tiene en cuenta el riesgo de morir. La Tasa Bruta de Reproducción sobrestima generalmente en 1.03 veces a la Tasa Neta de Reproducción lo cual es un valor pequeño, por lo que la Tasa Bruta de Reproducción es la mas utilizada por su sencillez de calculo y fiabilidad.

Ejercicios Resueltos

1- La población de Cuba durante 1998 fue de 11 122 308 habitantes y los nacidos vivos fueron 151 080, de ellos en instituciones de salud nacieron 150 897. Calcule e interprete la Tasas Bruta de Natalidad y el porcentaje de nacidos vivos en instituciones de salud de Cuba en 1998.

Tasa Bruta de Natalidad de Cuba, 1998:

$$\frac{\text{Nacidos Vivos}}{\text{Población Total}} \times 1\,000 = \frac{151\,080}{11\,122\,308} \times 1\,000 = 13.6$$

Por cada 1 000 habitantes ocurrieron 13.6 nacidos vivos en Cuba durante 1 998.

Porcentaje de nacimientos vivos ocurridos en instituciones de salud:

$$\frac{\text{Nacidos Vivos en instituciones de salud}}{\text{Total de Nacidos Vivos}} \times 100 = \frac{150\,697}{151\,080} \times 100 = 99.9$$

Por cada 100 nacimientos, 99.9 ocurrieron en instituciones de salud en Cuba durante 1 998

2- Calcule la tasa general de fecundidad, la tasa global de fecundidad y la tasa bruta de reproducción con los datos que se brindan a continuación correspondiente a una provincia durante 1993. Interprete los resultados.

Analice la curva de fecundidad de esa población.

Grupos de edad	Población femenina	Nacidos Vivos	Tasas Específicas por edad de Fecundidad.
15 a 19	85 236	968	11.3
20 a 24	96 184	20 067	208.6
25 a 29	123 505	10 017	81.1
30 a 34	80 938	4347	53.7
35 a 39	70 501	1596	22.6
40 a 44	74 800	294	3.9
45 a 49	72 036	3	0.04
Total	603 200	37292	61.8*

Por cada 1 000 mujeres entre 25 y 29 años ocurren 81.1 nacimientos. De esta forma se pueden interpretar las tasas específicas de todos los grupos de edad.

Tasa General de Fecundidad. Por cada 1000 mujeres de 15 a 49 años nacen 61.8 nacidos vivos

Tasa Global de Fecundidad = 5 (11.3 + 208.6 + 81.1 + 53.7 + 22.6
+ 3.9 + 0.04)/1 0= 5* 192.74/ 1 000= 1.9

En promedio cada mujer tiene 1.9 hijos al terminar la edad fértil asumiendo que se cumplen los supuestos mencionados anteriormente.

Tasa Bruta de Reproducción = 5 * 0.4878 * 192.74 / 1 000 = 0.47

En promedio cada mujer tiene 0.47 hijas al terminar la edad fértil asumiendo que se cumplen los supuestos mencionados anteriormente. No hay reemplazo poblacional.

La población estudiada debe tener condiciones de vida favorables y alto desarrollo socioeconómico porque tiene baja natalidad, la curva de fecundidad es tardía y no tiene reemplazo poblacional.

Ejercicios Propuestos

1- La población de Cuba durante 1999 fue de 11 142 691 habitantes y los nacidos vivos fueron 150 785, de ellos en instituciones de salud nacieron 150 590. Calcule e interprete la Tasas Bruta de Natalidad y el porcentaje de nacidos vivos en instituciones de salud de Cuba en 1998.

2- Calcule las tasas específicas de fecundidad, la tasa general de fecundidad, la tasa global de fecundidad y la tasa bruta de reproducción con los datos que se brindan a continuación correspondiente a una región dada durante 1997. Interprete los resultados.

Grupos de edad	Población femenina	Nacidos Vivos
15 a 19	80 436	1 223
20 a 24	91 647	30 458
25 a 29	115 418	20 015
30 a 34	78 356	5 258
35 a 39	65 326	1 636
40 a 44	69 223	301
45 a 40	22 458	2

6- Estadísticas de Morbilidad

6.1 Concepto de Estadísticas de Morbilidad

Es la información numérica sobre enfermedades, traumatismos y sus secuelas, incapacidades y otras alteraciones de la salud diagnosticadas o detectadas en la población durante un período de tiempo.

6.2 Sistema de Información Estadística de Morbilidad: Son los procedimientos encaminados a la recolección, procesamiento y presentación de información sobre estadísticas de morbilidad.

6.3 Utilidad: Interviene en todas las etapas de la gerencia en salud. La información sobre morbilidad contribuye a identificar aspectos importantes relacionados con los diferentes componentes del estado de salud de la población.

Es de suma importancia en la planificación, ejecución y evaluación de programas e intervenciones de salud.

En la planificación:

a) Permiten conocer las afecciones que aquejan a la población.

b) Permiten conocer el riesgo de enfermar.

c) Permiten conocer la gravedad de las afecciones.

En la ejecución:

La recolección activa de la información de morbilidad permite detectar los eventos que aparecen y tomar las medidas pertinentes para controlar o erradicar enfermedades y prevenir las que están a su alrededor.

Las estadísticas de morbilidad permiten conocer la carga de enfermedad presente en la población. La carga de enfermedad es un concepto incorporado por la Organización Mundial de la Salud a partir de cual se estiman no solo acciones curativas y preventivas sino también la calidad de vida, la esperanza de vida, el redimensionamiento de los servicios y sistemas de salud.

En la evaluación y proceso de toma de decisiones:

A través de la medición del cumplimiento e impacto de los programas se puede garantizar la ganancia de la salud. El proceso de toma de decisiones se hace más eficiente y efectivo con el uso de buenas estadísticas de morbilidad.

En la investigación:

En la investigación epidemiológica, sirve para determinar el modo de transmisión de la enfermedad, el periodo de incubación, la severidad de la infección y los aspectos inmunológicos así como para conocer el cuadro epidemiológico de un territorio. Son utilizadas también en otros tipos de investigación en áreas clínicas, ensayos biológicos, pruebas de medicamentos y vacunas, por citar algunos ejemplos.

En la docencia:

Son utilizadas sistemáticamente en la docencia de pregrado y postgrado.

En general, las estadísticas de morbilidad son útiles para:

a) Conocer el número de personas que sufren de una enfermedad en particular, con qué frecuencia y en cuánto tiempo.

b) La demanda que hacen esas enfermedades sobre los recursos médicos y que pérdidas financieras causan.

c) Fatalidad y gravedad de las enfermedades.

d) Si las medidas de prevención son eficaces.

e) Distribución de las enfermedades según edad, sexo, ocupación, etc. y comportamiento en el tiempo.

f) Relación entre el control de la enfermedad y la atención médica brindada.

6.4 Dificultades más importantes en el estudio de la morbilidad.

El estudio de la morbilidad y su medición por las estadísticas sanitarias, es un proceso complejo dada la propia complejidad de la morbilidad. La morbilidad es un fenómeno dinámico y con una carga importante de subjetividad. Por ejemplo: Una persona puede estar enferma y no percatarse de ello, por tanto no

demandara atención medica y permanecerá oculto a diferentes fines, entre ellos, el estadístico. Una persona puede estar enferma de varias patologías a la vez o poseer un diagnostico presuntivo que cambia o se corrobora al realizarse estudios específicos y he ahí una gran dificultad para la medición del evento.

Aspectos conceptuales a considerar en el estudio y medición de la morbilidad:

1) Proceso salud-enfermedad: Es un proceso dialéctico, no existe una línea divisoria claramente definida entre salud y enfermedad y es difícil determinar lo que cada persona considera como equilibrio biosicosocial.

2) Anormalidad y normalidad: Un individuo puede tener salud, es decir gozar de equilibrio biosicosocial y estar enfermo de algo que no le molesta.

Hay enfermedades que permiten establecer medidas para definir lo normal de lo anormal por ejemplo, se puede considerar como hipertensión cuando la tensión arterial está por encima de las cifras establecidas, sin embargo, una persona puede sentirse bien con una tensión arterial elevada.

3) Problemas del Diagnóstico: Se puede considerar que un individuo está enfermo atendiendo a:

a) Opinión del paciente de estar enfermo.

b) Que sea diagnosticado a través de examen físico del médico.

c) Que sea diagnosticado a través de pruebas complementarias.

4) Enfermos o Enfermedades: Un individuo puede estar enfermo de diarrea y asma al mismo tiempo, hay que definir si se recoge el número de enfermedades que en este caso son dos o de enfermos que es uno, es decir definir que medir: enfermos o enfermedades.

Si se recogen las enfermedades hay que tomar en consideración el momento de su comienzo, de esta forma puede ser de tres maneras distintas:

- La enfermedad comienza antes del periodo de recogida de datos y termina dentro de este.
- La enfermedad comienza y termina dentro del periodo de recogida de la información.
- La enfermedad comienza dentro del periodo de recogida de la información y termina después.

De acuerdo a la decisión que en cuanto a estos criterios se tomen determinaran que se medirá y conocerá realmente: casos nuevos o total de casos.

5) Frecuencia e Incidencia. Esto se relaciona con la problemática antes descrita a fin de establecer que tipo de medición se hará.

Las formas clásicas de medición en Morbilidad se basan en los siguientes dos conceptos:

Incidencia: Casos nuevos de una enfermedad en un periodo y lugar dado.

Prevalencia: Total de casos de una enfermedad en un periodo y lugar dado. Incluye los casos nuevos como los que ya existían antes del periodo de estudio y permanecen enfermos.

6) Pesquisaje o demanda: Hay que definir si se registran solamente los pacientes que solicitan atención médica en los servicios de salud o se hace una búsqueda activa de la enfermedad como es el caso de los programas de salud que implican entre sus acciones el pesquisaje masivo.

7) Consultas o Reconsultas: Es necesario definir si solo se registraran las consultas donde se hizo el diagnóstico o todas las que se hagan durante el transcurso de una enfermedad.

6.5 Fuentes de Información

Para obtener información sobre morbilidad rara vez resulta suficiente el uso de una sola fuente de información, lo más usual es la consulta de mas de una de las existentes. Por ejemplo las salidas del sistema de información de defunciones y defunciones perinatales que informan sobre las causas de muerte, si bien aportan conocimiento sobre morbilidad, es sobre la morbilidad más severa, la que termina generalmente con un desenlace fatal. Algo similar ocurre cuando se utiliza como fuente de información las salidas del sistema de egresos hospitalarios ya que se accede a la morbilidad que requiere hospitalización, también por lo general mas severa. Al consultar la fuente información que constituyen las salidas del sistema de enfermedades de declaración obligatoria, conocemos también parcialmente la morbilidad que aqueja a la población ya que los datos se refieren fundamentalmente a las enfermedades transmisibles.

Se enumeran a continuación las fuentes habituales de información de morbilidad:

MORBILIDAD GENERAL:

Mortalidad General
Diagnósticos de egresos hospitalarios
Diagnóstico de consultas ambulatorias
Exámenes masivos a la población
Enfermedades Transmisibles
Enfermedades Dispensarizadas
Otras sujetas a registros especiales (cáncer, tuberculosis)
Registros de enfermedades sujetas a pesquisaje.

MORBILIDAD ESPECÍFICA RELATIVA A PARTE O SEGMENTOS ESPECIALES DE LA POBLACIÓN:

Grupos de Edad:

- Menores de un año, menores de 5 años, mayores de 5 años, tercera edad, adolescentes y otros.
- Morbilidad perinatal, morbilidad de embarazadas, de adultos y de otros grupos de edad.
- Mortalidad y Morbilidad laboral y escolar.
- Morbilidad según sexo, escolaridad, zona de residencia, características geográficas y socio económicas.

6.6 Medición de la Morbilidad.

La medición de la morbilidad se realiza utilizando los indicadores de uso mas frecuente en la actividad de las estadísticas continuas:

Números absolutos
Proporciones y porcentajes

Razones
Tasas

Algunas tasas propias de la morbilidad

$$\text{Tasa de Incidencia} = \frac{\text{Número de nuevos casos de una enfermedad durante un periodo dado y para un lugar.}}{\text{Población en estudio}} \times 10^n$$

Expresa el riesgo de contraer una enfermedad en una población dada en un periodo de tiempo.

$$\text{Tasa de Prevalencia} = \frac{\text{Número total de casos de una enfermedad en un periodo dado y un lugar dado}}{\text{Población en estudio}} \times 10^n$$

Expresa el riesgo de padecer una enfermedad en una población dada en un periodo dado.
Una variante de la Tasa de Incidencia es la Tasa de Ataque que se mide cuando la población solo está expuesta durante un período limitado. Será de ataque primario cuando considera solo el numero de casos de inicio de un brote o epidemia.

$$\text{Tasa de Letalidad} = \frac{\text{Número de defunciones por una causa}}{\text{Número de enfermos por esa causa}} \times 100$$

Mide la severidad de la causa. Cuando se mide en la comunidad esta tasa sobre estima la severidad ya que por lo general el número de enfermos registrados siempre es menor que los existentes.

Otras tasas de morbilidad son:

$$\text{Tasa Bruta de Morbilidad} = \frac{\text{Número de enfermos por una causa dada región y periodo dado}}{\text{Total de población de esa región y periodo dado}} \times 10^n$$

Las Tasas Brutas de Morbilidad se pueden calcular para una causa en particular o para todas las causas. Son brutas o generales ya que el denominador incluye la población general.

Tasas Específicas de Morbilidad por sexo o edad:

Numerador: Número de enfermos de un grupo de edad o un sexo dado.
Denominador: Población del grupo de edad o sexo correspondiente al numerador.
Numerador y denominador tienen que ser de igual región, periodo y segmento poblacional.

6.7 SIE de Morbilidad en Cuba

En muchos de los sistemas de información estadísticos de salud, aparece incluido el registro, flujo y emisión de datos de morbilidad junto a otros tales como servicios o recursos. Existen sistemas de información que son dirigidos especialmente a este componente del estado de salud de la población o sea se han concebido casi exclusivamente para el conocimiento de la morbilidad en los diferentes niveles de atención.

Se enumeran y detallan, de estos últimos, los más relevantes:

1) SIE de Enfermedades de Declaración Obligatoria (EDO)

El registro primario es la tarjeta de EDO que se llena a los pacientes que padecen enfermedades que deben ser declaradas obligatoriamente por tratados internacionales o intereses del país. Esta tarjeta la llena el medico de asistencia con carácter obligatorio. El SIE de EDO se nutre además de los registros de laboratorios, historias clínicas, hojas de egreso, hojas de consulta externa. Ahí se pueden detectar EDO que podrían no haberse notificado por la tarjeta de EDO en la consulta externa o de urgencia. De esta forma el SIE gana en integridad y fiabilidad. Las tarjetas de EDO son utilizadas para confeccionar los modelos establecidos para el uso estadístico a nivel de unidad. Esta información pasa a los niveles inmediatos superiores tabulada y consolidada por edades, sexo y territorio. El sistema de EDO es un sistema de vigilancia epidemiológica y emite salidas de información de periodicidad tan corta como la semanal en los diferentes niveles del sistema nacional de salud. A partir de la notificación de las algo mas de 90 enfermedades sujetas a la notificación obligatoria del sistema de EDO, algunas son seleccionadas por situaciones de vigilancia especial o por alertas epidémicas y pasan a la emisión diaria a partir del sistema de información diario (SID), jerarquizado por el Instituto de Medicina Tropical Pedro Kouri (IPK).

2) Registro Nacional de Cáncer

Recoge todos los casos diagnosticados de cáncer. El registrador es el medico de asistencia. La máxima dirección del registro de cáncer se ubica en el INOR (Instituto Nacional de Oncología y Radio biología, Ciudad de La Habana). Es un registro de cierta complejidad en el que el medico de asistencia esta en la obligación de notificar los casos diagnosticados de cáncer y ciertas características biológicas de los pacientes así como ubicación anatómica y características morfológicas e histológicas del tumor. La información de este registro pasa a soporte electrónico a nivel de provincia en los departamentos de estadísticas, quienes envían los ficheros por vía electrónica al INOR para confeccionar la base de datos nacional. El Registro Nacional del Cáncer se nutre además de las salidas del sistema de información

estadístico de defunciones y defunciones perinatales. A partir de esta información se realizan importantes estudios sobre el cáncer.

3) SIE de Dispensarizados.

Se ocupa de la recolección, flujo, procesamiento y presentación de información sobre las personas que padecen de enfermedades sujetas a la dispensarización. Se notifican a nivel primario de atención, fundamentalmente por los médicos de la familia, y abarca en la actualidad las siguientes entidades: Diabetes Mellitus, Hipertensión Arterial, Asma Bronquial, Enfermedad Isquémica del Corazón, Insuficiencia Renal Crónica, Accidente Vascular Encefálico e Hipercolesterolemia. Las salidas de este sistema se ofrecen con periodicidad anual.

4) Encuestas periódicas.

Se aplican encuestas periódicas nacionales utilizando las técnicas del muestreo para conocer factores de riesgo (La mas reciente fue aplicada en 1995 y en el transcurso de este ano 2000 se repetirá) . Se aplico la Encuesta de satisfacción y uso de los servicios de salud (ENSUSS) en 1998 la que entre otros, ofreció datos de morbilidad padecida en los últimos dos años.

5) SIE de Morbilidad Laboral.

Su registro primario es el certificado medico por invalidez temporal expedido a trabajadores. La máxima dirección de este sistema se ubica en el Instituto Nacional de Salud del Trabajador.

6) SIE de egresos hospitalarios.

Este sistema ha pasado por varias modificaciones en su concepción a través del tiempo. En sus inicios se captaban, fluían y eran tabuladas la totalidad de las hojas de egreso de todos los egresados de todos los hospitales del país. Las salidas de información eran sumamente complejas, voluminosas y detalladas lo que hacia que su utilización decayera. De ahí se concibió su funcionamiento más económico y eficiente, a partir del uso del muestreo aplicado a las unidades emisoras de información y se adecuaron las variables a medir y la lista de morbilidad que serviría para la comunicación estadística con un pensamiento más racional y económico. En la actualidad el dato primario se obtiene de la historia clínica cerrada por el medico al egreso del paciente e informan solo 35 hospitales del país seleccionados a partir de un diseño muestral que permite hacer representativa esta información, al país en general.

Perspectivas de desarrollo de los SIE de Morbilidad en Cuba.

1. Se perfecciona la calidad del dato en cuanto a cobertura, integridad y calidad.
2. Se trabaja en la automatización para mejorar la oportunidad y procesos de validación del dato.
3. Se encuentran en fase de concepción y diseño otros SIE para la morbilidad relacionada con la urgencia y la emergencia medica, las enfermedades crónicas no trasmisibles, la discapacidad y otros.

4. Se desarrolla la capacidad de aplicación de encuestas periódicas a población abierta, a fin de obtener información sobre enfermedades padecidas.

Actualmente las estadísticas de morbilidad toman en consideración los factores de riesgo incorporando así los enfoques actuales de la epidemiología y la importancia del trabajo preventivo. Las encuestas aplicadas a fin de conocer los factores de riesgo presentes en la población, fueron mencionadas en párrafos anteriores. Algunas reflexiones con respecto a estos son:

Los ***factores de riesgo*** pueden influir en la aparición de las ***enfermedades*** que pueden provocar ***deficiencias*** que se refieren a pérdida, malformación o anomalía de un órgano, estructura, o de una función mental, psicológica, fisiológica o anatómica lo que puede conllevar a la ***incapacidad*** que es la restricción o la falta de habilidad para llevar a cabo una actividad y mantenerla en los límites considerados como normales para un ser humano. Depende de la capacidad de cada persona de adaptarse a su deficiencia. El ***handicap***, por su parte es producido por una deficiencia o una incapacidad y el individuo no puede cumplir su role social según su edad o sexo, lo que representa las consecuencias sociales de su deficiencia o incapacidad.

La identificación de los factores de riesgo es de importancia para desarrollar medidas que previenen la enfermedad (prevención primaria), pero en el otro extremo de la enfermedad están sus consecuencias. La prevención para impedir dichas consecuencias es llamada prevención secundaria.

Ejercicios resueltos

1. A continuación se brinda información estadística de Lepra en Cuba durante 1999. Con la misma calcule:

a) Tasa de Incidencia
b) Tasa de Prevalencia
c) Interprete los resultados.

DATOS:

Número de casos nuevos de Lepra	333
Total de casos de Lepra	597
Total de población	11 142 691

a) Tasa de Incidencia de Lepra en Cuba durante 1 999:
333/ 11 142 691 * 100 000 = 3.0
El riesgo de contraer lepra en Cuba durante 1 999 fue de 3.0 por cada 100 000 habitantes.

b) Tasa de Prevalencia de Lepra en Cuba durante 1 999:
597/ 11 142 691 * 100 000 = 5.4
El riesgo de estar enfermo de lepra en Cuba durante 1 999 fue de 5.4 por cada 100 000 habitantes.

2. En un hospital se encontró que ingresaron en el pasado año un total de 234 pacientes por infarto agudo del miocardio y fallecieron por esa causa 94 personas. ¿Cuál fue la tasa de letalidad en esa institución de salud?

Tasa de Letalidad por Infarto Agudo del Miocardio en el hospital el pasado año:

94/234 * 100 = 40.0

Por cada 100 enfermos de Infarto Agudo del Miocardio fallecieron 40 lo cual expresa alta severidad de la enfermedad.

Ejercicios Propuestos

1. Un territorio dado está confeccionando el Análisis de la Situación de Salud para lo cual ha recopilado información de morbilidad que aparece a continuación:

Total de personas diagnosticadas por cáncer	32
Mujeres de 15 a 49 años que tienen cáncer cervicouterino	11
Enfermos por Diabetes Mellitus	45
Fallecidos por Diabetes Mellitus	3
Total de Población	12125
Casos vistos por enfermedades diarreicas agudas	269
Total de Dispensarizados	749
Dispensarizados por hipertensión arterial	258
Población femenina de 15 a 49 años	3056

Calcule:

a) Tasa de Letalidad por Diabetes Mellitus
b) Tasa de Incidencia de Cáncer
c) Tasa de Morbilidad por Diabetes Mellitus
d) Tasa de Incidencia por cáncer en las mujeres de 15 a 49 años de edad
e) Tasa de atención brindada por Enfermedades Diarreicas Agudas
f) Porcentaje de dispensarizados por Hipertensión Arterial

Interprete los resultados obtenidos en cada inciso.

7. Estadísticas de Recursos y Servicios:

7.1 Conceptos de estadísticas de Recursos y Servicios

Estadísticas de Recursos: Información numérica cuantificable acerca de los recursos con que cuenta el Sistema Nacional de Salud para cumplir sus propósitos y elevar el estado de salud de la población.

Los recursos pueden ser:

- Humanos.
- Medios de producción.
- Bienes Producidos.
- Equipos.
- Unidades de salud.
- Locales de consultas.
- Camas para ingresos.
- Sillones estomatologicos.
- Instituciones y locales para la formación de personal calificado de salud.
- Medios de transporte.
- Otros.

Estadísticas de Servicios: Es la información numérica cuantificable sobre los servicios de salud que se prestan con los recursos disponibles para mejorar el estado de salud de la población.

7.2 Utilidad de las estadísticas de recursos y servicios:

Las estadísticas de recursos y servicios son de gran utilidad en la administración científica durante todas sus etapas.

Planificación: Según los principales problemas de salud se deben planificar los recursos y servicios necesarios para su solución.

En la ejecución se mide la eficacia y eficiencia de los medios en el proceso.

Para el análisis del costo beneficio de los servicios prestados con los recursos existentes.

7.3 Medición de Recursos y Servicios.

Se utilizan los indicadores habituales en la actividad de las estadísticas continuas:

Números absolutos
Razones
Proporciones y porcentajes
Tasas

Los indicadores de recursos se agrupan usualmente de la siguiente manera:

a) Indicadores que miden los recursos existentes.

Este grupo de indicadores tienen como propósito expresar el volumen de recursos existentes.

Ejemplo:

- Número de médicos.
- Médicos por habitantes.
- Camas por médicos.

Puede relacionarse el volumen de recursos existentes con la población beneficiada, o con los servicios prestados o relacionar recursos entre sí, Ej. Médicos por cama hospitalaria, enfermeras por médicos, estomatólogos por sillón.

b) Indicadores que miden el aprovechamiento de los recursos.

Este grupo de indicadores tiene como propósito expresar el uso que de los recursos se hace.

Ejemplo:

- Consultas por médicos, este indicador es usado para medir la productividad médica.
- Aprovechamiento de los salones de operaciones. Puede relacionarse el total de horas que estuvo ocupado el salón entre el total de horas planificadas o del total de salones, cuantos están funcionando y que tiempo.
- Aprovechamiento de la cama hospitalaria. Por su importancia y complejidad serán tratados mas adelante con profundidad.

Con frecuencia se relaciona información de recursos y de servicios a fin de obtener mayor expresividad en los indicadores. De hecho es conveniente incorporar a la practica del análisis de estos indicadores la relación entre recursos y servicios prestados, como única vía de medir la eficiencia del proceso de gestión así como percatarse oportunamente de las necesidades de cada comunidad para ajustar los recursos en función de los servicios cada vez que sea necesario.

<u>Indicadores que miden el aprovechamiento de la cama hospitalaria</u>

La cama hospitalaria es un recurso de importancia y guía para el quehacer en el nivel de atención secundario y terciario. A partir de este recurso se establecen plantillas, presupuestos y demás recursos para la actividad hospitalaria que actualmente gasta mas del 60 por ciento del presupuesto del sector salud.

Es de gran importancia para la administración de salud, medir el aprovechamiento de la cama a fin de un desempeño gerencial exitoso y un aprovechamiento optimo del alto presupuesto que la atención

hospitalaria lleva implícito. Es de suma importancia para evaluar la calidad y cantidad de servicios de hospitalización se ofrecen a la población.

Los indicadores de aprovechamiento de la cama hospitalaria llevan implícitos definiciones básicas para su construcción y calculo. Estas son:

Dotación normal de camas: Son las camas que dispone el hospital en un periodo de tiempo.

Cama real: Es aquella cama que se encuentra realmente instalada y dispuesta las 24 horas del día para recibir una persona, este ocupada o no. Se excluyen las camas de Cuerpo de Guardia, trabajo de parto, cuartos de emergencia o cuidados especiales, observación, reconocimiento, banco de sangre, investigaciones radiológicas o endoscopicas, intervenciones menores, recién nacidos normales y las camas del personal que reside en el hospital.

Egreso: Paciente que habiendo ocupado una cama real del hospital la abandona ya sea vivo o fallecido.

Día Cama: Es la disponibilidad de una cama real por 24 horas del día.

Días Cama: Suma de los día cama de un día.

Día paciente: Es la unidad de servicio prestado por el hospital a un paciente ingresado en un día. O sea es el periodo de servicio prestado a un paciente hospitalizado comprendido entre las horas de censo de dos censos consecutivos, siendo contado el día de alta solamente en el caso de que un paciente ingrese y egrese en el mismo día.

Los censos se realizan cada día en un hospital, generalmente en horas de la noche. Debe ser a una hora fija. En este se enumeran las camas ocupadas o no y los pacientes existentes. El censo o ronda nocturna del hospital es una actividad básica del movimiento hospitalario de una unidad.

Días paciente: Es la suma de los día paciente en un día. Un hospital habrá prestado en un día cualquiera, tantos días paciente como pacientes permanezcan ingresados a las doce de la noche o a la hora del censo, mas un día paciente por cada persona que ingresa y egresa en el mismo día.

Días Estadía: Es el tiempo transcurrido desde el momento que el paciente ingresa hasta que realiza su egreso. La unidad de tiempo que se utiliza es el día. Si el paciente permanece ingresado solo unas horas, tendrá a los efectos de la medición, un día de estadía. La estadía, al abandonar el paciente el hospital, se considerara el día de egreso o el de ingreso, nunca ambos.

Los indicadores de uso más frecuente en la medición del aprovechamiento del recurso cama son:

Promedio de Estadía (PE): Es el promedio de días de asistencia hospitalaria recibida por cada paciente en un periodo de tiempo dado. Mide cuantos días en promedio está hospitalizado un paciente desde su ingreso hasta su egreso.

El PE es el indicador más importante de la utilización de las camas. Es el más consistente y de mayor poder discriminatorio.

Intervalo de Sustitución (IS): Es el tiempo promedio que una cama permanece desocupada entre el egreso de un paciente y el ingreso de otro. Cuando un hospital tiene un IS de un día, se encuentra en dificultades para dar servicio a la comunidad.

Intervalo de Rotación (IR): Mide el numero de pacientes que en promedio rotan por una cama en un período determinado de tiempo.

Índice de Ocupación o Índice Ocupacional (IO): Establece la relación entre los pacientes ingresados y la capacidad real de las camas de un servicio u hospital. Es el indicador de menor poder discriminatorio. Nunca debe ser analizado aisladamente de los otros.

Promedio de Ocupación de Camas al Año (POCA): Número promedio de días que una cama se mantiene ocupada al año.

No es aconsejable analizar aisladamente cada uno de estos indicadores, la interpretación integral de los mismos es capaz de brindar una información mucho más clara del aprovechamiento de la cama hospitalaria.

El aprovechamiento optimo de la cama se reflejaría a partir de estos indicadores y de modo general de la siguiente manera: Promedio de Estadía corto, intervalo de sustitución de dos o más días, índice de rotación elevado y un también elevado índice ocupacional. Cada hospital o servicio según su tipo o nivel, tiene patrones de comportamiento particulares de estos indicadores. Por ejemplo, por lo general los hospitales gineco-obstétricos son de corta estadía, alto índice ocupacional y de rotación. Al descender la natalidad por ejemplo o incorporarse tecnologías medicas modernas que acortan el periodo de hospitalización, los indicadores muestran menor aprovechamiento de las camas y se impone redimensionar el servicio u hospital. El aumento de la capacidad resolutiva de las estructuras de la Atención Primaria de Salud debe acortar también los periodos de hospitalización en todo tipo de hospital, fundamentalmente en los Clínico Quirúrgicos y Generales. Los hospitales psiquiátricos tienen por lo general, larga estadía, alto índice ocupacional y bajo índice de rotación. Obviamente en esto influye la morbilidad que atienden.

Existe más de una fórmula para el cálculo de cada uno de estos indicadores según si se trata de unidades o servicios de corta o larga estadía. A continuación aparecen las más utilizadas.

$$P.E = \frac{\text{Días pacientes}}{\text{egresos}} \qquad P.E. = \frac{\text{Días estadía}}{\text{egresos}}$$

I.O = Días pacientes
Días camas

I.O = P.E x IR.
Promedio de camas reales x Días del periodo

I.S = Camas desocupadas
egresos

IS = % de desocupación x PE
IDO.

I.R = Egresos
Promedio de camas reales

IR = Días del periodo
PE (mas) IS

Los indicadores de servicio se agrupan de la siguiente manera:

a) Indicadores que expresan la cantidad de servicios prestados.
Ejemplos:

- Número de consultas
- Consultas por habitantes. Relaciona servicios con población beneficiaria.

b) Indicadores que expresan la calidad de los servicios prestados.
Ejemplos:

Mortalidad Bruta = Número de defunciones hospitalarias x 100
Total de egresos

Mortalidad Neta = Defunciones hospitalarias de más de 48 horas de hospitalización x 100
Total de egresos

Mortalidad anestésica = Defunciones x anestesia x 100
Total de casos anestesiados

7.4 SIE de Recursos y SIE de Servicios.

Comprenden los subsistemas que se ocupan de la recolección, procesamiento y presentación de datos sobre los medios y los servicios que se prestan en el Sistema Nacional de Salud para alcanzar sus objetivos.

Los sistemas de información estadística de recursos son jerarquizados en gran medida por el SIEN por su relación con los costos, financiamiento, y la planificación de interés gubernamental. Otros se incorporan al SIEC de salud. Los SIE de servicios se integran en su mayoría al SIEC de salud.

Ejemplos de SIE de recursos y servicios del SIEC de salud:

Movimiento Hospitalario: Es uno de los sistemas más voluminosos del SIEC de salud ya que comprende la recolección, procesamiento y presentación de indicadores que expresan las acciones que se realizan al nivel hospitalario de atención. Emite información con periodicidades de trimestre, semestre y ano.

Registro de Profesionales de la Salud: Este sistema se ocupa del registro de todo el personal profesional de salud en funciones laborales. Se basa en las plantillas de trabajadores de las unidades y departamentos de recursos humanos. Emite información con periodicidad anual.

Consulta Externa: El SIE de servicios externos como también se conoce, es un SIE voluminoso ya que abarca la actividad de la atención primaria de salud y la ambulatoria al nivel que esta se produzca. Los indicadores de salida son fundamentalmente de volumen de servicios prestados y algunos relativos a la calidad de los servicios prestados en ciertas áreas especiales de la actividad de consulta externa.

Transporte Sanitario: Abarca información de transporte de pacientes tales como ambulancias, su estado y aprovechamiento.

Servicios de Ópticas: Este SIE se especializa en la actividad de óptica en su vertiente de servicios prestados, en que tiempo y con qué frecuencia. Ofrece indicadores que miden fallas o éxitos en la gestión en general.

Actividad Quirúrgica: Este sistema abarca una de las actividades más importantes, complejas y costosas del sector salud, la que se realiza casi en su totalidad en el ámbito hospitalario. Ofrece indicadores de volumen de operaciones realizadas según tipo y especialidad. Este SIE posee definiciones para la medición del dato primario de cierta complejidad y dinamismo, dado este ultimo atributo y por el rápido desarrollo de las técnicas quirúrgicas que se han incorporado al desempeño de la cirugía para el bienestar del paciente y de la eficiencia en este campo.

Ejercicios Propuestos

1. Un territorio está realizando el Análisis de la Situación de Salud y tiene la información siguiente:

Número de médicos	360
Número de enfermeras	410
Total de Consultas Estomatológicas.	25805
Total de población	25639
Número de Estomatólogos	72
Obturaciones caídas	21

Diseñe y haga el cálculo de un indicador de:

a) Volumen de servicios de consultas estomatológicas prestados por habitantes.
b) Volumen de recursos humanos por habitantes.

c) Calidad de los servicios estomatológicos.
d) Aprovechamiento de los recursos humanos (Estomatólogos)
e)Interprete los resultados

a) Volumen de Servicios Prestados

Consultas por Habitantes:
25805/ 25639 = 1.0
Se hizo 1 consulta por habitante.

b) Volumen de recursos
Enfermeras por médicos:
410/360 = 1.1
Hay aproximadamente 1 enfermera por cada médico

c) Calidad de los servicios prestados
Obturaciones caídas por estomatólogos:
21/72 * 100 = 29.2
Hay 29.2 obturaciones caídas por cada 100 estomatólogos

d) Aprovechamiento de los recursos
Productividad médica (Estomatólogos).
25805/360 * 100 = 71.7
Cada estomatólogo realizó aproximadamente 72 consultas en el periodo analizado

2. Interprete los indicadores de aprovechamiento de la cama de Cuba. 1999

a) Promedio de Estadía	9.6
b) Índice Ocupacional	69.3
c) Índice de Rotación	26.2
d) Intervalo de Sustitución	4.3

En Cuba, durante 1999 el aprovechamiento de la cama hospitalaria fue como sigue:

a) Cada paciente estuvo ingresado en promedio 10 días.
b) El 69.3 % de las camas estuvo ocupada.
c) Rotaron en promedio por una cama hospitalaria 69.3 pacientes.

d) Cada cama permaneció vacía aproximadamente 4 días.

Ejercicios Propuestos:

Construya dos indicadores de cantidad o volumen de servicios prestados, calidad de servicios prestados, volumen de recursos existentes y aprovechamiento de recursos. Interprete los resultados.

Utilice como fuente de información los registros estadísticos de la institución donde trabaja o el Anuario Estadístico del Ministerio de Salud Publica.

8 Estadística de Vivienda y Saneamiento

8.1 Conceptos de Estadísticas de Vivienda y Saneamiento

Estadísticas de Vivienda: Información numérica sobre la cantidad y calidad de las los lugares donde residen las personas. O sea es la información numérica de las viviendas donde el hombre habita y sus características.

Estadísticas de Saneamiento: Información numérica sobre las características ambientales y sanitarias del medio así como las actividades de control de saneamiento que se realizan en el seno de la población.

8.2 Sistemas de Información Estadísticas de Viviendas y Saneamiento

La información de vivienda: Se obtiene a través de los censos. Se realizan encuestas periódicas para la actualización de la información.

Los datos que se recogen usualmente son:

Tipo de vivienda
Abasto de agua
Características del servicio sanitario
Tipo de piso, paredes y techos. Tipología constructiva.
Estado constructivo de la vivienda.
Luz y ventilación
Hacinamiento. Numero de habitaciones para dormir y numero de personas que los usan. Numero totales de habitaciones de la vivienda y personas que viven en la vivienda.
Otros.

SIE de Saneamiento: Se integran al SIEC, se les denomina Subsistema de Información Estadística del Cuadro Higiénico y abarcan los procedimientos relacionados con la recolección, flujo, procesamiento y

presentación de información sobre agua, disposición y recogida de desechos sólidos, de desechos líquidos y contaminación ambiental de los lugares donde el hombre realiza sus actividades cotidianas (fábricas, escuelas, zonas de residencia).

Ejercicios Propuestos:

1. Explore en la institución donde trabaja, la información que se utiliza para analizar la situación de la vivienda. Busque cinco indicadores e interprételos.
2. Explore en la institución donde trabaja la información que se utiliza para analizar la situación del saneamiento ambiental. Busque cinco indicadores e interprételos.

Bibliografía Consultada:

1. Castañeda Abascal I y col. Indicadores más utilizados para medir la mortalidad. Monografía. La Habana. Facultad de Salud Pública, 1995.

2. Silva Ayzcaguer LC. Cultura estadística e investigación científica en el campo de la salud. Madrid. Díaz de Santos, 1997

3. Castañeda Abascal I y Gran Álvarez M. Generalidades de Estadísticas de Salud para la Maestría de Sicología de la Salud. Monografía. La Habana. Facultad de Salud Pública, 1998.

4. Anuario Estadístico 1999. Cuba. Dirección Nacional de Estadísticas del MINSAP. 2000.

5. Gran Alvarez M. Calidad de la Información Estadística. Aspectos Conceptuales. Indicadores de Salud Publica. Selección de Artículos. Facultad de Salud Publica. ISCM La Habana. 1987.

6. Indicadores de Salud Publica. Selección de Artículos. Facultad de Salud Publica. ISCM La Habana. 1987.

7. Clasificación Estadística Internacional de Enfermedades y Problemas relacionados con la Salud. Volumen I, II y III. Publicación científica No. 554. OPS / OMS Washington, DC. EUA. 1995.

8. Sistemas de Información y Tecnología de Información en Salud. Desafíos y Soluciones para América Latina y el Caribe. Programa de Sistemas de Información sobre Servicios de Salud. División de Desarrollo de Sistemas y Servicios de Salud. OPS. OMS. Washington, D.C. Abril 1998.

9. Jaspers Faijer, D. Evolución Futura de la Mortalidad. Tendencias de la Mortalidad por sexo y edad en América Latina, 1950 – 1995. CELADE. Chile. 1995.

10.La Salud Pública en Cuba. Hechos y Cifras. Dirección Nacional de Estadística. MINSAP. 1999.

11.Guidelines for monitoring the availability and use of obstetric services. UNICEF, UNFPA, OMS. New York. October, 1997.

12.. Family – building and Family Planning Evaluation. Department of Economic and Social Affairs Population Division. United Nations. New York. 1997

13. Informe sobre la salud en el mundo. La vida en el siglo XXI. Una perspectiva para todos. Organización Mundial de la Salud. Ginebra. 1998.

Ministerio de Salud Pública
Escuela de Salud Pública

Breve Introducción al Análisis Demográfico

Autores: Prof. Lic. Lorenzo I. Herrera León
Master en Estadística Matemáticas
Profesor Auxiliar. Investigador Auxiliar
Egresado del Centro Latinoamericano de Demografía (CELADE)

Dra. Isabel M. Barroso Utra
Especialista de primer grado en Bioestadística

Ciudad de La Habana
1999

Indice

1. Introducción.

El conocimiento de la población, entiéndase ésta como la que corresponde a un país o a una comunidad, y de sus características esenciales es vital para cualquier propósito en el cual esté involucrada ésta.

> *Existe un sinnúmero de actividades y acciones que tienen por objeto a la población, como por ejemplo las diferentes producciones de bienes y los diversos servicios públicos, entre los que se cuentan los relacionados con la salud pública, la educación, el empleo, las actividades recreativas y culturales, entre otos. A su vez, estos servicios son proporcionados por individuos que son parte integrante de esa población, es decir, por médicos, maestros y profesores, etc. Como se ve, la población tiene un doble carácter: productor y consumidor de bienes y servicios.*

El hecho de conocer los mecanismos de crecimiento, cambio y en general los determinantes de la dinámica poblacional proveen a gobernantes, políticos y a aquellos que deben tomar grandes decisiones, de herramientas muy útiles e información estratégica para la planificación y para la certera conducción del país. El caso del sector de la Salud Pública no escapa de esta óptica, ya que el mismo brinda a la población innumerables y valiosos servicios que es necesario planificar. Aún más, es imperativo conocer el resultado de estas acciones.

¿Cómo saber a qué velocidad crece la población de nuestro país?. ¿Cuáles son las enfermedades por las que mueren con más frecuencia las personas?. ¿Cómo se mide el impacto de éstas sobre la población?. ¿Cómo se producen los nacimientos, los matrimonios y divorcios, etc?. Las personas cambian de trabajo , los niños y jóvenes estudiantes promueven a diferentes grados en la enseñanza. Las personas se trasladan de una región a otra. Enferman y, curan o mueren. Las enfermedades y los accidentes producen secuelas y discapacidades. ¿Cómo manejar esta trama de eventos y situaciones que acontecen?. ¿Cómo saber la velocidad de transmisión de una enfermedad infecciosa?. ¿Cómo saber que los servicios prestados, las acciones de salud, etc, están surtiendo el efecto esperado?. Para la determinación del estado de salud de una población es indispensable responder a la mayoría de estas interrogantes, al igual que para evaluar la efectividad de las acciones de salud, como campañas de vacunación, de educación y promoción, y para una buena conducción del análisis de la situación de salud de una comunidad.

A tenor con lo dicho y como un fiel reflejo de una realidad que se impone vienen con un acento oportuno las palabras de Lotka, ese gran biólogo polaco, padre de la Demografía matemática:

"... . Las condiciones que se presentan en una población concreta son siempre excesivamente complicadas. Aquel que no haya captado claramente las relaciones necesarias entre las características de una población teórica sujeta a hipótesis simples, no sabrá desenvolverse con las relaciones mucho más complicadas que existen en una población real*." (1)*

La Demografía es una disciplina que puede ayudarle a comprender y resolver muchos de los problemas e interrogantes planteados arriba. Etimológicamente se deriva del griego ***Demos*** *que significa pueblo y* ***Grafía****, descripción; es decir, el estudio o descripción de la población.*

El presente capítulo pretende ser una mínima introducción al problema del estudio de la población, y brindar al estudiante los elementos básicos para el manejo de las complejas relaciones que se verifican en el seno de una población.

1.1 Algunos conceptos básicos

Población: Es una colección de objetos o individuos de la misma clase. Por ejemplo, podemos hablar de la población de libros en una biblioteca; la población de peces en un océano; la población humana en una ciudad, etc. En el seno de una población se producen una serie de procesos como, el aumento o disminución de sus miembros, cambios internos en algún atributo de varios de sus miembros por ejemplo, disminución de los libros de color rojo y aumento de los azules en la biblioteca, así como otros.

Las poblaciones humanas, no sólo son una colección pasiva de individuos sino que las mismas están conformadas por grupos entre los cuales se establecen relaciones y leyes. Además, ese colectivo humano está por lo general establecido en un área geográfica determinada (a excepción de los pueblos nómadas), es decir, que constituye un asentamiento.

En este curso nos dedicaremos a estudiar las poblaciones humanas; es decir, sus características más importantes y la forma de medir y realizar ciertos análisis en las mismas. Existen tres fenómenos fundamentales, también llamados variables demográficas, que producen cambios cuantitativos en la población: mortalidad, fecundidad y migraciones. Existen otras variables que también provocan cambios, pero nos concentraremos en las tres mencionadas.

Mortalidad: Se refiere a las defunciones como componente del cambio poblacional. La ocurrencia de la muerte de una persona es un hecho individual y puede ser catalogada como algo casual, mientras que al estudiarse la totalidad de las muertes que se producen en una población, nos percatamos de que existen diferencias según la edad, el sexo, el nivel educacional, ingreso, etc. Es decir, la proporción en que ocurren las muertes en una edad joven, es diferente a la que se observa en una edad adulta; así mismo, las proporciones de defunciones en cada sexo, observadas en un año, son distintas.

La mortalidad no es sólo un hecho biológico, sino un fenómeno con un gran componente socio-económico.

Natalidad: Se refiere a los nacimientos como componente del cambio de la población (el comentario anterior hecho en el caso de la mortalidad es válido aquí). Cuando se estudian los nacimientos de una población para un período de tiempo dado, se detectan diferencias por grupos sociales, nivel de ingreso, edad al casarse, disponibilidad y uso de anticonceptivos, nivel educacional, desarrollo económico, etc.

Fecundidad: .Es la capacidad reproductiva (real), de hombres, mujeres o parejas de una población. Es un concepto distinto al de fertilidad, que se refiere a la capacidad potencial (fisiológica) de producir un nacido vivo. Es oportuno señalar que el término reproductivo no se refiere a todos los nacimientos, sino sólo a aquellos cuyo resultado es un nacido vivo. La natalidad es consecuencia de la fecundidad y otros factores como la distribución de la población por edades.

Migración: Se refiere al movimiento de personas a través de una división política (frontera) para establecer una nueva residencia permanente. Se divide en internacional (movimiento entre países) e interna (entre regiones de un país). A los migrantes se les llama inmigrantes (inmigración) con respecto al país destino y emigrantes (emigración) con respecto al país origen.

1.2 Tasas brutas de mortalidad y natalidad. La ecuación compensadora.

Las tasas brutas o crudas miden la frecuencia de un fenómeno dado en la población, en un período de tiempo que generalmente se toma como un año.

La tasa bruta de mortalidad (TBM) se expresa como el cociente del total de defunciones y la población a mitad de año, multiplicada por 1000.

$$TBM = \left(\frac{D}{\overline{N}} \right) * 1000$$

donde:

D representa el total de defunciones

$\overline{N}$ la población total a mitad de período.

La tasa bruta de natalidad es igual a:

$$TBN = \left(\frac{B}{\overline{N}} \right) * 1000$$

donde:

B representa el total de nacidos vivos del período.

Cualquier otra tasa bruta se calcula de manera similar. Se suelen llamar tasas brutas porque contienen en el denominador la población total y debido a ello, el efecto de la composición por edad de la población está confundido con la medición de la frecuencia del fenómeno. Más adelante se verá esta cuestión.

Existe una relación importante entre la población en dos momentos. Si partimos del total de personas en un instante y le agregamos los nacidos vivos de un período, le sustraemos las defunciones de dicho período, le sumamos y restamos las entradas y salidas de personas, respectivamente, llegamos al total de población al final de ese período. Esto puede ser expresado mediante una fórmula, conocida con el nombre de ecuación compensadora.

$N^t = N^0 + B^{(0,t)} - D^{(0,t)} + I^{(0,,t)} - E^{(0,,t)}$

donde:

0 y t representan los instantes de inicio y final del período,

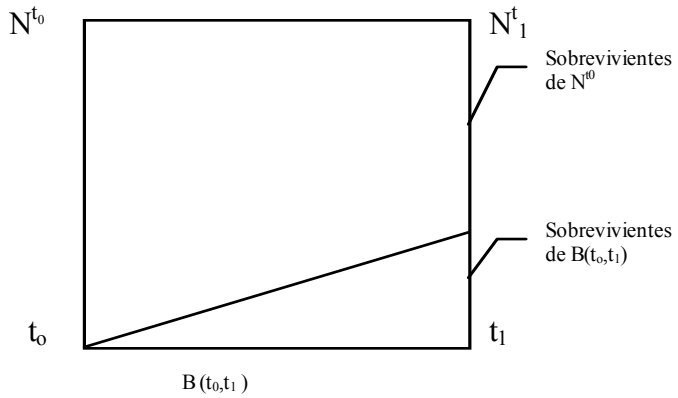
N, B, I, E, denotan población, nacimientos, inmigrantes y emigrantes, respectivamente.

En caso de una población cerrada a la migración (no hay entradas ni salidas), la ecuación se reduce al saldo de nacidos vivos y muertes, es decir:

$N^t = N^0 + B^{(0,t)} - D^{(0,t)}$

Entre los múltiples usos de esta ecuación se tienen:

- actualización de la población, partiendo de un monto inicial y los registros vitales y de migración, de un período.
- cálculo retrospectivo del monto poblacional para un instante en el pasado, partiendo de la población del presente y los registros vitales y de migración.
- cálculo de los nacidos vivos o las defunciones, de un intervalo de tiempo, partiendo de dos censos de población y el registro de migración.
- se puede evaluar la calidad de dos censos partiendo del supuesto que la calidad de los registros es buena o viceversa.



El esquema precedente muestra como se produce el cambio en una población cerrada (sin migración). La población en el momento t es el resultado de la sobrevivencia de los habitantes del momento inicial y de los nacidos vivos del período.

1.3 Distribución de la Población.

Se conoce como distribución de la población, la clasificación o agrupación de ésta según las categorías de una o más variables. Si tomamos la variable nivel de escolaridad, por ejemplo, efectivos de ésta se agruparían en: primaria, media, superior y universitaria. Según la zona de residencia, tenemos: urbana y rural. También la clasificación puede operarse con el sexo, en este caso masculino y femenino; la edad, menor de un año, 1-4 años, 5-9 años, etc.

A la distribución por edad y sexo de la población, se le denomina también, *composición* por edad y sexo de la población o estructura de la población por sexo y edades. La edad y el sexo son por excelencia variables básicas en el estudio de una población. Ellas brindan una descripción con mucho detalle de cualquier fenómeno demográfico que acontece en su seno. La proporción de varones y mujeres en cada grupo de edad, tiene vital importancia a la hora de trabajar con muchos indicadores y realizar comparaciones, así como en la magnitud de su crecimiento.

Muchas poblaciones poseen una alta concentración de personas jóvenes, lo cual es indicativo de que esos países posean una tasa de natalidad elevada, por lo común son países subdesarrollados. Por el contrario, los países con alto desarrollo socioeconómico, exhiben una proporción de personas adultas y ancianas mucho mayor, consecuencia directa de una natalidad muy baja a través de varias décadas.

Al proceso gradual en el que la proporción de adultos y ancianos aumenta en una población, mientras disminuye la de niños y jóvenes, se le denomina envejecimiento de la población. Los criterios para clasificar a las poblaciones según el grado de envejecimiento, parten de definir la edad de comienzo. Por ejemplo, el siguiente corresponde a un criterio establecido por las Naciones Unidas en 1989, toma para el análisis, la población de 65 años en adelante.

Categoría	% de pob. de 65 y más años
muy envejecida	16 % y más
envejecida	13% y menos de 16%
envejecimiento avanzado	10% y menos de 13%

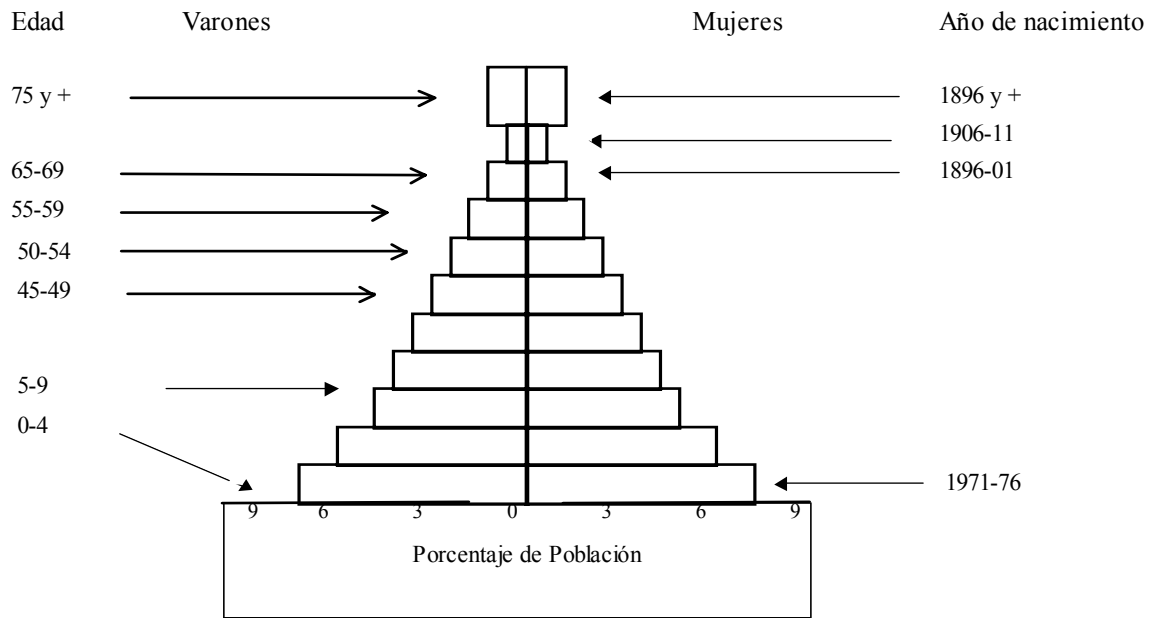
Incipiente	7% y menos de 10%
población madura	4% y menos de 7%
población joven	menos de 4%

En dependencia de la estructura por edad, los patrones de morbi-mortalidad serán diferentes. En una población muy joven prevalecerán los problemas de salud típicos de niños y adolescentes y las causas de muerte estarán acorde a dichos problemas. Así mismo, la demanda de servicios de salud estará matizada por el predominio de la pediatría.

Por otra parte, en una población envejecida, nos enfrentamos a entidades de corte crónico-degenerativo, en su mayoría no transmisibles como las enfermedades del corazón, los tumores malignos, senilidad, etc. Las especialidades médicas como la geriatría adquieren gran relevancia.

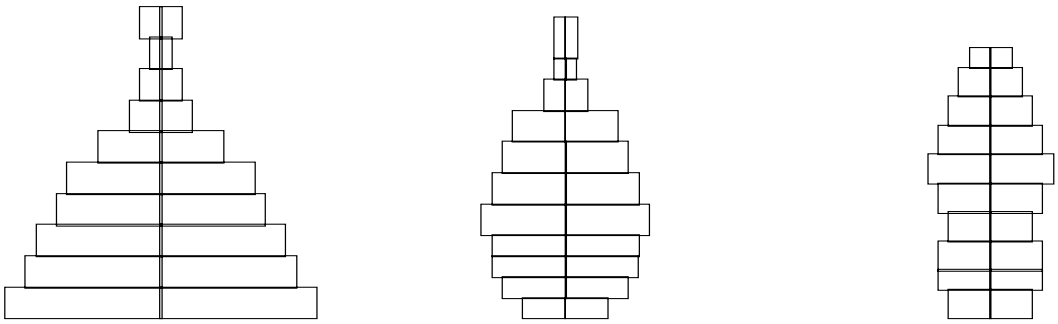
1.4 Pirámide de población

Existe una forma gráfica muy ilustrativa para representar la distribución por edad y sexo de una población, conocida como pirámide de población, ya que se asemeja notablemente a esta figura geométrica. Es un histograma doble (uno para el sexo masculino y otro para el femenino), donde se representa el número o proporción de varones y mujeres en cada grupo de edades.



A ambos lados del gráfico, se inscriben leyendas para señalar la edad y el año de nacimiento de los individuos, en cada grupo de edad, que en este ejemplo se han asumido quinquenales y referidos al mes de septiembre de 1976.

Perfiles generales de las poblaciones



Pirámide Expansiva

Pirámide Constrictiva

Pirámide Casi Estacionaria

Las poblaciones de los países tienden a agruparse en tres perfiles generales, de acuerdo a su composición por edades.

El perfil expansivo muestra una base ancha, que denota una alta natalidad, característico de poblaciones jóvenes. El constrictivo, exhibe menor proporción de jóvenes y denota un grado avanzado de envejecimiento. Por último, el modelo casi estacionario corresponde a países con un gran envejecimiento y un ritmo de crecimiento prácticamente nulo. Como ejemplos podrían citarse países como Guatemala, México en el primero; Estados Unidos, Canadá, en el segundo; Suecia y en general, los países nórdicos en el último grupo.

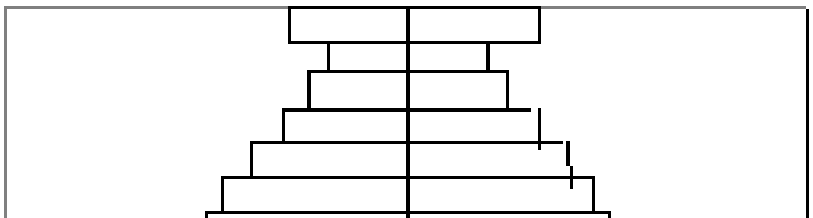
La pirámide ofrece gran información sobre la población, tanto presente como pasada; las huellas de acontecimientos importantes quedan impresas en ella. A través del análisis de este gráfico, podemos darnos cuenta, por ejemplo, de cuándo comenzó el descenso de la natalidad en un país; si en épocas pasadas existió una fuerte corriente migratoria a favor de un sexo determinado; qué proporción representa la población en edades laborales, etc.

1.5 Indice de masculinidad

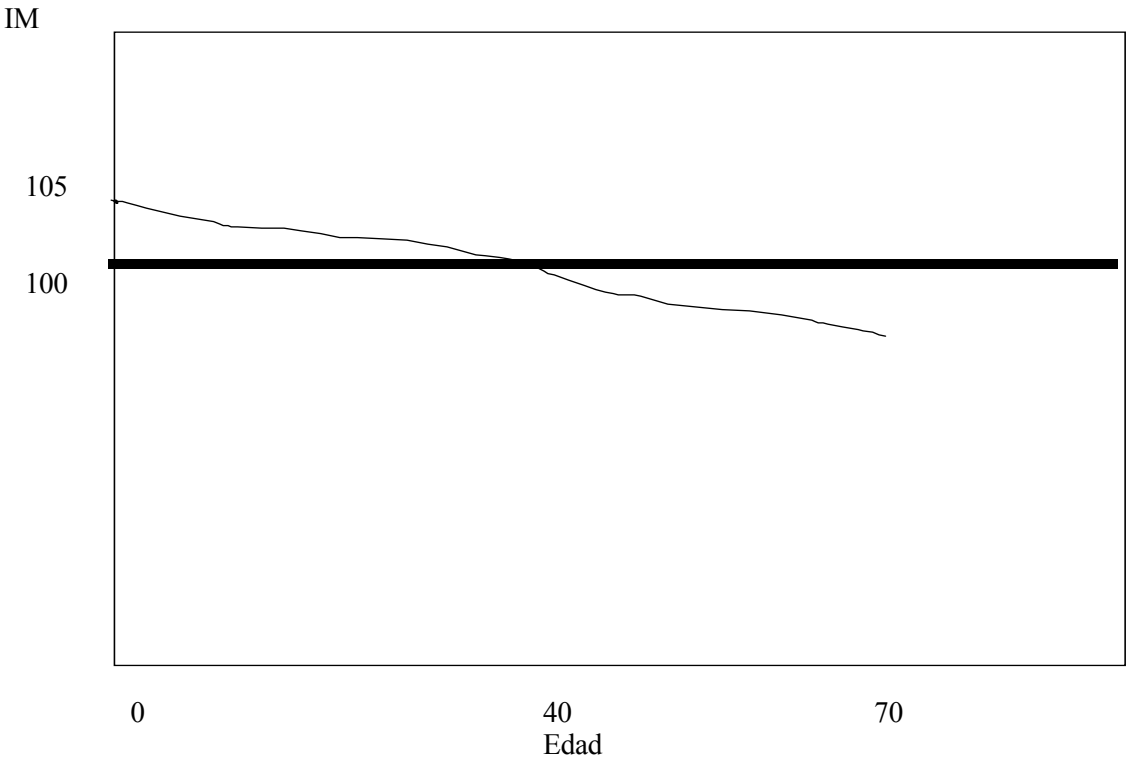
El índice de masculinidad, conocido también como relación de masculinidad, es la relación entre varones y mujeres en una población, que de ordinario se expresa por 100 mujeres. Su cálculo puede realizarse con los totales generales de cada sexo y también por edades. Al nacimiento, dicha relación posee un valor próximo a 105, debido a que nacen más varones, luego, por el efecto de la mortalidad, el índice se torna decreciente y llega a estar por debajo de 100 alrededor de los 40 años.

La gráfica siguiente muestra un comportamiento ideal del índice, toda vez que en general las migraciones distorsionan la curva, haciendo incluso cambios en su apariencia.

Pirámide de población, Cuba, 1995



Indice de masculinidad por edades.



Razón de Dependencia: es la relación entre las personas en edades dependientes (menores de 15 y mayores de 64 años) y las personas económicamente productivas (15-64 años), en una población, expresada por 100. Hay que tomar en cuenta, que no necesariamente las edades productivas son las mismas para todos los países, e incluso, muchas veces existen personas que trabajan aún siendo niños, lo que escapa de la óptica de este indicador. A continuación, se brinda un cuadro con información sobre la población de Cuba el 30 de junio de 1996.

Grupo de edades	Población
< 1 año	149641
1-4	640521
5-14	1636709
15-49	6078092
50-64	1500837
65 y +	1000066
15-64	7578929
<15 y 65 +	3426937
Total	11005866

Razón de dependencia para Cuba en 1996 = (3426937/7578929)*100= 45.21.

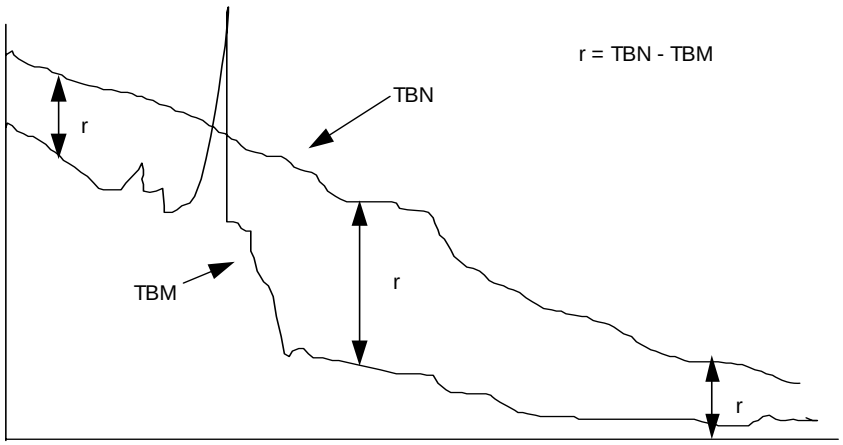
En nuestro país, cada cien personas en edades laborales, deben mantener a alrededor de 45 individuos que están fuera del segmento productivo. Esto nos da una medida de la carga o presión que ejerce la población no productiva sobre la productiva.

Recomendamos tomar con cautela cualquier análisis hecho en base a este indicador, sólo perseguimos ilustrar su uso. Téngase presente que es una cuestión muy peliaguda la definición de la población productiva y no productiva.

1.6 Transición Demográfica

Es el proceso a través del cual, la humanidad ha pasado de altas tasas de natalidad y mortalidad (crecimiento lento) a bajas tasas de natalidad y mortalidad (crecimiento lento, nuevamente); donde el descenso de la mortalidad, de ordinario ha precedido el de la natalidad. En una etapa intermedia, la tasa de crecimiento ha sido mayor, lo que produjo un aumento de la población no observado hasta entonces.

Tasas brutas de natalidad, mortalidad y crecimiento.



Años

En el gráfico precedente se ha esquematizado la transición demográfica: en los inicios, ambas tasas eran muy elevadas, con valores cercanos entre sí, lo que resultó en una tasa de crecimiento (diferencia entre TBN y TBM) pequeña; posteriormente, la mortalidad lideró el descenso cuando aún la natalidad permanecía elevada, produciéndose un exceso notable de los nacimientos sobre las muertes y por ende una aceleración del incremento. Al final, las dos tasas brutas han descendido a valores más pequeños y también lo ha hecho su ritmo de crecimiento.

Aunque aparentemente se llega a una situación de similitud, los dos momentos extremos presentan diferencias cualitativas importantes, sobre todo en lo concerniente al patrón de morbi-mortalidad: de un predominio de entidades de corte transmisible, se desemboca a un panorama en el cual el hombre se enfrenta a las enfermedades crónico- degenerativas, aparejado con la posesión de una alta tecnología sanitaria.

Ejemplo:

Con la información de la siguiente tabla, correspondiente a la columna encabezada con el título Población, construya la pirámide de población para Cuba, correspondiente al 31 de diciembre del año 1998. Calcule el índice de masculinidad por edad y grafíquelo.

Población				Porcentaje	
Edad	Masculino	Femenino	Total	Masculino	Femenino
0-4	379627	355506	735133	3,41	3,19
5-9	433163	408556	841719	3,89	3,67
10-14	437101	415207	852308	3,92	3,73
15-19	360676	345942	706618	3,24	3,11
20-24	417428	406935	824363	3,75	3,65
25-29	552049	544909	1096958	4,96	4,89
30-34	561823	563399	1125222	5,04	5,06
35-39	471100	478369	949469	4,23	4,29
40-44	327808	337374	665182	2,94	3,03
45-49	338568	347729	686297	3,04	3,12
50-54	302149	317931	620080	2,71	2,85
55-59	258367	259351	517718	2,32	2,33
60-64	210982	217736	428718	1,89	1,95
65y+	521863	568227	1090090	4,68	5,10
Total	5572704	5567171	11139875	50,02	49,98

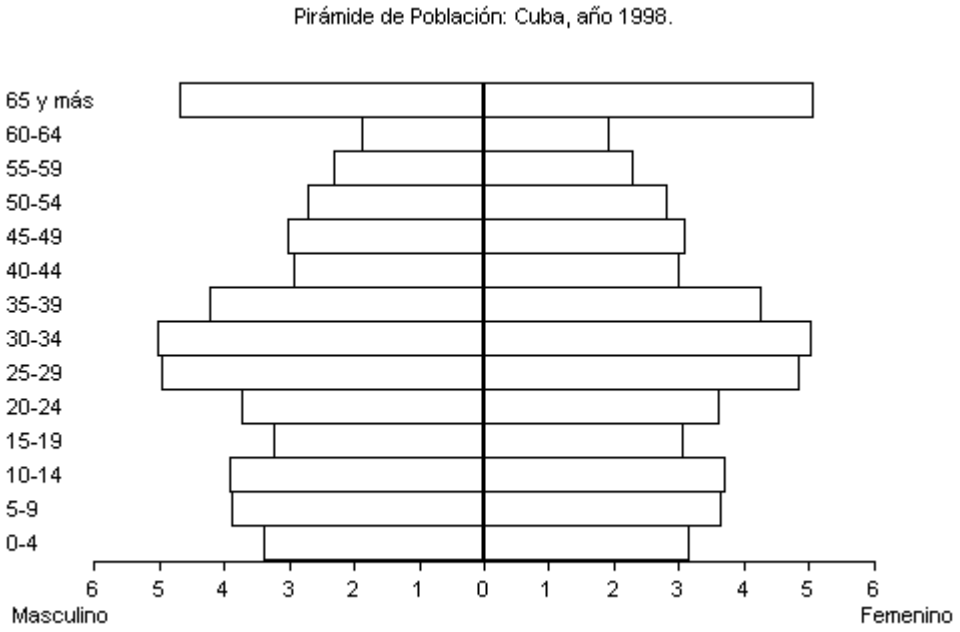
La pirámide que se va a construir es del tipo que toma a la población total (todas las edades y los dos sexos reunidos) como el ciento por ciento (100%).

El algoritmo a seguir consiste en calcular, para cada grupo de edades dentro de cada sexo, el porcentaje de población que representa con respecto al total general. Por ejemplo, el porcentaje de población que corresponde al grupo de 0-4 años del sexo masculino sería,

(379627 / 11139875) * 100 = 3.41%.

Las dos últimas columnas de la tabla muestran los valores correspondientes para cada combinación de edades y sexo.

Ahora se procede a construir el gráfico, o sea la pirámide propiamente dicha. Para ello se construyen dos histogramas o gráficos de frecuencia, uno para cada sexo, tomando la frecuencia igual a los valores que correspondan de las dos últimas columnas de la tabla anterior y como clases los intervalos de edades. Luego los dos gráficos se juntan de manera que a la izquierda aparezca el sexo masculino y a la derecha el femenino (podría ser al revés); el eje vertical común a ambos representa las clases o intervalos de edades; el eje horizontal exhibe los valores de la frecuencia de cada sexo, (ver gráficos en el texto).

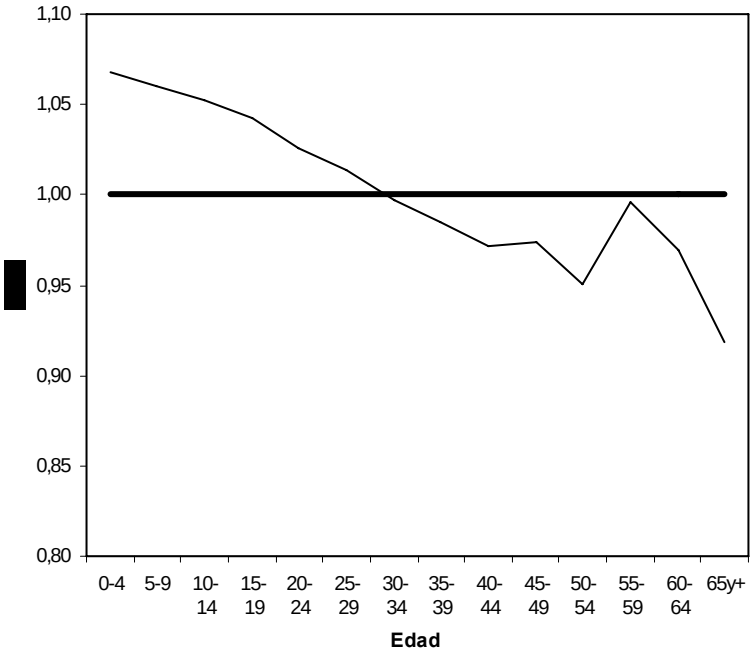


Para el cálculo del índice o razón de masculinidad, sencillamente dividimos la cantidad de hombres entre la cantidad de mujeres, para cada grupo de edades y luego construimos un gráfico de línea.

Edad	Masculino	Femenino	R de M
0-4	379627	355506	1.07
5-9	433163	408556	1.06
10-14	437101	415207	1.05
15-19	360676	345942	1.04
20-24	417428	406935	1.03
25-29	552049	544909	1.01
30-34	561823	563399	1.00
35-39	471100	478369	0.98
40-44	327808	337374	0.97

45-49	338568	347729	0.97
50-54	302149	317931	0.95
55-59	258367	259351	1.00
60-64	210982	217736	0.97
65y+	521863	568227	0.92
Total	5572704	5567171	1.00

Razón de masculinidad. Cuba, año 1998.



La línea trazada por el eje vertical con valor 1 se inscribe para referencia, por ejemplo, cuando la curva está sobre ésta, entonces los varones superan a las mujeres y viceversa.

Resumen del capítulo

El estudiante debe fijar los conceptos tratados toda vez que este capítulo es básico para los que vienen luego.

Ejercicio propuesto:

a) Diga, a su juicio, la importancia que para usted tiene la transición demográfica.

b) Por qué es importante conocer la distribución de la población por algunas variables como edad, sexo, zona de residencia, etc.

2. Representación en el tiempo de hechos demográficos.

2.1 El diagrama de Lexis

Para la representación de hechos demográficos, se utiliza un diagrama o esquema muy sencillo, conocido como diagrama de Lexis. Está conformado en el primer cuadrante del plano, el cual se fracciona en pequeñas cuadrículas. En el eje horizontal se inscribe el tiempo calendario y en el vertical la edad. Existen varios conceptos asociados a este diagrama como son:

- generación o cohorte
- línea de vida
- edad cumplida
- edad exacta.

Generación o cohorte: Es un grupo de personas que comparten una experiencia demográfica simultáneamente, el cual se observa durante un cierto tiempo. Las personas nacidas en determinado año, así como los individuos que contrajeron matrimonio en el año 1995, son ejemplos de cohortes.

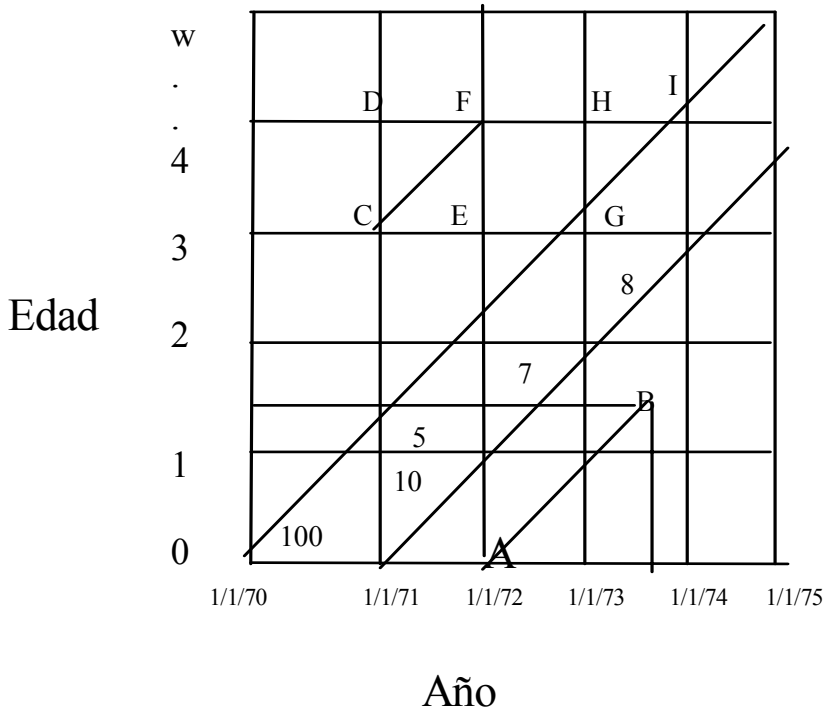
Línea de vida: La vida de una persona se representa a través de una línea oblicua (45^0 con la horizontal), que parte de las coordenadas {fecha de nacimiento, edad exacta cero año}, del individuo y termina cuando éste muere, es decir, en el punto {fecha de muerte, edad al morir}.

Edad Cumplida: En Demografía, cuando hablamos de edad cumplida, nos referimos a la edad del último cumpleaños. Una persona nacida el día 30 de junio de 1970, tendrá 0 año cumplido hasta su próximo cumpleaños el día 30 de junio de 1971, ya que suponemos el momento de su nacimiento como el cumpleaños número cero. En 1975, si aún continúa con vida, celebrará su quinto cumpleaños, a partir de donde su edad cumplida será 5 años hasta su próximo aniversario, en 1976.

Edad Exacta: Se refiere a la edad de una persona con la mayor exactitud posible. Aquí, a diferencia del concepto de edad cumplida, sí nos interesa la precisión, con día, mes y año. La edad cumplida es la edad exacta, a la cual se le trunca las fracciones de año. El niño del comentario anterior, antes de celebrar su primer cumpleaños, pasará por diferentes edades exactas: 1 mes, 2 meses y tres días; 6 meses y 2 semanas, etc.

El gráfico siguiente, muestra la constitución del diagrama. El segmento AB corresponde a la línea de vida de una persona que nació el 1/1/72 y murió aproximadamente en 30/6/73 con edad cumplida 1 año. Su edad exacta en ese momento era de 1 año y 6 meses. Ahora bien, como la escala de edades que estamos utilizando es de un año, sólo contaremos las edades exacta en 0, 1, 2, 3, etc. Es bueno apreciar que la edad cumplida y exacta es la misma edad de la persona, lo único que sucede es que la primera procede de la segunda, cuando a ésta se le quitan las fracciones de año. Un individuo siempre tiene una edad cumplida y una exacta.

Diagrama de Lexis



Sin embargo, al hacer una representación en el diagrama, la población de edad cumplida se inscribe en los segmentos verticales y la población de edad exacta, en segmentos horizontales. Veámoslo con un ejemplo: los segmentos CD, EF, HG, representan la población de edad cumplida 3 años en los momentos 1/1/71, 1/1/72 y 1/1/73, respectivamente. Por otra parte, los segmentos CE y EG están asociados a la población de edad exacta 3 años, en los años calendarios 1971 y 1972. De igual manera, los segmentos DF y FH, representan la población de edad exacta 4 años en 1971 y 1972.

Algunas propiedades

1. La población de edad cumplida corresponde a una fecha exacta (día, mes, año) en el eje del calendario, mientras que para el eje vertical es un intervalo. Está formada por un conjunto de líneas de vida que cortan a un segmento vertical en el mismo instante.
2. La población de edad exacta es, con respecto al eje del tiempo, un intervalo y, al eje de edades, un instante. Las líneas de vida que cruzan el segmento horizontal, lo hacen en diferentes momentos.
3. Cuando una línea de vida cruza un segmento vertical, cambia el calendario; es decir, se pasa al otro año.
4. Cuando una línea de vida cruza un segmento horizontal, se celebra un cumpleaños.
5. La población de edad cumplida se puede obtener de un censo o de un registro.
6. La población de edad exacta no se registra, por lo que hay que estimarla.

En el diagrama también se ha simbolizado una cohorte, la de los nacidos vivos del año 1970. Como se aprecia, dada la imposibilidad de trazar todas las líneas de vida, sólo se muestran la inicial y la final. Dentro de esa franja se supone que están contenidas las restantes.

Las áreas o superficies se utilizan para inscribir las defunciones. Por ejemplo, el cuadrado CDFE denota las defunciones ocurridas en el año 1971 de personas con edad cumplida 3 años. Dicha superficie está dividida por una línea discontinua, en dos triángulos, uno superior y otro inferior. Esto significa que las defunciones representadas corresponden a dos generaciones sucesivas: una nacida en 1967 y otra en 1968.

Podemos actualizar la cohorte si conocemos las defunciones que ocurren en cada edad y por supuesto, los nacidos vivos que le dieron origen. Veamos el siguiente

Ejemplo:

Suponga que en la cohorte de 1970 señalada en el diagrama, se tienen 100 nacidos vivos; entre las edades exactas 0 y 1 mueren 10 individuos; entre 1 y 2, fallecen 12 personas; 5 en 1971 y 7 en 1972. En 1973 mueren 8 individuos con edad cumplida 2 años. Deseamos calcular la población de edad exacta 3 años.

La población de edad exacta 1 año será igual a 100- 10= 90; la población de edad exacta 2 años es 90- 12= 78. Finalmente, la población de edad exacta 3 años es 78 - 8 = 70.

Con este proceder se puede llegar hasta la total extinción de la cohorte, que ocurre a una edad que hemos denotado w y de ella en adelante no existen sobrevivientes. Puede ser que no se conozca su valor con exactitud, pero de lo que sí estamos seguros es que existe. Algunos demógrafos la han situado en 110 años, otros en 120, como límite biológico de la vida humana.

Por último, es oportuno señalar, que aunque el diagrama presentado exhibe una escala simple en ambos ejes, también puede usarse otra cualquiera como quinquenal, mensual, de días, etc., sólo hay que tener presente que sea la misma en los dos ejes, transversal y longitudinal

Existen dos formas básicas de análisis: transversal y longitudinal. El primero se distingue por el hecho de que no hay variación en el calendario y sí puede haberla en el eje edad. Cuando realizamos un estudio de la mortalidad por edad, para un año dado o grupo de años, estamos en presencia de un estudio transversal o de momento. Por otra parte, el estudio longitudinal se destaca porque hay variación tanto en el eje horizontal como en el vertical. El seguimiento de una o varias cohortes, durante determinado tiempo, hace que necesariamente, haya cambios en el calendario y en la edad; éste es un estudio longitudinal.

2.2 Notación

Existen varias maneras de denotar los distintos componentes de la población. Comenzaremos por la población de edad cumplida, para ella se utiliza la letra mayúscula N. Si queremos expresar la correspondiente a un grupo de edad de determinada amplitud n, que comienza en una edad exacta x,

escribimos ${}_nN_x$; esto nos refiere a la población que tiene edades cumplidas entre las edades exactas x y x+n. En el caso que n =1, sólo escribimos N_x. Si quisiéramos agregar un símbolo para población media, procederíamos así:

${}_n\overline{\mathbf{N}}_x$

donde la barra indica población media. Pueden agregarse algunos superíndices para el sexo, año y demás. La población media de un año se ubica el día 30 de junio o primero de julio del año en cuestión.
La población de edad exacta se denota con E_x y también se pueden agregar súper índices para sexo, año, etc.

Ejemplo:

E_5^{75} denota la población de edad exacta 5 años, en 1975.

${}_5N_5^{t}$ denota la población de edad cumplida entre las edades exactas 5 y 10 años, en la fecha t.

${}_3N_2^{1/1/75}$ denota la población de edad cumplida entre las edades exactas 2 y 5 años, el 1/1/75.

Los nacidos vivos se escriben con la letra B (del inglés birth). La totalidad de los nacidos vivos del año 1973 se escribe como B^{1973}.

Para las defunciones, se utiliza una notación parecida a la de la población de edad cumplida. Ahora le toca ocupar su lugar a la letra D (del inglés death). Pongamos de ejemplo las muertes ocurridas en el año 1975, de personas con edad cumplida entre las exactas 5 y 10 años: ${}_5D_5^{1975}$.

Existe otra notación muy utilizada, conocida como censal. Es a la que más estamos acostumbrados y se utiliza para población de edad cumplida. Aparece en las tablas que sobre población hay en los anuarios y censos. La población de edad cumplida entre las edades exactas 5 y 10, por ejemplo, que escribimos como ${}_5N_5$, ahora la denotamos por $N_{5\text{-}9}$. En general, la población de edad cumplida comprendida entre las edades exactas x y x+n, es ahora $N_{x,\ x+n-1}$. El problema es que las edades exactas limitan al grupo de población de edad cumplida, no incluyéndose la edad exacta límite superior en el grupo, con la notación anterior.

Para hacer más clara la explicación, digamos que cuando escribo ${}_5N_5$, incluyo la población de las edades simples 5, 6, 7, 8 y 9, pero no la de 10 años, pues esta es el límite superior del grupo de edad. Por otra parte, si escribo $N_{5\text{-}9}$, entonces incluyo las poblaciones de 5, 6, 7, 8 y 9; considerando que la edad 9 años llega casi hasta 10, pero no la incluye. Lo más importante es saber que ambas notaciones son equivalentes: la demográfica y la censal.

Resumen del capítulo

Es importante tener clara las diferentes definiciones y conceptos.

Línea de vida: comienza con el nacimiento de la persona y termina con la muerte de ésta.

Edad exacta es aquella que se tiene exactamente en el momento del cumpleaños y se representa por un segmento horizontal en el diagrama de Lexis.

La edad cumplida es la misma edad exacta mientras no se llegue al próximo cumpleaños. Se representa por un segmento vertical en el diagrama de Lexis.

Ejercicio propuesto:

Construya un diagrama de Lexis donde represente la siguiente información

a) población de edad cumplida 3 años para los años calendarios 1970 hasta 1973.

b) Represente la población de edad exacta 0 año y 1 año para 1971.

Utilizando la notación demográfica represente a

a) población de edad cumplida 11 años

b) población de edades cumplidas entre las edades exactas 7 y 13 anos.

3. Cálculo de tasas

En esta sección trataremos de dar una visión algo más rigurosa de lo que es una tasa.

Dado un determinado evento demográfico (nacimiento, muerte, etc.) que puede ser experimentado por los individuos de una población, en un intervalo de tiempo (t_0, t_1), de duración (t_1-t_0) años o fracción de año, calculamos las siguientes cantidades:

$$\text{Incremento Absoluto} = \Delta^{(t\,o,\,t\,1)} = H^{(t\,0,\,t\,1)} = H^{t\,1} - H^{t\,o}$$

donde H^t denota la cantidad acumulada de individuos que han experimentado el evento en cuestión hasta un momento t dado.

El incremento absoluto es la frecuencia absoluta del evento en el intervalo de tiempo; es decir, la cantidad de personas que han experimentado el evento en el intervalo de tiempo.

$$\text{Incremento Medio Anual} = \frac{\text{Incremento Absoluto}}{\text{Longitud intervalo}}$$

El incremento Medio Anual nos dice cuánto corresponde de incremento (en promedio) a cada año del intervalo para el cual se desea calcular la tasa. Finalmente, la tasa anual se tiene mediante el cociente del Incremento Medio Anual y el valor de la población correspondiente a un momento dentro del intervalo. Esto es:

$$\text{Tasa del evento en el período } (t_0, t_1) = \frac{\text{Inc.Medio.Anual}}{N^{t0}}$$ (multiplicado por una potencia de 10, que puede ser : 100, 1000, 10000, etc).

N^{t0} representa la población al inicio del período; pudo haberse tomado la final del intervalo $N^{t1,}$ o $\overline{N}$, es decir, la población media del intervalo.

Al dividir el Incremento Medio Anual por un valor de referencia, en este caso N^{t0}, estamos relativizando dicho incremento; es decir, lo hacemos comparable al de otro país con características de tamaño poblacional muy diferente.

La tasa así calculada se interpreta como qué fracción (o qué frecuencia, expresada en porcentaje, o por mil, etc, en dependencia de por qué potencia del número 10 se ha multiplicado la tasa) representa el Incremento Medio Anual con respecto al total de la población para el momento inicial t_0 (o para el momento final t_1 si se usa como referencia la población final N^{t1}, o con respecto a la población de mitad del intervalo si se ha usado la población media).

Comencemos ahora describiendo las principales tasas utilizadas en Demografía.

3.1 Tasa bruta de Natalidad.

Se definió como el cociente entre los nacidos vivos y la población media, para un año t dado.

$$TBN = \left(\frac{B^t}{\overline{N}} \right) * 1000$$

Según la definición anterior debemos calcular el incremento medio anual, pero por tratarse de un intervalo de un solo año, el mismo es:

Incremento Medio Anual = (Nacimientos acumulados hasta el final del año t - Nacimientos acumulados hasta inicio del año t)/1

$$= \text{(Nacimientos acaecidos durante el año t)}/1 = \frac{B^t}{1}$$

Si tomamos como población de referencia la de mitad de período $\overline{\mathbf{N}}$, entonces llegamos a la fórmula anterior. Sólo resta multiplicar por 1000. El caso de la tasa bruta de mortalidad es similar; el estudiante puede comprobarlo, siguiendo los pasos anteriores.

El hecho de dividir el incremento absoluto del período por la longitud del mismo (expresada en años o fracción de año), persigue calcular una tasa anual. De no hacerse así, lo que se obtiene es una tasa referida al período como un todo.

Ejemplo:

En el trienio 1994-1996 se produjeron en Cuba 434711 nacidos vivos; la población media de ese trienio fue 10978148 personas, que corresponde al 30 de junio de 1995. Con esta información se decidió calcular la tasa bruta de natalidad:

$$\text{Tasa de natalidad} = \frac{434711}{10978148} * 1000 = 39.60.$$

Es cierto que esa es una tasa bruta de natalidad, pero no es una tasa anual sino trienal. Para que sea una tasa anual debemos dividir los nacidos vivos del trienio por 3, esto es:

$$\text{Tasa bruta de natalidad (anual)} = \frac{\frac{434711}{3}}{10978148} * 1000 = 13.20.$$

Esta tasa es anual, pero calculada con información de un trienio. Es representativa de cada uno de los tres años, como promedio. Si multiplicamos este último resultado por 3, obtendremos el valor anterior de 39.60. Está implícito el supuesto de uniformidad en la distribución de los nacidos vivos en el período. Las tasas brutas de natalidad de cada uno de los tres años, calculadas con la información propia de cada año fueron: 13.44 en 1994, 13.41 en 1995 y 12.75 en 1996.

Cuando nos es dada una tasa de natalidad, digamos, de 13.2 como la anterior, asumimos que la población media se estratifica o fragmenta en una serie de grupos de 1000 personas cada uno, y por cada uno de estos grupos, se producen, en promedio, 13.2 nacidos vivos en el transcurso del año; esta es una forma fácil de interpretación; para cualquier tasa bruta es similar, por ejemplo, de mortalidad, incidencia, etc.

También la tasa bruta de mortalidad se pueden calcular para cada sexo por separado; para cada causa, etc, en esos casos la información a utilizar se refiere a un sexo dado o una causa dada.

3.2 Tasa de Crecimiento de la Población.

Siguiendo la tónica de definición de tasa, ahora atañe proceder a estimar el ritmo al que la población crece. Para ello es necesario, en primer lugar, conocer cual es el evento que experimentan los individuos en la población. Respecto a ello puede decirse que el crecimiento está dado por el saldo entre los nacimientos y las muertes y el resultado de la migración. Por tanto, ahora no existe un evento propiamente dicho el cual experimentan los individuos, sino varios eventos que dan como resultado el incremento de la población (nacimientos de algunos, muerte de otros y movimientos de entrada y salida de varios).

Se procede calculando el Incremento medio Anual experimentado por la población, en un período de tiempo dado (t_0, t_1).

Incremento Medio Anual = $\dfrac{N^{t1} - N^{to}}{t_1 - t_0} = \overline{\Delta N}(t_0, t_1)$

Ahora, en dependencia de la población de referencia que se tome, así será el modelo de crecimiento de la población:

- Población al inicio del período, N^{t_0}

$$r = \frac{\Delta \overline{N}_{(t_0, t_1)}}{N^{t_0}} \times 100 \qquad \text{(I)}$$

En este caso, el supuesto básico implícito, es que la población crece linealmente, esto es: la ecuación que describe el comportamiento de la población en función del tiempo es una recta de la forma

$$N^t = N^{t_0}(1 - rt_0) + (N^{t_0} r) \cdot t \qquad \text{(II)}$$

si t_0=0

$$N^t = N^0 + (N^0 \cdot r \cdot t)$$

donde el primer sumando después del signo = es el intercepto de la recta con el eje vertical y el sumando más a la derecha, es el producto de la pendiente y la variable independiente t.

Con la ecuación (I) calculamos el valor de la tasa anual de crecimiento de la población, con la (II) estimamos el monto de la población para un momento futuro t, conocida la tasa r. Debe usarse para períodos de tiempo no muy largos sino más bien cortos, de 1 ó 2 años.

Ejemplo:

La población de cierto país el 1/1/1994 fue de 10043164, y seis meses más tarde era de 10057200. Calcular la tasa de crecimiento anual

Para calcular la tasa anual de crecimiento, basándonos en información de un semestre, procedemos de la siguiente manera:

Incremento Medio Anual = (10057200 - 10043164) / 0.5 = 28072,

r = (28072/10043164) * 100 = 0.28

Si imaginariamente, agrupamos la población del 1/1/94, en muchos conjuntos de 100 personas cada uno, por cada uno de ellos se producirán 0.28 personas más, en el transcurso del año.

- Población de referencia, la de mitad de período, $\overline{\mathbf{N}}$. Al igual que el anterior partimos de:

En este caso la tasa queda como:

$$r = \frac{\Delta \overline{N}^{(t_0, t_1)}}{\overline{N}} * 100$$

La ley que rige el crecimiento de la población en función del tiempo la describe la siguiente ecuación:

$$N^t = N^{t_0} \frac{(1 + \frac{tr}{2})}{(1 - \frac{tr}{2})}$$

Ya esto no describe una recta sino otro tipo de curva. Al igual que el caso anterior, es recomendable para períodos cortos de 1 ó 2 años y dados un valor de r (sin multiplicar por 100) y otro de t, conocido N^{to}, puede calcularse el monto futuro de la población para la fecha t.

Tomemos los datos del ejemplo anterior y estimemos la población 3 meses después del 1/1/94, el 31/3/94.

$$N^{31/3/94} = 10043164 * \frac{(1 + \frac{0.25 * 0.0028}{2})}{(1 - \frac{0.25 * 0.0028}{2})} = 10050197$$

Hemos puesto r = 0.0028 y t = 0.25 años ya que 3 meses es la cuarta parte de un año.

Hagamos el pronóstico para dos años a partir de la fecha 1/1/94, usando la misma tasa de crecimiento:

$$N^{1/1/96} = 10043164 * \frac{(1 + \frac{2 * 0.0028}{2})}{(1 - \frac{2 * 0..28}{2})} = 10099564 .$$

- Crecimiento Geométrico (Interés Compuesto).

Se asume que la población crece de acuerdo a una ley exponencial dada por la ecuación:

$$N(t) = N(0) * (1 + r)^t$$

Con el uso de logaritmo y algunos procedimientos algebraicos, se despeja r

$$r = \sqrt[t]{\frac{N(t)}{N(0)}} - 1$$

Es decir, conocidos los valores inicial y final de la población y la longitud del período (en años), podemos calcular r. Nótese que está involucrada una raíz de orden t, que cubre sólo el cociente.

El supuesto básico en este caso es que el crecimiento porcentual es constante. Veamos esto.

Si partimos del momento cero, la población en el momento t =1 sería:

N(1) = N(0)*(1 + r), lo que implica que

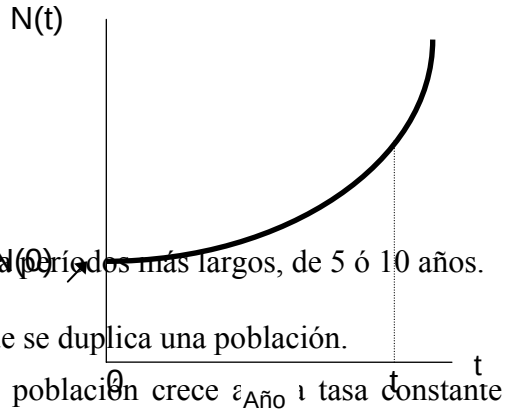
$\frac{N(1)}{N(0)} = 1 + r$.

Si t =2, entonces $\frac{N(2)}{N(1)} = 1 + r$

Si t =3, $\frac{N(3)}{N(2)} = 1 + r$.

Se observa que el cociente entre la población en determinado año y el precedente es constante e igual a 1+ r. Es decir, la población de un año representa el (1+r)100 % de la del año anterior.

Gráficamente, la forma que toma la función es como esta



Puede ser aplicada a períodos más largos, de 5 ó 10 años.

3.2.1 Tiempo en que se duplica una población.

Suponiendo que la población crece a una tasa constante r, se puede determinar el tiempo que demora la misma es duplicar, triplicar, etc., su tamaño.

En la fórmula $N(t) = N(0) * (1 + r)^t$, la población en el momento t debe ser igual al doble de la inicial N(0), por tanto

$2*N(0) = N(0)*(1+r)^t$

al despejar t, resulta que

$$t = \frac{\ln 2}{\ln(1 + r)}$$

Ejemplo:

Con la tasa del ejemplo anterior, r = 0.0028, estimaremos el tiempo que demora la población en duplicar su tamaño.

$$\mathbf{t} = \frac{\ln 2}{\ln(1 + 0.0028)} = 248 \text{ años.}$$

El crecimiento de una población puede descomponerse en tres vertientes fundamentales, a saber: crecimiento natural, mecánico y total o neto.

El crecimiento natural es el que se da sólo por el saldo entre los nacidos vivos y las defunciones. Mide la capacidad de crecimiento de la población en función de su natalidad y mortalidad.

El crecimiento mecánico esta asociado al saldo migratorio y el total o neto involucra a ambos.

De la ecuación compensadora conocemos que

$N^1 - N^0 = B^{(0,1)} - D^{(0,1)} + I^{(0,1)} - E^{(0,1)}$

para un intervalo de un año. Si dividimos ambos miembros de la ecuación por la población media del año, nos queda

$$\frac{N^1 - N^0}{\overline{N}} = \frac{B^{(0,1)}}{\overline{N}} - \frac{D^{(0,1)}}{\overline{N}} + \frac{I^{(0,1)}}{\overline{N}} - \frac{E^{(0,1)}}{\overline{N}}$$

lo que equivale a

r = (TBN - TBM)+(T inm - Temi)

la tasa de crecimiento total (r), es igual a la tasa bruta de natalidad menos la tasa bruta de mortalidad, más la tasa de inmigración (T inm) menos la tasa de emigración (T emi). A su vez, la diferencia entre la tasa de natalidad y mortalidad nos proporciona el ritmo de crecimiento natural y la diferencia entre las de inmigración y emigración, da la tasa de crecimiento mecánico.

TASA DE CRECIMIENTO NETO ó TOTAL = (TASA DE CRECIMIENTO NATURAL + TASA DE CRECIMIENTO MECANICO)

En Cuba, en el período comprendido entre el 30/6/ 1995 y 30/6/96, la tasa de crecimiento total (por 1000) fue 2.5, y la de crecimiento natural de 6.3 (TBN =13.4 y TBM = 7.1), lo que significa que el crecimiento mecánico es negativo (predominio de emigración).

3.3 Tasa bruta de mortalidad.

Al igual que se procedió para la tasa de natalidad, así es el cálculo de la tasa de mortalidad. Se obtiene primero el incremento total del período que se quiere analizar, luego el incremento medio anual y por último se divide por la población de referencia.

La tasa bruta de mortalidad puede calcularse para una provincia o zona, para un sexo dado, para una causa o grupo de causas dadas, simultáneamente para un sexo y una causa, etc, con tal de que se tome la información que corresponda, es decir, si es por ejemplo para el sexo femenino, debe tomarse sólo las defunciones de la población femenina y la población de referencia de este sexo.

Ejemplo:

En 1996, las defunciones totales de la provincia de Matanzas ascendieron a 5420 y la población media era de 630192 habitantes, por tanto su tasa bruta de mortalidad fue de $\frac{5420}{630192} = 0.0086$ ó 8.6 defunciones por mil habitantes.

3.3.1 Tasas por edad, sexo y causa.

Para el cálculo de las tasas por edades se sigue un procedimiento similar al explicado para las tasas brutas, lo único que referido al grupo de edades; es decir, se calcula el incremento medio anual y se divide entre la población de referencia, todo dentro del grupo. A continuación un

Ejemplo:

Las defunciones correspondientes al grupo de edades 1-4 en el año 1996 fueron 391 y la población media de dicho grupo en ese año de 640521. La tasa de mortalidad viene dada por

$\overline{\Delta D}(1/1/96,31/1/96) = \frac{391}{1} = 391$ (incremento medio anual de las defunciones del grupo 1-4)

En el ejemplo anterior se estima la tasa de mortalidad de ese grupo con información de un año calendario.

Tasa del grupo 1-4 = $\frac{391}{640521} = 0.00061$ ó 0.61 por mil habitantes.

Si se desea obtener la tasa del grupo 1-4 del sexo masculino, sólo se tomarían las muertes ocurridas en esas edades y de ese sexo, dividido por la población media de los varones de 1-4 años en 1996. Las tasas por edades se pueden combinar con un sexo determinado, una causa de muerte o grupo de causa, etc.

Las tasas donde participan todos los grupos de edades en su cálculo; es decir, aquellas que involucran a toda la población se les llama tasas brutas. Mientras que aquellas restringidas a un grupo de edad se les denomina específicas. Uno puede definir una tasa lo más específica que desee; por ejemplo, al calcular la tasa de mortalidad del grupo quinquenal de edades 5-9 años, hemos especificado la misma; pero si hacemos el cálculo por edades simples dentro del mismo grupo, tendremos una tasa para la edad 5 años, otra para 6, para 7 y así sucesivamente. Cada una de estas tasas es más específica que la correspondiente al grupo quinquenal como un todo, lo que quiere decir que ellas medirán mejor el riesgo de muerte y estarán mucho menos afectadas por la distribución por edades de la población. Muy pronto daremos una explicación más abundante del significado de esto.

3.3. 2 Notación**.**

Existe una notación para las tasas específicas de mortalidad. Recordemos que las defunciones de un grupo de edades se escribían como

${}_nD_x$, donde la x significa la edad exacta de comienzo del intervalo de edades y n la amplitud del mismo.

Para la tasa específica del mismo grupo pondremos

${}_nm_x = \frac{{}_n\mathbf{D_x}}{{}_n\overline{\mathbf{N}}_\mathbf{x}}$, es decir, el cociente de las defunciones del grupo y la población media. Pueden agregarse índices para el sexo, causa y demás.

Fácilmente se ve que

${}_nD_x = {}_n\overline{\mathbf{N}}_\mathbf{x} * {}_nm_x$. Las defunciones del grupo es el producto de la tasa correspondiente por la población media.

Por tanto, la tasa bruta de mortalidad se escribe entonces como

$$TBM = \frac{\sum_{0} {}_n\overline{N}_x * {}_nm_x}{\overline{N}} = \sum \left(\frac{{}_n\overline{N}_x}{\overline{N}}\right) * {}_nm_x$$

el factor entre paréntesis es la proporción de la población por edades (la pirámide que resulta de unir los dos sexos), luego, la tasa bruta es la suma de los productos de la proporción de población en cada grupo de edades y la tasa específica del grupo.

Esta fórmula es muy elocuente, si dos países o regiones tienen iguales tasas por edad

(por ende igual nivel de mortalidad) y sus respectivas distribuciones por edades

(pirámides) son diferentes, las tasas brutas resultantes serán también diferentes.

La comparación del nivel de la mortalidad entre dos o más países se ve afectada por la diferencia en estructura entre los países, enturbiando los resultados. Dicho en otras palabras, cuando se compara el nivel de la mortalidad entre países a través de la tasa bruta, las diferencias encontradas se deben no sólo a las diferencias puras en mortalidad, sino a una mezcla de mortalidad y estructura de población. Veamos un ejemplo práctico.

A continuación se brindan los valores de la proporción de población por edades de los países A y B. Suponemos que ambos países tienen idénticas tasas de mortalidad por grupos de edades.

Grupo de Edades	Tasas de mortalidad	Porcentaje de Población del país A	Porcentaje de población del país B
Menor de 1 año	31.0	1.43%	1.43%
1-4 años	0.5	6.06%	6.06%
5-14 años	0.4	14.98%	7.85%
15-49 años	1.5	56.31%	13.37%
50-64 años	8.4	13.37%	14.98%
65 y más	55.4	7.85%	56.31%
Total		100.00%	100.00%

Nota: La tasa de mortalidad está calculada por 1000 habitantes.

La tasa bruta de mortalidad del país A será:

TBM^A = (0.031 * 0.0143) + (0.0005 * 0.0606) + (0.0004 * 0.1498) + (0.0015 * 0.5631) + (0.0084 * 0.1337) + (0.0554 * 0.0785) = 0.00685 o 6.85 defunciones por mil habitantes.

La tasa total para el país B, se calcula de forma similar:

TBM^B = (31.0 * 1.43 + 0.5 * 6.06 + 0.4 * 7.85 + 1.5 * 13.37 + 8.4 * 14.98 + 55.4 * 56.31) / 100 = 31.2 defunciones por mil habitantes.

En el primer caso tomamos en cuenta el hecho de que las tasas por edad están multiplicadas por mil y que la distribución está dada en por ciento, por lo que se corrieron los puntos decimales hacia la izquierda 3 y 2 lugares respectivamente, multiplicándose el resultado por mil. En el caso B, se hicieron los cálculos con los valores tales como aparecen en la tabla y al final se dividió por 100, con lo que la tasa de B quedó automáticamente multiplicada por mil. Ambos caminos son equivalentes. Nótese la gran diferencia de resultados por el solo hecho de poseer los países diferentes estructuras por edad.

Es conveniente dejar claro que casi siempre que se da una serie de tasas específicas de mortalidad por edades, se acostumbra poner en el grupo de cero año la tasa de mortalidad infantil en vez de la tasa específica para esa edad. A la hora de tipificar debe tenerse presente que lo que se usa es ésta última.

3.3.3 Tasas por causas de muerte.

En definitiva, el cálculo de tasas por las distintas causas o grupos de éstas, no tiene un tratamiento diferente al visto hasta ahora. Sea una tasa bruta o específica, se procede como hasta el momento.

En Cuba durante 1996 se produjeron las siguientes defunciones por las causas señaladas,

Causa	Defunciones	Tasas $x10^5$
Enfermedades del Corazón	22660	205.9
Tumores Malignos	15112	137.3
Enfermedades Cerebro-vasculares	7945	72.2
Accidentes	5653	51.4
Influencia y neumonía	4462	40.5
Enfermedades de las arterias, arteriolas y capilares	3512	31.9
Diabetes. Mellitus	2573	23.4
Suicidio y lesiones auto inflingidas	2009	18.2
Bronquitis, enfisema, asma	1027	9.3
Cirrosis y otras enfermedades crónicas del hígado	923	8.4
Sub- total	65876	598.5

Las tasas fueron calculadas mediante el cociente de las defunciones entre la población media del año 1996 (11005866 habitantes). Note que las tasas son aditivas, es decir, se pueden sumar y da como resultado la tasa de todas esas causas juntas. Esto es lícito hacerlo, ya que el denominador de las tasas es común.

De igual manera se procede cuando se trabaja con los grupos de edades. De inmediato traemos un ejemplo, con información de Cuba por edad y causa.

Tasas de Mortalidad por grupo de edades y causas seleccionadas. Cuba, 1986 y 1996.

	1988		1996	
	Tumores malignos	Muertes Violentas	Tumores Malignos	Muertes Violentas
0-4 años	6.8	29.6	6.4	22.9
5-14 años	5.2	24	5.6	15.7
15-39 años	13.7	78.6	15.8	63.1
40-64 años	170.5	81.2	166.4	72
65 y más	955.3	229.7	985.5	324.2
Total	127.5	79.6	141	79.3

Aquí las tasas se calculan con las defunciones de cada grupo de edades y causa, con la población media del grupo, todo para la fecha en cuestión. Los resultados por cada grupo se pueden sumar, dentro del mismo año, ya que el denominador es idéntico: en 1996, la tasa de Tumores Malignos y Muertes Violentas para el grupo

5-14 años (5.6 + 15.7 = 21.3), mide el riesgo de muerte por las dos causas como un todo. Sin embargo, en la última fila de la tabla, la tasa total de cada causa en los años señalados, no es una simple suma, sino la media ponderada con los valores de la población de cada grupo de edades.

3.4 Diferencia porcentual de las tasas.

Este índice es útil cuando queremos conocer cuánto ha aumentado o disminuido una tasa entre dos momentos, en forma relativa.

$$\text{Diferencia porcentual} = \frac{\text{tasa}_0 - \text{tasa}_1}{\text{tasa}_0} * 100 = 100 - \frac{\text{tasa}_1}{\text{tasa}_0} * 100$$

donde tasa_0 se refiere al valor inicial y tasa_1 al final.

¿Cuál es el significado de esta diferencia?.

Ella representa el porcentaje de decremento (incremento) de la tasa respecto al valor inicial de la misma. Cuando la expresión $\mathbf{tasa_0 - tasa_1}$ es positiva**,** quiere decir que disminuyó el riesgo de muerte pues la tasa final es menor que la inicial. Equivalentemente, en la segunda igualdad el valor 100 representa el nivel de partida, al que se le resta el porcentaje de la tasa en el momento 1 con relación al momento 0, resultando en el porcentaje que ha disminuido la tasa final en relación a la inicial. Tomemos la tasa de Tumores Malignos del grupo 15-39 años; en 1986 era de 13.7 y en 1996 de 15.8. La diferencia porcentual experimentada es:

$$\text{Diferencia Porcentual} = \frac{13.7 - 15.8}{13.7} * 100 = -15.33\ \%$$

Como el valor resultante es negativo, concluimos que la tasa en 1996 fue 15.33% mayor que en 1986.

En este mismo grupo de edades, las Muertes Violentas pasaron de 78.6 a 63.1, por tanto la diferencia porcentual es de :

$$\text{Diferencia Porcentual} = \frac{78.6 - 63.1}{78.6} * 100 = 19.72\%$$

La causa en cuestión disminuyó en 10 años su tasa en casi un 20% con respecto al valor del año 1986.

También con un simple cociente de la forma

$$\frac{63.1}{78.6} * 100 = 80.28\%,$$

nos percatamos que la tasa del final de período es el 80.28% del 100% que es la tasa de 1986, o sea, casi 20% menor que aquella. Ambas maneras de enfocar el problema son equivalentes, lo que en la segunda hay que hacer la resta de 100 (o mejor dicho, la suma algebraica) mentalmente, mientras en la primera sale directamente el porcentaje de disminución (aumento).

3.5 Sobre mortalidad Masculina .

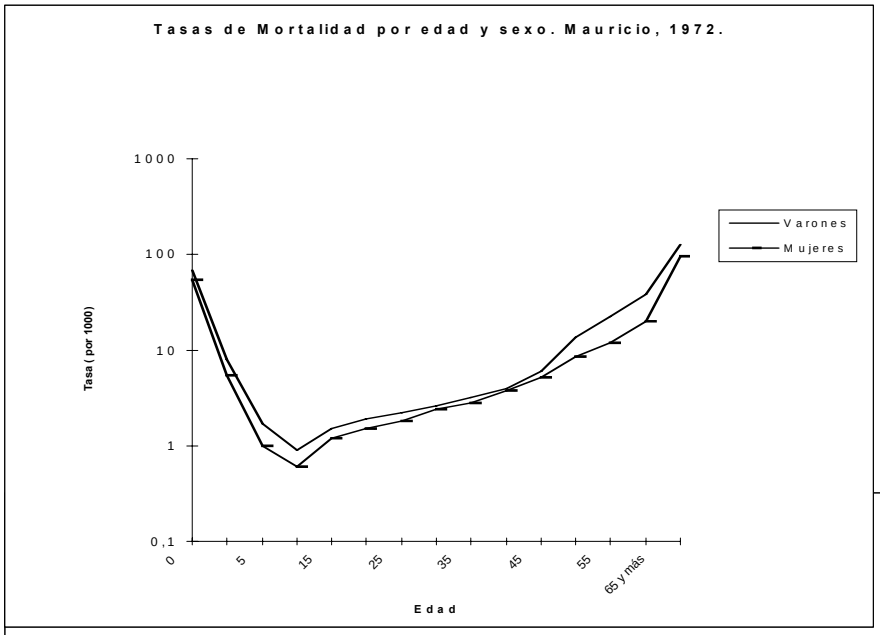
El comportamiento de la mortalidad por edad si bien varía de un país a otro, tiene, en general, un patrón bastante definido con independencia de la región de que se trate. Usualmente, se describe la curva de las tasas específicas de mortalidad en función de la edad, como una letra “J” alargada, siendo el riesgo muy

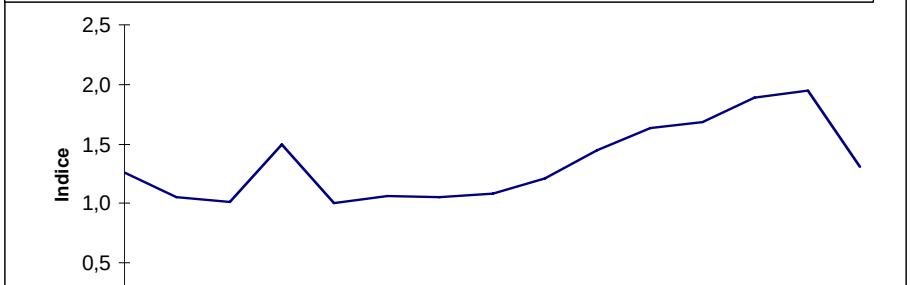
elevado en las primeras edades, con un mínimo entre los 5 y 15 años, y luego crecientes hasta el final de la vida.

Otra peculiaridad es que los riesgos de morir en el sexo masculino son más elevados que en el femenino, midiéndose el exceso de mortalidad de los varones a través de la relación entre las tasas de hombres y de mujeres. A esta relación se le ha dado el nombre de Índice de Sobremortalidad Masculina.

Tasas de Mortalidad por edad y sexo. Mauricio, 1972

Grupo de Edad	Varones	Mujeres	Sobre mortalidad Masculina
0 año	68.2	54.1	1.3
1-4 años	5.7	5.4	1.1
5-9 años	1.0	1.0	1.0
10-14 años	0.9	0.6	1.5
15-19 años	1.3	1.2	1.1
20-24años	1.6	1.5	1.1
25-29 años	1.9	1.8	1.1
30-34 años	2.6	2.4	1.1
35-39 añ0s	3.4	2.8	1.2
40-44 años	5.5	3.8	1.4
45-49 años	8.5	5.2	1.6
50-54 años	14.5	8.6	1.7
55-59 años	22.5	11.9	1.9
60-64 años	38.5	19.8	1.9
65 y más	126.1	96.3	1.3





Note que la relación entre las tasas de los sexos siempre permanece por encima del valor 1, lo que es indicativo de que los hombres poseen un riesgo mayor de mortalidad general que las mujeres. No obstante, se ha encontrado en Cuba, un exceso de riesgo femenino en causas de muerte como la Diabetes Mellitus, donde las mujeres parecen tener cierta desventaja con respecto a los hombres.

Usualmente, el gráfico de las tasas por edad se representa con el eje vertical en escala logarítmica; que tiene la ventaja de abarcar un rango mayor por el hecho de contraer la escala. Si se usara una escala aritmética, entonces el eje vertical sería muy largo pues hay que representar desde valores muy pequeños hasta muy elevados.

3.6 La mortalidad infantil.

La medición del riesgo de muerte en los menores de un año suele hacerse con la conocida Tasa de Mortalidad Infantil, que algunos expertos opinan no es realmente una tasa sino que es más parecida a una probabilidad. No obstante tampoco es la única manera de estimar el riesgo de muerte en esa edad.

La forma clásica de la tasa de mortalidad infantil relaciona las defunciones de menores de un año, dentro de un límite temporal de un año, con los nacidos vivos del propio año calendario y se multiplica por 1000.

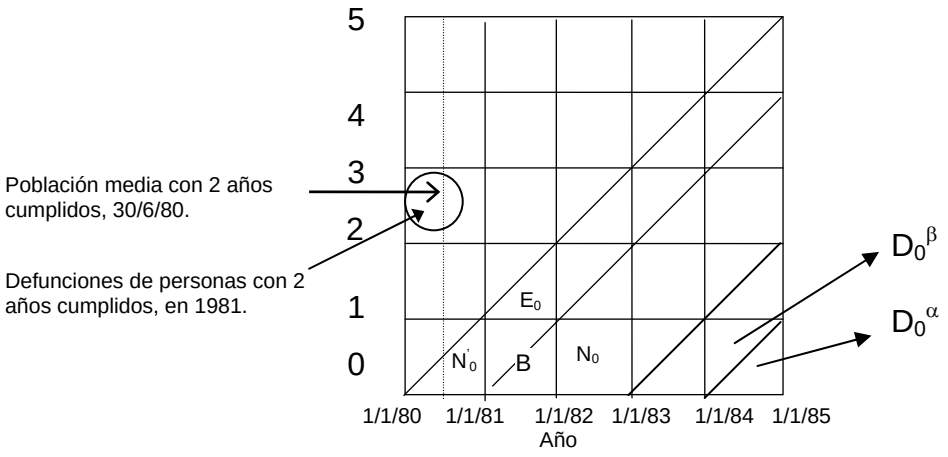
$$TMI = \frac{D^{t}_{<1}}{B^{t}} * 1000$$

La otra manera de medir es con la tasa específica de mortalidad referida a la edad cumplida de cero año.

$$m_0 = \frac{D_0{}^{t}}{N_0}$$

pero esta última forma es menos usada.

Es importante tener en cuenta que aunque se refieren al mismo segmento de edad, miden el riesgo de manera distinta. La primera lo hace con relación a los nacidos vivos y la segunda con respecto a la población de edad cumplida cero año.

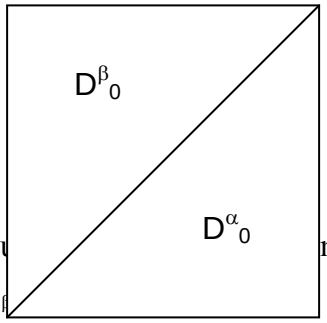


Se puede refinar la medida y estimar el riesgo a través de un procedimiento que toma en cuenta la experiencia de las dos cohortes sucesivas que participan en la mortalidad del año (recuérdese que en un cuadrado hay defunciones de dos generaciones contiguas).

El método se debe a Greville y es conocido por ese nombre. Parte de la descomposición de la probabilidad de sobrevivir entre dos edades exactas en un producto de dos probabilidades, esto es

$$P_0^{\alpha} = \frac{N_0}{B} \quad y \quad P_0^{\beta} = \frac{E_1}{N'_0}$$

E_1



Por último la tasa bu mplemento del producto de las dos probabilidades anteriores

$$TMI = 1 - P_0^{\alpha} * P_0^{\beta}$$

Ahora bien, para el cálcul $B = E_0$ necesario partir de la población N'_0 y disponer de las defunciones D^{β}_0 (ver diagrama anterior). Los valores de N_0 y N'_0 pueden obtenerse mediante relaciones del tipo

$N_0 = B - D^{\alpha}_0$ y $N'_0 = B - D^{\beta}_0$

o por los registros o censos existentes.

3.6.1 Componentes de la mortalidad infantil

La mortalidad dentro del primer año de vida, se diferencia según se esté próximo o no al momento del nacimiento. Clásicamente se divide en tres momentos: mortalidad neonatal precoz; mortalidad neonatal tardía y post- neonatal.

La primera comprende el lapso desde el nacimiento hasta antes de 7 días; la segunda, desde los 7 días hasta 27 y la tercera a partir de 28 días hasta los 11 meses y 29 días. Estos intervalos son de edades cumplidas; por ejemplo, el primer intervalo va desde la edad exacta 0 año (momento del nacimiento) hasta los 6 días cumplidos, lo que significa que llega casi hasta la edad exacta 7 días pero sin incluirla.

La mortalidad neonatal precoz se asocia con causas de muerte de las llamadas endógenas como las malformaciones congénitas, mientras que, por ejemplo la post- neonatal con causas de muerte más ligadas al medio ambiente tales como infecciosas y parasitarias, en general, evitables.

A medida que se produce un progreso en las condiciones higiénicas y medioambientales, se espera que la mortalidad de los menores de un año vaya concentrándose más cerca del parto, o sea en el segmento

neonatal, donde imperan entidades muchas de las cuales son de corte " no evitables" con el conocimiento actual. No obstante, debido al avance tecnológico en los cuidados intensivos, muchos niños que engrosarían la mortalidad neonatal ahora lo hacen en la post- neonatal. En Cuba, en 1970 las muertes neonatales ocupaban el 63.5%; en 1985, 61.8% y para 1996 se experimenta un ligero incremento hasta 63.3%. Las tasas por componentes se calculan como el cociente de los fallecidos en el intervalo de edad específico y los nacidos vivos.

En 1996 se produjeron en Cuba 140276 nacidos vivos. En ese propio año hubo 1109 defunciones infantiles, distribuidas de la siguiente manera: 456 en el período neonatal precoz; 250 en el neonatal tardío y 403 en el post- neonatal. Las respectivas tasas por componentes fueron:

Tasa mortalidad neonatal precoz = (456/ 140276)* 1000 = 3.25

Tasas mortalidad neonatal tardía = (250/140276)*1000 = 1.78

Tasa mortalidad post- neonatal = (403/ 140276)*1000 = 2.87

Estas tasas son aditivas, o sea que se pueden sumar y obtener la tasa total de mortalidad del menor de un año, en este caso 7.9.

Otro indicador muy utilizado y que mide el riesgo de mortalidad del producto de la concepción alrededor del parto es la tasa de mortalidad perinatal. Tiene por componentes la mortalidad fetal tardía (1000 gramos y más de peso del feto) y la neonatal precoz. Se calcula como el cociente de las defunciones perinatales (numerador) y los nacidos vivos más las defunciones fetales tardías (denominador).

Tasa mortalidad perinatal =

[(defunciones neonatales precoces + defunciones fetales tardías) / (nacidos vivos + defunciones fetales tardías)] *1000

El denominador de esta tasa pretende estimar el número de embarazos viables, es decir, aquellos que no han sido objeto de abortos. En 1996 la tasa de mortalidad perinatal de Cuba era de 12.4 defunciones perinatales por 1000 nacidos vivos y defunciones fetales de 1000 gramos y más.

3.6.2 Factor de separación y tasa de mortalidad infantil por cohorte

Ya hemos visto que en un cuadrado se representan las defunciones de dos generaciones sucesivas. Dicho cuadrado puede dividirse en dos triángulos, uno superior y otro inferior. En el superior se representan las defunciones de la cohorte precedente y en el inferior las de la subsiguiente.

Las muertes del triángulo superior se denotan por ${}_n\mathbf{D}_x^{\beta}$ y las del inferior por ${}_n\mathbf{D}_x^{\alpha}$.

El factor de separación de las muertes, se define como la proporción de las defunciones del triangulo superior respecto al total de muertes del cuadrado y se denota por ${}_nf_x$.

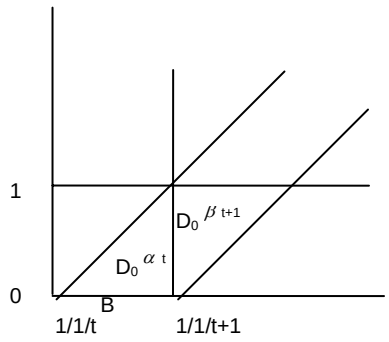
$$ {}_nf_x = \frac{{}_n\mathbf{D}_x^{\beta}}{{}_n\mathbf{D}_x} . $$

Para su cálculo es necesario conocer de los fallecidos, al menos dos de los tres elementos siguientes: fecha de nacimiento, edad al morir, fecha de muerte.

Es útil para reconstruir cohortes cuando sólo conocemos las muertes totales por edad. En muchas ocasiones no conocemos las defunciones de los triángulos inferior y superior pero podemos asumir los factores de separación de otra población con características similares o usar los de otros años para la misma población, etc.

La tasa de mortalidad infantil por generación resulta fácil de calcular cuando conocemos las muertes correspondientes al triángulo inferior del año en cuestión y las del triángulo superior del siguiente año. Esto es:

$$TMI = \frac{D_0^{\alpha} + D_0^{\beta}}{B} * 1000$$



Ejercicio resuelto:

Calcule la tasa de crecimiento de una población, sabiendo que su tiempo de duplicación es de 60 años.

Sabiendo que $t = \frac{\ln(2)}{\ln(1+r)}$ es la fórmula que da el tiempo de duplicación de la población que crece a una tasa constante r, según la ley del interés compuesto, se puede despejar r.

Despejando ln(1+r) y aplicando antilogaritmo queda que

$$\ln(1+r) = \frac{\ln(2)}{t} = \frac{1}{t} \cdot \ln(2) = \ln(2)^{\frac{1}{t}}$$

$$1 + r = 2^{\frac{1}{t}}$$

$r = 2^{\frac{1}{t}} - 1$ o sea, la raíz de orden t del número 2, a lo que luego se le resta 1.

Sustituyendo t = 60 y con el auxilio de una calculadora científica, se tiene

r = 1.16%.

Resumen del capítulo

Deben recordarse los pasos fundamentales para el cálculo de una tasa:

- Incremento absoluto (diferencia absoluta entre dos momentos)
- Incremento medio anual (incremento absoluto entre longitud del intervalo en años)
- La tasa propiamente dicha (Incremento medio anual entre la población)

Cuando usamos la fórmula del interés compuesto, el cálculo de la tasa es diferente.

Ejercicio propuesto:

Durante el primer semestre del año 1999, la población de cierta comunidad se incremento en 100 000 personas. Al calcular la tasa de crecimiento, se dividió dicho incremento entre la población media del primer semestre. ¿Cómo Ud. interpretaría dicha tasa?.

4. Ajuste de Tasas.

En una sección anterior de este documento, vimos que cuando hacíamos comparaciones del nivel de la mortalidad, a través de la Tasa Bruta de Mortalidad, existía el inconveniente de que la distribución de la población por edades ejercía un efecto confusor; es decir, enturbiaba la comparación, no permitiendo decidir si la diferencia observada se debía a diferencias en los niveles de mortalidad de los países o regiones comparados. Para controlar este efecto adverso, existen dos métodos de ajuste que son aplicables a cualquier tasa bruta, sea de incidencia (la mortalidad es un tipo de incidencia), prevalencia, de natalidad, etc.

A grosso modo, el ajuste, también llamado tipificación o estandarización, es una técnica de amplio uso. No sólo se realiza por edad, también por otras variables para las cuales se dispone de la distribución de la población.

Los dos métodos más utilizados se conocen como Método Directo y Método Indirecto. Pasemos de inmediato a describir el Método Directo.

4.1 Método de ajuste Directo

Básicamente, lo que suele hacerse es tomar una distribución por edad de la población tipo o estándar y recalcular todas las tasas brutas que se desean comparar utilizando dicha distribución. Para aplicar este método es necesario contar con las tasas de mortalidad específicas por edad de los países.

Recordemos que:

$$TBM = \sum \left(\frac{{}_nN_x}{N}\right) * {}_nm_x$$

donde el término entre paréntesis es la proporción de población en el grupo de edades

(x, x + n).

Si dos países tienen tasas brutas de mortalidad TBM^A y TBM^B, para tipificarlas por el método directo, primero seleccionamos una distribución por edad estándar o tipo, que puede ser la de un tercer país, digamos. Acto seguido, sustituimos en la fórmula anterior los valores correspondientes y realizamos el cálculo. Hemos controlado el efecto adverso que provoca el hecho de que A y B posean diferentes distribuciones por edad, imponiendo una común a ambos, la escogida como tipo. Al ser iguales las mismas, entonces se puede ver con facilidad en la fórmula anterior, que la diferencia encontrada entre las tasas brutas recalculadas se debe sólo a las diferencias en los niveles de mortalidad entre ambos países. Algún ejemplo con cifras resultará ilustrativo.

Ejemplo:
Tasas de mortalidad por edad y distribución de la población, países A y B

	País A		País B	
Grupos de edades	Tasas Mortalidad	Distribución	Tasas Mortalidad	Distribución
<1	22.3	1%	29.3	3%
1-4 años	0.5	6%	0.5	7%
5-14 años	0.4	15%	0.3	10%
15-49 años	1.5	56%	1.1	30%
50-64 años	8.4	13%	8.7	20%
65 y más	55.4	8%	56.1	30%
Total	6.7	100%	19.8	100%

Las tasas totales de A y B son 6.7 y 19.8 por mil, pero si utilizamos la distribución de A como tipo vemos que:

$TBM^{A|A} = TBM^A$,
la tasa de A es la misma cuando usamos la distribución de A.
$TBM^B = TBM^{B|A} = (29.3*0.01+0.5*0.06+0.3*0.15+1.1*0.56+8.7*0.13+56.1*0.08) = 6.6$,
la tasa de B calculada con la distribución por edad de A es similar a la de A. Hemos hecho comparables las dos tasas brutas, controlando el efecto de la distribución. Fíjese que se ha corrido el punto decimal en los porcentajes, pues de no hacerse así, entonces la tasa quedaría multiplicada por 100 además de por 1000, o sea por 100000.

Si se toma la distribución de B como tipo, entonces la tasa total o bruta de A sería
$TBM^{A|B} = 19.7$ por mil personas, muy similar a la bruta de B (antes de tipificarla) y la tasa ajustada de B es la misma (19.8) pues se está usando su propia distribución. También puede usarse como tipo la distribución promedio de los dos países.

Algo interesante: no sé si se han dado cuenta que los valores de las tasas ajustadas cambian según se disponga de una u otra distribución estándar, este es un resultado importante. Otra cuestión, la tasa ajustada o estandarizada es de hecho un índice resumen, que no tiene el "defecto" de la tasa bruta.
Cuando un país tiene sus tasas específicas por edad superiores al de otro país en todas las edades, no hay dudas sobre quién tiene un nivel de mortalidad más elevado, no obstante, se calculan para ambos tasas ajustadas con el propósito de tener un resumen de la mortalidad para ambos.

Diferencias socio-económicas que determinan la diferencia de niveles de mortalidad

Si tuviéramos que dar una respuesta sobre la diferencia entre las tasas ajustadas de los países A y B, diríamos que la diferencia entre las mismas se debe a una serie de factores socioeconómicos que prevalecen en cada uno de ellos y que determinan sus respectivos niveles de mortalidad, con la salvedad de que se excluye la distribución por edad de la población. Dicho en otros términos, esa diferencia está determinada por las condiciones socioeconómicas que incluye el grado de desarrollo, nivel de instrucción, recursos para la salud y muchos otros más, pero no por la diferencia en la estructura por edad de la población entre A y B.
Lo dicho en el párrafo precedente nos prepara para admitir el siguiente acápite.

4.2 Estandarización de tasas controlando el efecto de distribuciones de población diferentes de la edad

En el siguiente ejemplo, se da la tasa de mortalidad infantil por zonas de residencia (urbana y rural), así como el porcentaje de población en cada una, para los años 1980 y 1990; datos hipotéticos.

Año	Zona Urbana	Zona Rural	Ambas Zona
1980	21.0 (40%)	26.0 (60%)	24.0 (100%)
1990	22.0 (90%)	27.0 (10%)	22.5 (100%)

La tasa de mortalidad infantil total se calculó como la media ponderada de las dos zonas:

$$TMI^{1980\ (\text{A. Zonas})} = (21.0*0.40 + 26.0*0.60) = 24.0$$

En este ejemplo se da una paradoja, los riesgos de muerte infantil han aumentado en cada zona, al pasar de 1980 a 1990, mientras que el riesgo total ha disminuido. ¿ qué ha sucedido aquí?. Sencillamente que ha variado la distribución de la población por zona de residencia. La urbanización pasó de 40 % en el 1980 a 90% en 1990, por lo que el promedio ponderado en el último año tiende a acercarse más al valor con mayor peso, en este caso el de la parte urbana.

$$TMI^{1990\ (A.\ Zonas)} = (\ 22.0*0.90 + 27.0*0.10) = 22.5$$

Peso de ponderación mayor

Si ajustamos la tasa total de 1990 utilizando la estructura de 1980 nos queda

$$TMI^{1990(A.Zonas)} = (\ 22.0*0.40 + 27.0*0.60) = 25.0$$

Como observamos, luego del ajuste, el riesgo en 1990 es superior al de 1980.

4.3 Ajuste de tasas por más de una variable

Hasta ahora sólo hemos visto la aplicación del método directo, para el control de una variable confusora, la edad. A continuación veamos una tabla con información sobre incidencia de tuberculosis según la localización (pulmonar y extra pulmonar) de dos países, A y B.

Incidencia de Tuberculosis por edad y localización

Grupo de edades	País A				País B			
	Pulmonar		Extra pulmonar		Pulmonar.		Extra Pulmonar.	
	1	2	3	4	5	6	7	8
	Tasa	Distrib. %	Tasa	Distrib.%	Tasa	Distrib.%	Tasa	Distrib.%
<15	1.6	40	0.8	45	0.2	31	0.2	25
15-44	14.2	30	1.6	35	3.4	53	0.6	60
45-64	26.6	20	2	15	8.4	12	0.7	10
65 y más	42.3	10	2.3	5	19.6	4	1.8	5
Total	14.5	100	1.3	100	3.7	100	0.6	100

Distrib.Local.	51%	49%	44%	56%
Tasa ajustada(edad)	14.5	1.4	4.7	0.6

Ambas localizaciones 7.2 2.7

$\sum col(1) * col(2) = 14.5$ $\sum col(2) * col(3) = 1.4$ $\sum col(5) * col(2) = 4.7$ $\sum col(7) * col(2) = 0.6$

Tasa ajustada por edad y localización

A B

14.5 * 0.51 + 1.4 * 0.49 = 7.2 4.7 * 0.51 + 0.6 * 49 = 2.7

Existen dos variables de clasificación: la edad y la localización. Primero ajustaremos las tasas totales por la edad y luego por la localización. Tomaremos la distribución tipo por edad correspondiente al país A (pulmonar) y luego la distribución por localizaciones de B.
En la tabla se ha señalado en el procedimiento de cálculo las columnas que intervienen.

La $TBI^{A|A} = TBI^{A}$, la tasa de A (pulmonar) ajustada con la distribución por edades de A es igual a la tasa original, 14.5. Las tasas ajustadas por edades se dan en la fila encabezada con " Tasa Ajustada(edad)". Luego aparece en la última fila las tasas totales (ambas localizaciones) como una media de las tasas por localizaciones ajustadas por la edad, ponderadas por los porcentajes de enfermos en cada una de ellas, sus valores 7.2 y 2.7. Estos valores son ligeramente diferentes a las tasas totales (ambas localizaciones), de cada país antes de los ajustes, que no aparecen en la tabla pero cuyo cálculo es sencillo: para A,

14.5*0.51+1.3*0.49=7.2

y para B,

3.7*0.44+0.6*0.56=2.4.

Aspectos a tener en cuenta con el Método Directo de tipificación.

1. Se utiliza cuando si se conocen las tasas específicas por edad de los países que se van a comparar. Estas tasas deben calcularse de manera tal que posean confiabilidad estadística, sobre todo en poblaciones no muy pequeñas.

2. Los resultados del ajuste dependen de la población tipo utilizada, incluso puede llegarse a conclusiones contradictorias cuando se procede con una u otra distribución. Por ejemplo, puede que un país presente menor nivel de mortalidad que otro cuando se fija determinada estructura estándar y por el contrario, al cambiarla por otra, resulte con el nivel mayor. No sucede siempre pero debemos estar alertas.

3. Se pueden comparar dos o más países.
4. En principio se puede usar cualquier distribución estándar, pero muchas veces es recomendable tomar la promedio de los países involucrados.

4.4 Diferencia entre dos tasas brutas.

Si tenemos dos países, A y B, podemos descomponer la diferencia entre sus tasas brutas de mortalidad de la siguiente manera:

Sean TBM^A y TBM^B las respectivas tasas, si se toma la distribución por edades del país A como tipo, entonces tenemos que

$TBM^{B|A} = \sum {}_nm_x * {}_nC_x^A$, donde C_x^A simboliza la distribución por edades de A.

A través de algunos manejos algebraicos, se llega a

$$TBM^A - TBM^B = \sum C_x^A * ({}_nm_x^A - {}_nm_x^B) + \sum {}_nm_x^B * (C_x^A - C_x^B)$$

La primera sumatoria a la derecha de la igualdad representa una media ponderada de las diferencias de las tasas específicas de A y B, donde los factores de ponderación lo constituyen la proporción de población en cada grupo de edades de A. Esta es la componente Mortalidad.

La segunda sumatoria es una suma de los productos de las diferencias entre las proporciones de población por edad de A y B y las tasas específicas por edad de B. A esta se le llama la componente Estructural.

Ejercicio resuelto:
Comparar las tasas de mortalidad de Cuba para los años 1989 y 1996, según el procedimiento descrito.

Los cálculos están indicados en las columnas correspondientes de la tabla siguiente:

Edad	Tasa1989 (1)	Tasa 1996 (2)	Dist. 1996 (3)	Dist 89 (4)	Comp. Mort. (5)=[(2)-(1)]*(3)	Comp. Estru. (6)=(1)*[(3)-(4)]
<1 año	21.8	7.4	0.01	0.02	-0.2	-0.1
1-4 año	0.7	0.6	0.06	0.07	0.0	0.0
5-14 año	0.4	0.3	0.15	0.15	0.0	0.0
15-49 año	1.7	1.7	0.55	0.57	0.0	0.0
50-64 año	8.8	8.4	0.14	0.12	-0.1	0.2
65 y más	48.4	54.9	0.09	0.08	0.6	0.3
Total	6.6	7.3	1.00	1.00	0.3	0.4

Diferencia de Tasas = 0.7

Observe que la diferencia entre las tasas brutas es igual a 0.7 y también que la suma de las componentes estructural y mortalidad da esa diferencia. Es interesante ver que la parte estructural representa alrededor del 57% (0.4/0.7) de la diferencia total, de ahí la importancia de hacer el ajuste.

4.5 Método de ajuste indirecto

La tipificación indirecta se utiliza también para hacer comparaciones del nivel de la mortalidad entre países aunque tiene sus peculiaridades. Siempre que sea posible, preferiremos el método directo de estandarización, para lo cual precisamos de las tasas por edades y de una distribución por edades tipo. Cuando no contamos con las tasas específicas por edades, sea porque no tenemos la información para su cálculo, sea porque se trata de poblaciones pequeñas en las cuales no serían confiables las estimaciones de los riesgos por edades, usamos el procedimiento indirecto.
¿En que consiste el ajuste indirecto de tasas?

Si en la tipificación directa nosotros anulábamos el efecto de la diferencia entre las distribuciones por edades de los países que se confrontaban, con la imposición de una común tipo, ahora se procede a estimar este efecto.
Se supone que conocemos la tasa bruta de mortalidad, así como la distribución por edades del país objeto de estudio.
Primeramente se toma un país para el cual son conocidas las tasas específicas por edades, el cual será el patrón o tipo que utilizaremos. Luego se aplican estas tasas tipo a la población por edades del otro país, con lo que se calcula una especie de tasa bruta pero con la mortalidad del país patrón. A continuación hacemos el cociente entre la tasa bruta del patrón y la tasa bruta del país objeto pero con las tasas por edades tipo.

Llamémosle A al país objeto de estudio y C al patrón o tipo.

$$I = \frac{TBM^{C}}{\sum {}_{n}m_{x}^{C} * {}_{n}N_{x}^{A} / N^{A}}$$

Este índice mide el efecto de la distribución de A , tomando como referencia a C. Podemos rescribir el mismo de la siguiente forma

$$I = \frac{\sum {}_{n}m_{x}^{C} * {}_{n}N_{x}^{C} / N^{C}}{\sum {}_{n}m_{x}^{C} * {}_{n}N_{x}^{A} / N^{A}}$$

La diferencia entre el numerador y denominador estriba solamente en que en el último las tasas de C se multiplican por la estructura de A. Por tanto el valor del índice dependerá del efecto de la distribución de A.

Si I es mayor que 1 (I > 1) significa que el denominador es menor que el numerador, lo que es equivalente a decir que la estructura por edades del país A tiene un efecto que tiende a bajar la tasa bruta de C. Consecuentemente, podemos pensar que la propia tasa de A está más baja a causa del efecto favorable de su estructura por edades. En estas circunstancias la tasa bruta de A debería ser aumentada en cierta medida para hacerla comparable con la tasa bruta de C (con el objetivo de emparejarla con C en cuanto a efecto estructura) lo que se logra multiplicándola por el índice I.
La tasa bruta de A ajustada por el método indirecto resulta, $TBM^{A|\,ajust} = I * TBM^{A}$.
Ahora estamos en condiciones de comparar la tasa bruta de C (país patrón) con la de A (país objeto) .
Nótese algo importante, el efecto de la distribución por edades de A (si eleva o disminuye la tasa) no podemos medirlo con respecto al propio país A, por eso se toma como referencia a C.
Si el valor de I hubiera resultado menor que 1 (I < 1), entonces es a causa de que la distribución del país A eleva la tasa de C (denominador mayor que el numerador) y en este caso debemos disminuir la tasa bruta de A multiplicándola por un valor menor que 1 para hacerla comparable con la del país C.

El caso en que I es igual a 1(I = 1) significa que la estructura de A tiene el mismo efecto que la de C y por tanto no es necesario la corrección anterior; las dos tasas se comparan sin dificultad.

Vamos a suponer que nos interesa un tercer país llamado B, que contamos con su tasa bruta de mortalidad y la distribución por edades. Procedemos en la misma forma que con A; calculamos el índice I usando ahora la estructura de B y luego multiplicamos la tasa total de B por dicho índice. Ahora ambos países han sido equilibrados en cuanto al efecto de sus distribuciones por edad; es decir, se han igualado los mismos al multiplicar las respectivas tasas brutas por los correspondientes índices. Ya podrán compararse los países A y B.

No es necesario aplicar los dos métodos, con uno solo basta. Siempre que sea posible es preferible utilizar el método directo de ajuste.

De inmediato vamos a tratar otra forma equivalente del ajuste indirecto. Tomemos la fórmula anterior y escribámosla de esta manera

$$TBM^{A|ajust} = \frac{\sum {}_nm_x^{\,C} * {}_nN_x^{\,C} / N^{C}}{\sum {}_nm_x^{\,C} * {}_nN_x^{\,A} / N^{A}} * \sum {}_nm_x^{\,A} * {}_nN_x^{\,A} = \frac{\sum {}_nm_x^{\,A} * {}_nN_x^{\,A} / N^{A}}{\sum {}_nm_x^{\,C} * {}_nN_x^{\,A} / N^{A}} * \sum {}_nm_x^{\,C} * {}_nN_x^{\,C} / N^{C}$$

Índice I_A | TBM^A | Cociente entre TBM^A y $TBM^{(C|A)}$ (REM) | TBM^C

El miembro derecho de la ecuación es el cociente entre la TBM^A y la tasa bruta de C pero con la estructura de A, todo ello multiplicado por la TBM^C.

Prestaremos atención al cociente entre TBM^A y $TBM^{(C|A)}$. En este caso lo que hemos evaluado es el efecto de la mortalidad de C sobre la tasa bruta de A; es decir, se responde a la interrogante ¿ cuántas veces mayor (menor) sería la tasa total de A si este país tuviera los riesgos por edades observados en el país C?. Este cociente se denomina Razón Estandarizada de Mortalidad (REM). Al multiplicarlo por la TBM^C obtenemos también la tasa ajustada para A.

Es bueno dejar claro que la REM no es la tasa ajustada, no obstante, con ella se realizan análisis comparativos. La REM multiplicada por TBM^C si es la tasa ajustada de A por la técnica indirecta, al igual que también lo es I_A multiplicado por TBM^A. Veamos el siguiente ejemplo:

	Tasas de C	País A		País B		Defunciones de C
	(1)	(2)	(3) = (1)*(2)	(4)	(5)= (1)*(4)	
grupo de edades	${}_nm_x^{\,C}$	${}_nN_x^{\,A}$	${}_nD_x^{\,A*}$	${}_nN_x^{\,B}$	${}_nD_x^{\,B*}$	${}_nD_x^{\,C}$
0-14	0.006	20000	120	40000	240	210
15-64	0.005	55000	275	55000	275	275
65 y+	0.100	25000	2500	5000	500	1000
Total		100000	2895	100000	1015	1485
TBM	14.9	18.7	29.0	13.0	10.2	14.9

En el cuadro anterior, se observan las tasas específicas de mortalidad de C así como la población por edades de los países A y B. Para ajustar la tasa de A por el método indirecto procedemos como a continuación:

cálculo del índice I_A

$$I_A = \frac{TBM^C}{\sum {}_nm_x{}^C * {}_nN_x{}^A / N^A} = \frac{14.9}{29.0} = 0.51$$

El denominador del índice se calculó como la multiplicación de las columnas (1) y (2), lo que resultó en la columna (3) que representan las defunciones esperadas para el país A si tuviera los riesgos por edades de C o también las defunciones esperadas del país C si tuviera la distribución por edades de A. Entonces la TBM de A sería 29.0

Ahora calculemos la REM, $REM = \frac{TBM^A}{\sum {}_nmx^C * {}_nNx^A / N^A} = \frac{18.7}{29.0} = 0.64$

El denominador de la REM es idéntico al de I_A. Como la REM es menor que 1, concluimos que si el país A tuviera las tasas de mortalidad por edades iguales a las de C, entonces su tasa bruta sería 29.0 y no 18.7. A tiene una mortalidad general menor que C. Si la REM hubiera dado mayor que 1, entonces diríamos que la mortalidad del país A es superior a la de C.

$TBM^{A|AJUST} = I_A * TBM^A = REM * TBM^C$

$9.53 = 0.51 * 18.7 = 0.64 * 14.9$

Como se ha visto, el ajuste indirecto se puede realizar de dos maneras, a través del índice I y a través de la REM.

Igualmente se procede con el país B. En la tabla se muestran las defunciones esperadas para B si tuviera las tasas por edades de C.

$$I_B = \frac{14.9}{10.2} = 1.46$$

$$REM_B = \frac{13.0}{10.2} = 1.27$$

$$I_B * TBM^B = REM_B * TBM^C$$

$TBM^{B|AJUST} = 1.46 * 13.0 = 1.27 * 14.9 = 18.9$

El I_B mayor que 1 nos dice que la distribución por edades de B tiende a disminuir la tasa de C; por otra parte, la REM_B también mayor que 1 nos está informando que si el país B tuviera las mismas tasas por edades que C, entonces su mortalidad general sería menor que la observada.

La comparación a través de la REM se hace entre el país en estudio y el patrón, esto es, entre A y C o entre B y C, pero no debe relacionarse A con B. Esto es debido a que no se controlan los efectos de las distribuciones por edades de estos dos países.

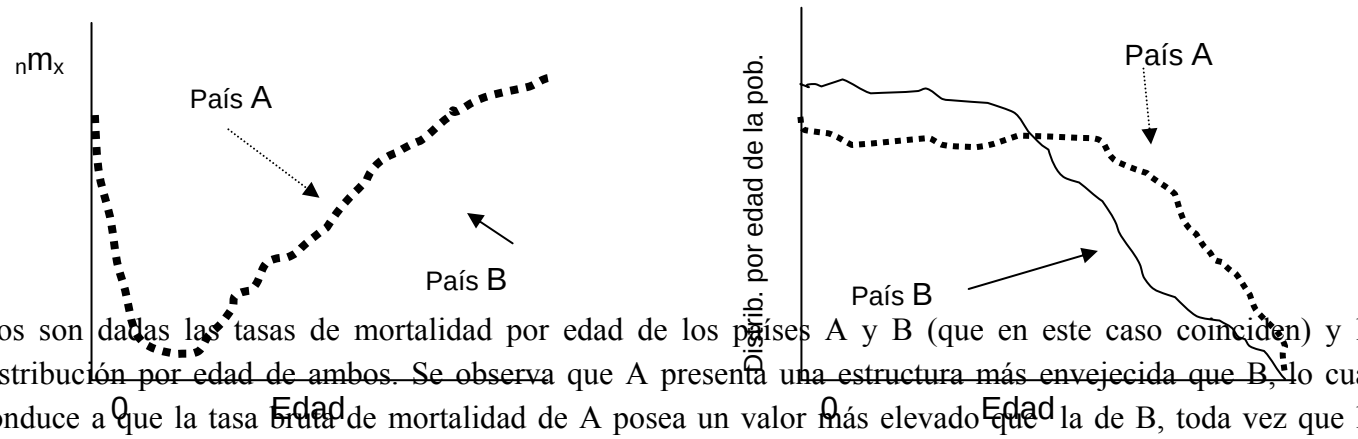
Mediante las fórmulas se percibe mejor esta problemática:

$$\text{Sea } REM_A = \frac{\sum {}_nm_x{}^A * {}_nN_x{}^A / N^A}{\sum {}_nm_x{}^C * {}_nN_x{}^A / N^A} \quad \text{y} \quad REM_B = \frac{\sum {}_nm_x{}^B * {}_nN_x{}^B / N^B}{\sum {}_nm_x{}^C * {}_nN_x{}^B / N^B}$$

Los denominadores de ambas REM representan tasas ajustadas del país C mediante tipificación directa, donde las estructuras por edades tipos corresponden a los países A y B respectivamente. Sería algo muy equivocado decir que ambas son comparables ya que salta a la vista que las distribuciones por edades de A y B casi con seguridad serán diferentes. Por otra parte, si nos detenemos a observar la composición de la REM, nos percatamos de que la única diferencia entre el numerador y el denominador es que en el primero aparecen las tasas específicas de A (B) y en el segundo las de C, por lo cual uno estaría tentado a pensar que la REM no contiene el efecto de la estructura por edades de A o de B, ya que habría una compensación entre numerador y denominador. Realmente esto no se puede afirmar categóricamente.

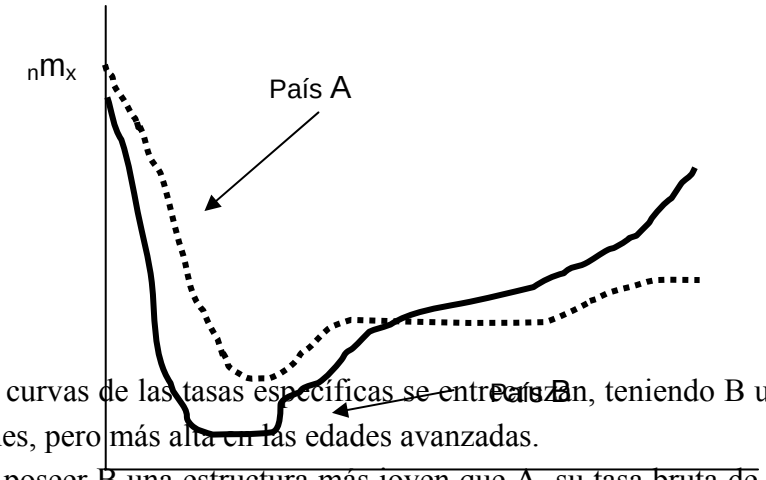
Como usos de la REM podemos enumerar aquellos estudios en que se desea medir brechas en materia de mortalidad, entre dos países, lo que se logra tomando un juego de tasas de mortalidad por edades patrón del país con mejores condiciones . Además, en casos de poblaciones pequeñas, para las cuales no contamos con tasas específicas por edades o si existen no son confiables.

Veamos ahora este problema con una nueva óptica: la tasa bruta no es más que un promedio ponderado (media para datos agrupados) de las tasas específicas de mortalidad por edad, donde los factores de ponderación o pesos lo constituyen las proporciones de población en cada grupo de edad. Esto hace que este promedio pueda variar en dependencia de las edades donde esté más concentrada la distribución de la población Veamos a continuación una ilustración de lo anterior.



Nos son dadas las tasas de mortalidad por edad de los países A y B (que en este caso coinciden) y la distribución por edad de ambos. Se observa que A presenta una estructura más envejecida que B, lo cual conduce a que la tasa bruta de mortalidad de A posea un valor más elevado que la de B, toda vez que la distribución de aquel está más concentrada en edades más altas y arrastra el promedio ponderado hacia los valores de las tasas específicas de las edades adultas y ancianas. Es decir, aún siendo los riesgos específicos iguales en ambos países (igual nivel de mortalidad), se concluiría que la mortalidad en A es más elevada que en B, lo cual no es cierto. La diferencia en estructura (pirámide) nos hace llegar a una conclusión falsa.

Supongamos ahora que existen diferencias en cuanto a los riesgos específicos de morir entre los dos países y que las tasas por edad corresponden a las señaladas en el próximo gráfico:



En este caso las curvas de las tasas específicas se entrecruzan, teniendo B una mortalidad más baja que A en las edades jóvenes, pero más alta en las edades avanzadas.

Por el hecho de poseer B una estructura más joven que A, su tasa bruta de mortalidad será menor pues en el tramo de las edades jóvenes los riesgos de muerte de B son inferiores.

Si pretendemos ajustar la tasa bruta de B, digamos, con la distribución por edad de A, entonces es muy probable que el resultado favorezca a este último (menor tasa para A), ya que su distribución al estar más concentrada en edades altas arrastraría el promedio ponderado de B hacia valores de tasas específicas, que son superiores a las de A.

El problema de la tipificación directa pude ser enfocado también mediante el cociente de las tasas específicas por edad del país en estudio y el patrón.

Sea $_n\lambda_x = \frac{_nm_x^A}{_nm_x^C}$, por despeje se obtiene que

$_nm_x^A = {}_n\lambda_x * {}_nm_x^C$,

y la tasa bruta de A ajustada con la distribución por edad de C resulta en

$TBM^{(A/C)} = \sum {}_n\lambda_x * {}_nm_x^C * {}_nN_x^C / N^C = \sum {}_n\lambda_x * {}_nD_x^C / N^C$,

Si hacemos el cociente entre la tasa de A ajustada y la tasa bruta de C, con el pretexto de comparar ambas, nos queda la siguiente expresión:

$$\frac{TBM^A}{TBM^C} = \frac{\sum {}_n\lambda_x * {}_nD_x^C}{\sum {}_nm_x^C * {}_nN_x^C} = \sum {}_n\lambda_x * (\frac{{}_nD_x^C}{D^C}) = \bar{\lambda};$$

que ha resultado en un promedio ponderado de los $_n\lambda_x$, donde los pesos están dados por la proporción por edad de las defunciones del país patrón C.

Con un enfoque similar se puede tratar la Razón Estandarizada de Mortalidad (REM); esto es:

la tasa bruta de mortalidad de A la podemos escribir como

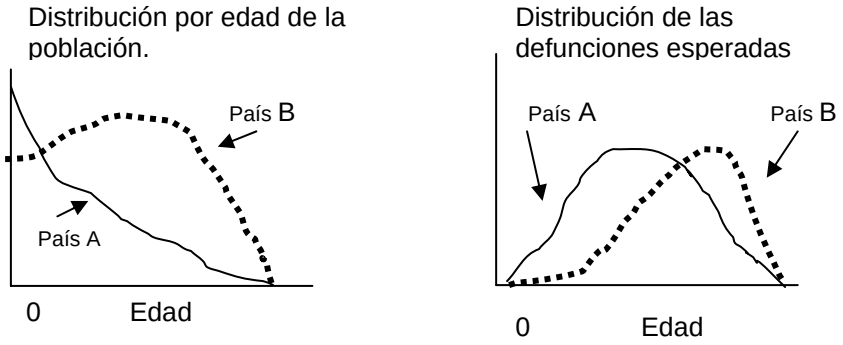
$TBM^A = \sum {}_n\lambda_x * {}_nm_x^C * {}_nN_x^A / N^A$

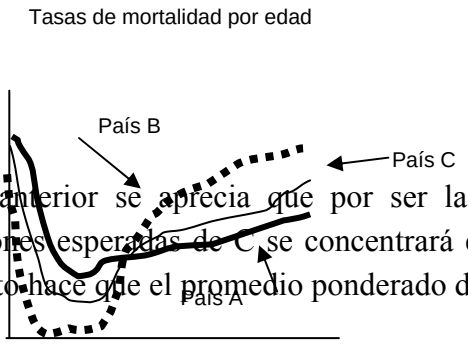
Luego la REM se escribiría como

$$REM = \frac{\sum {}_nm_x^{A} * {}_nN_x^{A}}{\sum {}_nm_x^{C} * {}_nN_x^{A}} = \frac{\sum {}_n\lambda_x * {}_nm_x^{C} * {}_nN_x^{A}}{D^{C/esperada}} = \frac{\sum {}_n\lambda_x * {}_nD_x^{C/Esperada}}{D^{C/Esperada}} = \sum {}_n\lambda_x * (\frac{{}_nD_x^{C/Esperada}}{D^{C/Esperada}}) = \bar{\lambda}$$

La REM es también un promedio ponderado de los cocientes de tasas específicas de mortalidad por edad; ahora los pesos o factores de ponderación están dados por la proporción de defunciones esperadas de C (con la distribución por edad de A). Es fácil advertir que las ponderaciones dependen implícitamente de las tasas específicas de C y la distribución por edad de A. Si se calcula nuevamente la REM para otro país B utilizando a C como patrón, entonces dichas ponderaciones dependerán de las tasas específicas de C y de la distribución por edad de B; resulta entonces que dos o más REM no deben compararse ya que el efecto de la distribución por edad no se controla.

Veamos un ejemplo:





De la información visual anterior se aprecia que por ser la estructura por edad de A muy joven, la distribución de las defunciones esperadas de C se concentrará en edades más jóvenes mientras que para B será en edades más altas; esto hace que el promedio ponderado de los cocientes de tasas específicas entre A y C tienda a ser mayor que 1

Ya que A supera en riesgo de muerte a C en ese tramo. Como B exhibe una estructura muy envejecida, entonces las defunciones esperadas de C en este caso se concentrarán en las edades avanzadas y también el λ promedio será mayor que 1 pues en ese tramo el riesgo de muerte en B es mayor.

Ahora ¿ que pasaría si A y B tuviesen ambos una estructura por edad muy envejecida ?. Resultaría que el $\lambda^{A/C}$ sería menor que 1 (mortalidad de A menor que la de C en ese tramo) y $\lambda^{B/C}$ mayor que la unidad (mortalidad de B superior a la de C) ya que en las edades altas las tasas de B superan a las de C y lógicamente el cociente es mayor que 1.

Resumen del capítulo

Debe tenerse presente por qué es necesario realizar el ajuste de las tasas brutas o crudas: controlar el efecto de la distribución de la variable de clasificación, que por lo general es la edad.

El método directo es preferible al indirecto. Para el primero debe disponerse de una distribución tipo y las tasas por edad conocidas; para el segundo debe tenerse un juego de tasas de mortalidad por edad tipo y conocer las tasas crudas a comparar.

Ejercicio propuesto:
Si Ud. desea comparar la mortalidad general de un área correspondiente a un consultorio médico con la del municipio al cual pertenece dicho consultorio, realizaría un

a) ajuste o tipificación directa, tomando como estándar cualquiera de las dos distribuciones.
b) ajuste indirecto tomando como tasas patrón las del municipio.

5. Otros Indicadores

5.1 Años de vida potenciales perdidos (AVPP).

Entre los diversos indicadores utilizados en el estudio de la mortalidad se cuenta con los AVPP. Está concebido para hacer una medición del tiempo perdido a causa de la mortalidad en una población. Esto es , la cantidad de años que dejan de vivir las personas con respecto a lo que hubieran vivido de no haber muerto prematuramente.
Comencemos la explicación con el siguiente ejemplo:

Grupo de edades	Edad media del intervalo (1)	Defunciones (2)	AVPP (3) = 65- (1)	AVPP en cada grupo (4) = (3) x (2)
<1	0.5	1109	65-0.5=64.5	71530.5
1-4	3	391	65-3=62	24242
5-14	10	540	65-10=55	29700
15-49	32.5	10109	65-32.5=32.5	328542.5
50-64	57.5	12601	65-57.5=7.5	94507.5
65 y más		54883		
Total		79654		548522.5

El cuadro anterior resume información de Cuba para 1996. Están dadas la edad media de cada intervalo de edades, las defunciones por grupos de edades, los AVPP para un individuo fallecido y finalmente el cálculo en cada grupo de edades. Describiremos a continuación cada columna de la tabla.

En la columna (1) se ha insertado la edad media del intervalo de edades, que se calcula como la semi -suma de los extremos reales de cada uno, esto es, por ejemplo para el grupo 5-14, procedemos 5+15 = 20 entre 2 que es 10, toda vez que el límite real superior de ese grupo es 15 y no 14 (la edad cumplida 14 llega casi hasta 15, pero sin tomarlo). Observe que para el intervalo final abierto (65y+) no hemos puesto la edad media ya que no conocemos la amplitud de éste y por ende no se incluye en los cálculos. La columna (2) refleja las defunciones de cada grupo y pienso que no necesita explicación.
La columna (3) es la diferencia entre 65 (edad que se ha tomado como referencia) y la edad media de cada intervalo. Si se considera la edad de 65 años como el límite establecido para el cálculo, entonces esta

columna expresa la cantidad de años perdidos por un individuo que ha fallecido en determinado grupo de edades. Ubiquémonos en el intervalo 1-4 al cual le corresponde una diferencia de 62 en la columna (3); esto significa que cada persona fallecida con edad contenida en dicho intervalo, deja de vivir 62 años. En la columna (4) este valor se ha multiplicado por la cantidad de defunciones de ese grupo con el propósito de estimar el número de años perdidos para todos los fallecidos en ese intervalo. Posteriormente se han sumado los valores de los diferentes grupos lo que resultó en el total de Años Potenciales Perdidos (548 522.5).
El límite superior usado (65 años) pudo haber sido otro, como por ejemplo la esperanza de vida al nacimiento u otra edad, digamos 80 años, siempre que ésta sea igual o mayor que las del último grupo de edades considerado. El investigador escoge el límite según sus intereses aunque en ocasiones está condicionado por la disposición de los datos.

Los AVPP constituyen una alternativa en la manera de medir el impacto de la mortalidad en la población diferente a como lo hacen las tasas, que dan como resultado una frecuencia determinada. El AVPP da sus resultados en años- persona.

Es bueno dejar claro que esta medida depende de la estructura por edad de la población toda vez que las defunciones en cada grupo están determinadas por la tasa específica y la población del mismo; por tanto se pudiera aplicar la técnica de ajuste o estandarización directa para refinar este indicador.
En términos analíticos, el AVPP puede escribirse de la siguiente manera:

$$AVPP = \sum (K - \bar{X}) * {}_nD_x = \sum (K - \bar{X}) * {}_nm_x * {}_nN_x$$

donde $\bar{X}$ y K representan respectivamente, la edad media del intervalo de edades y el límite escogido para señalar la prematuridad de la muerte.
En la segunda igualdad se ha escrito la misma fórmula pero ahora se han expresado las defunciones como el producto de la tasa y la población de edad cumplida del grupo, con el objetivo de resaltar la posible necesidad de tipificar o controlar la distribución por edad de la población.
Es factible calcular una tasa de AVPP con solo dividir éste entre la población media del año correspondiente al segmento de edad considerado; siguiendo el ejemplo anterior, si se sabe que la población media de Cuba, menor de 65 años en 1996 fue de 10 005 800 habitantes, entonces la tasa de AVPP para este segmento de población sería:

Tasa de AVPP $= \frac{548522.5}{10005800} * 1000 = 54.82$ años potenciales perdidos por cada 1000 personas de esa población media.

Los AVPP pueden calcularse para diferentes segmentos de la población como sería el tramo de edades que va desde 15 a 50 años y no necesariamente hay que comenzar en la edad cero; además pueden obtenerse por sexo y causa de muerte.

Veamos el siguiente ejemplo:

Grupo de Edad	Defunciones	por AVPP	por AVPP en cada

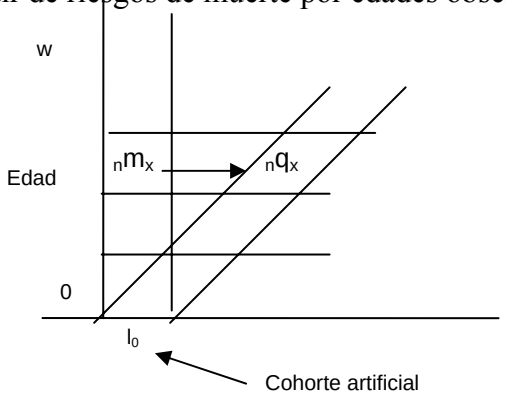
edades	media	accidentes		indivi-duo	grupo	
		1980	1996		1980	1996
<1 año	0.5	84	45	64.5	5418	2902.5
1-4	3	143	119	62	8866	7378
5-14	10	412	242	55	22660	13310
15-49	32.5	1796	2062	32.5	58370	67015
50-64	57.5	415	633	7.5	3112.5	4747.5
Total		2850	3101		98426.5	95353

En la tabla precedente se ha calculado el AVPP para la causa de muerte Accidentes, resultando un total de años de vida potenciales perdidos por esta causa mayor en 1980 que en 1996 . Las respectivas tasas de AVPP fueron 10.06 y 8.66 por mil personas . Los resultados anteriores son concordantes con el hecho de que la edad media de las muertes por Accidentes en 1980 fue 37.7 y en 1996 de 39.6, por lo que parece lógico que la pérdida de años de vida sea mayor para 1980.

5.2 La Tabla de Mortalidad o de Vida.

La Tabla de Mortalidad, también conocida como tabla de vida es un instrumento muy útil para hacer análisis de la mortalidad, pues ofrece una descripción muy detallada de la mortalidad según la edad. Su aspecto exterior es un cuadro numérico brinda diferentes índices entre los que se cuenta la esperanza de vida. Para su construcción se parte de las tasas específicas de mortalidad por edades observadas en determinado período calendario, que generalmente es un trienio, o sea una información transversal, y el cálculo se realiza para cada sexo por separado, integrándose luego para obtener una tabla para los dos sexos juntos.

El paso previo más importante es la transformación de las tasas específicas de mortalidad en probabilidades de muerte; luego éstas se aplican a una cohorte hipotética de nacimientos de tamaño l_0, con lo que se simula un análisis longitudinal a partir de riesgos de muerte por edades observados transversalmente.



Las componentes que conforman la tabla de vida se denominan funciones. Las más importantes son:

- Probabilidades de muerte entre dos edades exactas, (${}_nq_x$).
- Sobrevivientes a edades exactas, (l_x).
- Defunciones entre dos edades exactas, (${}_nd_x$).
- Tiempo vivido entre dos edades exactas, (${}_nL_x$).
- Tiempo vivido a partir de la edad exacta x, (T_x).
- Esperanza de vida a una edad exacta x, (e_x^0).

.

A continuación presentamos la tabla de vida del sexo femenino de Cuba, en el período 1986-87.

Grupo de edades (x,x+n-1)	Probab. de muerte $_nq_x$	Sobrevivientes l_x	Defunciones $_nd_x$	Tiempo vivido entre x y x+n $_nL_x$	Tiempo viv. a partir de x T_x	esperanza de vida a edad x e_x^o
0	0.011490	100000	1149	99079	7634345	76.34
1-4	0.002833	98851	280	394719	7535266	76.23
5-9	0.001775	98571	175	492381	7140547	72.44
10-14	0.001738	98396	171	491628	6648166	67.57
15-19	0.004429	98225	435	490101	6156538	62.68
20-24	0.004837	97790	473	487774	5666437	57.94
25-29	0.005364	97317	522	485291	5178663	53.21
30-34	0.006013	96795	582	482559	4693372	48.49
35-39	0.007546	96213	726	479326	4210813	43.77
40-44	0.010609	95487	1013	475067	3731487	39.08
45-49	0.016512	94474	1560	468736	3256420	34.47
50-54	0.025023	92914	2325	459105	2787684	30.00
55-59	0.036958	90589	3348	445045	2328579	25.70
60-64	0.054309	87241	4738	425008	1883534	21.59
65-69	0.081791	82503	6748	396646	1458526	17.68
70-74	0.131978	75755	9998	355379	1061880	14.02
75-79	0.214578	65757	14110	295027	706501	10.74
80-84	0.317443	51647	16395	217421	411474	7.97
85 y más	1.00000	35252	35252	94053	194053	5.50

Probabilidad de morir entre dos edades exactas ($_nq_x$).

Representa la probabilidad que tiene un individuo de la cohorte de morir entre las edades exactas x y x+n. En el último grupo su valor siempre es igual a 1, ya que se sobreentiende que llegado a una edad determinada la generación se extinguirá.

Sobrevivientes a edades exactas (l_x) y defunciones ($_nd_x$)

A medida que la cohorte ficticia transita por los diferentes grupos de edades (proceso de envejecimiento), va perdiendo efectivos a causa de la mortalidad, produciéndose las correspondientes defunciones en cada grupo y un número de sobrevivientes en cada edad exacta al final del intervalo de edades. Nótese que para la edad exacta 85 años, el número de sobrevivientes es igual a la cantidad de muertes del intervalo abierto.

Tiempo vivido entre dos edades exactas x y x+n, ($_nL_x$) y a partir de x, (T_x).

Cada miembro de la cohorte vive cierto número de años dentro de cada intervalo de edades. Si la persona no fallece en un grupo dado, vivirá un número de años en el mismo igual a la amplitud de éste; en caso de fallecer dentro del grupo, también vivirá cierta cantidad de años, aunque menor que la amplitud de éste. A su vez, la cantidad de años vividos por los integrantes de la generación se puede contar desde una edad exacta x, hasta el final de la vida (T_x). En el grupo abierto final de la tabla se observa que ambas medidas coinciden.

Esperanza de vida a la edad exacta x (e_x^o)

Es el número de años que en promedio vivirá cada miembro de la cohorte, a partir de la edad exacta x. Esto presupone que el individuo tiene que llegar vivo a esa edad, para los que fallecen antes de x, por supuesto, esta medida no cuenta.

Se destaca por su importancia la esperanza de vida a la edad exacta 0 año o sea en el instante del nacimiento. Como en ese momento todos los miembros de la cohorte están vivos, esta medida es aplicable a todos ellos. Nos dice el número de años que en promedio vivirá cada miembro de la generación a partir de su nacimiento, bajo el supuesto que se someten a los riesgos de muerte por edad observados en un momento. Esto último es muy importante entenderlo bien.
La esperanza de vida al nacimiento o a cualquier otra edad, no es el tiempo de vida real de las personas que nacen en determinado año o llegan a determinada edad, sino que es el tiempo de vida de las personas sometidas a las condiciones de mortalidad observadas en un momento. Es útil para medir el impacto de las condiciones de mortalidad de un momento. Se espera que con el transcurso del tiempo, los riesgos de muerte cambien, generalmente hacia la mejoría, por lo que los individuos que nacen en determinado año vivirán sus vidas sometidos a riesgos por edad cambiantes en el tiempo, luego su tiempo real de vida será mayor que el calculado mediante la tabla de vida de momento.

Por otra parte, la cohorte inicial con la que se construye la tabla no está constituida por los nacidos vivos de un año o período de tiempo determinado, sino por un conjunto de nacidos vivos con propiedades similares, cuya identidad no interesa y que no tienen por qué pertenecer a un calendario dado; lo que sí está bien ligado al tiempo son los riesgos de muerte (tasas de mortalidad por edades del período calendario para el cual se confecciona la tabla).
Resumen del capítulo:
Los AVPP miden la pérdida ocasionada por el efecto de la mortalidad en términos de años- personas. La tasa de AVPP, sin embargo, da el número de años dejados de vivir por cada 1000 habitantes. Es útil para realizar análisis de la mortalidad de forma diferente a como se hace con las tasas.
Por otra parte, la tabla de mortalidad es un instrumento valiosísimo; el mismo ofrece un indicador de suprema importancia, a saber, la esperanza de vida a diferentes edades, en especial al nacimiento. Dicho indicador resume las condiciones de mortalidad para un momento determinado y no está afectado por la influencia del efecto de la distribución de la edad cuando se comparan varios países.
Ejercicio propuesto:
¿Por qué la esperanza de vida al nacimiento no mide el tiempo real de vida de las personas?.
6. FECUNDIDAD

Antes de adentrarnos en el estudio de la fecundidad, es conveniente dejar claros algunos conceptos y definiciones.

Fertilidad

Se refiere a la capacidad fisiológica de una mujer, hombre o pareja para producir un nacido vivo. No necesariamente tiene que producirse el nacido vivo, es una propiedad potencial. Una mujer puede que sea fértil pero no fecunda, en dependencia de que tenga la capacidad fisiológica para ello pero no haya tenido hijos.

Fecundidad

Puede verse como la procreación real de un individuo, pareja o grupo. La fecundidad significa un paso posterior a la fertilidad. Es necesario ser fértil para luego ser fecundo, es decir, producir nacidos vivos.

Natalidad

Es la fecundidad pero vista de forma más global a nivel de la población, es decir, la producción de nacidos vivos a nivel poblacional, los cuales constituyen un elemento importante en el cambio poblacional.

La fecundidad es una de las variables demográficas más importantes como ya se ha apuntado. Su comportamiento depende, a su vez, de variables de corte socio- económico como el nivel de desarrollo del país, nivel de instrucción de la pareja, ingreso, aspectos culturales como las tradiciones de la región o país. Existe otra serie de variables que se ha dado en llamar "intermedias", más cercanas a la fecundidad y que dependen de las anteriores en una secuencia más o menos causal: la edad al casarse o unirse, frecuencia del coito en las parejas, conocimiento y uso de anticonceptivos, ideales en cuanto al tamaño de la familia, etc.

Entre las variables demográficas más importantes también se ha citado a la mortalidad, sin embargo ésta ostenta algunas diferencias con la fecundidad que deben reconocerse explícitamente:

- La fecundidad es el aspecto positivo de la vida, es decir, aporta individuos a la sociedad, mientras que la mortalidad es contraria o sea merma a la sociedad.
- La mortalidad es un evento irrepetible, sin embargo la fecundidad no. Una mujer o una pareja pueden ser varias veces fecunda.
- La mortalidad es un evento que puede acontecer a una persona de cualquier edad, sin embargo la fecundidad tiene definido su período fértil tanto para el hombre como para la mujer.

Veamos ahora algunas medidas muy usadas para medir la fecundidad.

Tasa bruta de natalidad. Como se ha visto con anterioridad, es la frecuencia de nacidos vivos en la población durante un período de un año.

$$TBN = \frac{B^t}{\overline{N}} * 1000$$

Esta medida aunque es fácil de calcular, ya que no demanda información sofisticada, tiene el inconveniente de que en el denominador incluye a la totalidad de la población, donde existen personas que no están a riesgo de procrear, como son los niños y los ancianos.

Una medida que aporta un grado de refinamiento mayor es la Tasa de Fecundidad General (TFG), que sólo contempla en el denominador a las mujeres en edades fértiles (desde 15 a 49 años); la población usada es la que está expuesta al riesgo de concebir. Aunque la fecundidad se puede medir por pareja y también por los hombres, es preferible hacerlo a través de las mujeres, ya que es más fácil registrarla.

En 1996 Cuba exhibió una $TFG = \frac{140276}{3011616} = 46.58$ nacidos vivos por cada mil mujeres en edad fértil.

No obstante este índice mejora mucho la medición comparativamente con la TBN, internamente dentro del grupo de 15 - 49 años pueden darse algunas distorsiones debido a la diferencia que puede existir en la distribución por edades dentro de dicho grupo, entre distintas poblaciones.

Nosotros podemos definir medidas más específicas en cuanto a la edad. Así tendremos las tasas específicas de fecundidad por edades que permiten medir el riesgo de producir un nacido vivo en una edad o grupo de edades. Se denotan como

$$ {}_nf_x = \frac{nBx}{\overline{nNx}^{fem}} $$

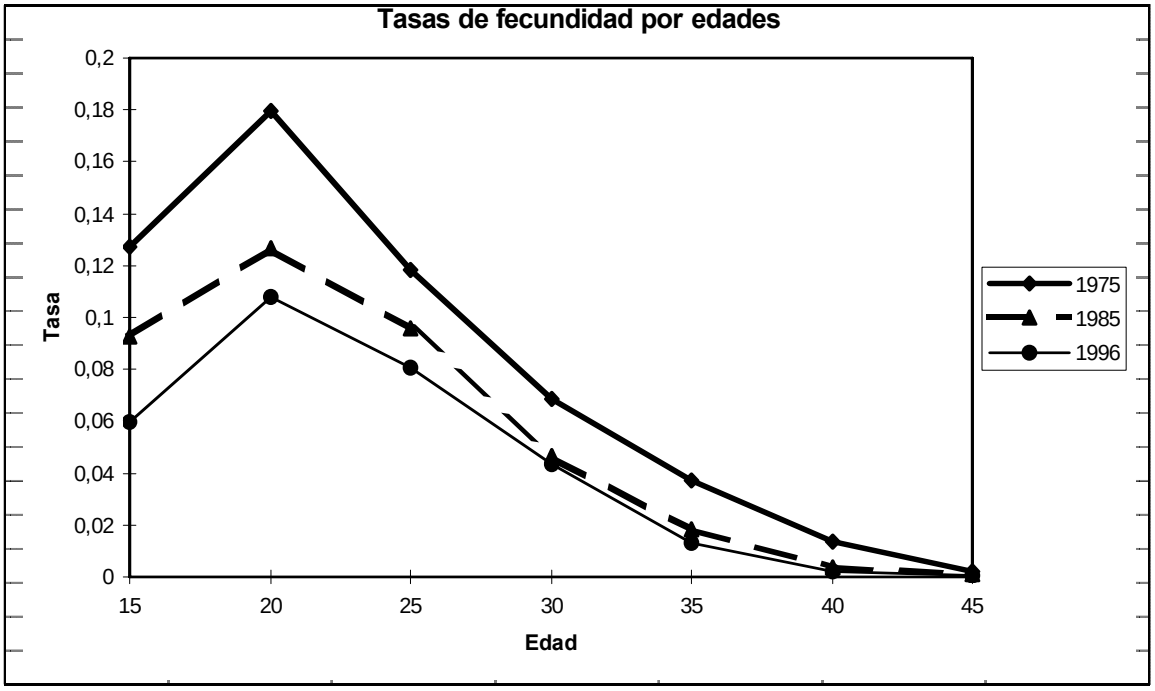
es decir, el cociente entre los nacidos vivos correspondientes a mujeres de determinado grupo de edades y la población media femenina de dicho grupo. No debe confundirse el factor de separación con la tasa de fecundidad ya que tienen la misma notación.

En 1975 la tasa específica de fecundidad del grupo 15-19 años fue de 127.3 nacidos vivos por mil mujeres de ese grupo de edades, mientras que en 1996 era sólo de 59.9.

Tasas de Fecundidad por edad y años seleccionados

Edad de la madre	1975	1980	1985	1990	1996
15-19	0.1273	0.0863	0.0929	0.0775	0.0599
20-24	0.1796	0.1168	0.1268	0.1139	0.1076
25-29	0.1183	0.0709	0.0957	0.0974	0.0804
30-34	0.0688	0.0374	0.0465	0.0561	0.0436
35-39	0.0374	0.0162	0.0185	0.0175	0.0132
40-44	0.0135	0.0046	0.0039	0.0033	0.0023
45-49	0.0023	0.0018	0.0012	0.0003	0.0004
Total (TGF)	2.74	1.67	1.93	1.83	1.537
TBR	1.33	0.81	0.94	0.89	0.75

Cuando las tasas de fecundidad por edades tienen su valor máximo en el grupo de 20-24 años, se dice que la fecundidad es de cúspide temprana; si es en el de 25-29, se denomina de cúspide tardía y cuando el gráfico tiene aspecto de meseta (los valores de los grupos 20-24 y 25-29 son más o menos similares) de cúspide dilatada.



El aspecto del gráfico de las tasas por edades tiene gran importancia ya que nos informa del comienzo de formación de la familia en distintas sociedades. Una curva de cúspide tardía nos dice que las mujeres han pospuesto su maternidad debido a diferentes factores que podrían enumerarse como : una edad al casarse mayor o que esperan completar su formación profesional y/o afianzar su situación económica y laboral antes de tener sus hijos, etc. Una cúspide dilatada podría deberse a un esparcimiento de los hijos por diversos motivos; la cúspide temprana denota una mayor precocidad en el completamiento del número de hijos deseados. En el gráfico anterior puede observarse el cambio acaecido entre la situación del año 1975 y 1996: en éste último el grupo de 25-29 años ha adquirido mayor importancia por lo que se presume una tendencia a una fecundidad de cúspide dilatada.

Las tasas de fecundidad por edades nos brindan gran detalle sobre el comportamiento de este fenómeno, pero sería bueno contar con otras medidas de resumen que a su vez sean refinadas en el sentido que no estén afectadas por la estructura por edad de la población. Con ese propósito definimos la Tasa Global de fecundidad (TGF).

La tasa global de fecundidad es el número de hijos que en promedio tendrán cada una de las mujeres de una cohorte hipotética al final de su período reproductivo, si tuvieran sus hijos de acuerdo a las tasas de fecundidad por edades observadas en un momento o año y que no estuvieran sometidas al riesgo de muerte hasta tanto no hayan finalizado el periodo reproductivo.

Las tasas por edades se aplican a la cohorte de mujeres. En un grupo de edades (x, x+n), la tasa ${}_nf_x$ se aplica n veces; por ejemplo, supongamos que la cohorte de mujeres comienza su vida reproductiva a la edad exacta 15 años, entonces la tasa ${}_5f_{15}$ se aplicará a las mujeres de edad cumplida 15 años, luego a las de edad cumplida 16 años, así hasta las de edad cumplida 19 años, toda vez que ${}_nf_x$ es una tasa promedio para cada

edad simple dentro del intervalo. Luego pasamos al grupo 20-24 años, donde se sigue la misma práctica. Se va transitando por los diversos grupos hasta llegar al último, 45-49, donde finaliza el cálculo.

En términos de fórmulas, la TGF se puede expresar como:

$$TGF = n * \sum nfx ,$$

donde la sumatoria se extiende desde la edad 15 años hasta 45, es decir, primero la x se pone igual a 15, luego igual a 15+ n, 15+2n, ..., hasta la edad exacta 45 años, o sea dando saltos de n años . Si fuera n=5, tendríamos grupos quinquenales de edades y los saltos serían de 5 en 5. En la tabla anterior en la fila total se dan los valores para la TGF en los años señalados. Note como el valor ha descendido de algo más de 2 hijos por mujer en 1975 a alrededor de uno y medio hijo en 1996.

Existe un importante supuesto implícito cuando calculamos la TGF y es que esas mujeres de la cohorte hipotética no estarán sometidas a los riesgos de la mortalidad hasta después de haber finalizado su período fértil.

Aquí hemos aplicado la experiencia de un momento (información transversal) a una cohorte (estudio longitudinal).

6.1 Tasa Bruta reproducción (TBR)

Otra medida refinada de la fecundidad lo constituye la tasa bruta de reproducción. Muy parecida a la TGF, sólo se diferencia de ésta en que mide la cantidad de hijos hembras, mientras aquella lo hacía para el total de hijos. Más que una medida de fecundidad se considera una medida de reproductibilidad: una mujer se reproduce a sí misma de forma simple cuando tiene una hija, si tiene más de una hija se reproduce de manera ampliada.

En fórmula, la TBR es $TBR = K * TGF$

donde K es la proporción de nacimientos femeninos del total de éstos.

Conocido el índice de masculinidad al nacimiento podemos calcular K de la siguiente manera, sea B el total de nacidos vivos de todos los sexos juntos, sean B^F y B^M las cantidades de nacidos vivos correspondientes al sexo femenino y masculino respectivamente.

El índice de masculinidad al nacimiento (edad exacta 0 año) será entonces

$$I_0 = \frac{B^M}{B^F}$$

fácilmente se encuentra una relación para K..

Despejando los nacidos vivos masculinos en la fórmula anterior queda que el total de nacidos vivos es $B^T = B^M + B^F = I_0 * B^F + B^F = B^F (1 + I_0)$, luego la proporción de nacimientos femeninos en el total de nacidos vivos de los dos sexos reunidos es

$$K = \frac{B^F}{B^T} = \frac{B^F}{B^F(1+I_0)} = \frac{1}{(1+I_0)}$$

Por ejemplo, si I_0 es igual a 1.05 , entonces $K = \frac{1}{(1+1.05)} = 0.4878$

La TBR también tiene implícito el supuesto de que las mujeres están inmunes a la mortalidad hasta tanto no hayan rebasado el período reproductivo.

6.2 Tasa Neta de Reproducción

Su diferencia con la TBR consiste en que toma en cuenta el hecho de que las mujeres estarán sometidas a los riesgos de muerte durante su vida. Es el número de hijos hembras que tendrá cada mujer de una cohorte hipotética si estuviera sometida a los riesgos de fecundidad y mortalidad observados en un momento. Siempre es menor que la TBR pues precisamente toma en cuenta que la mujer estará sometida a los riesgos de muerte desde su nacimiento. Es una medida de reproducción de la población más rigurosa que la TBR. Cuando es mayor que 1 significa que una generación de mujeres garantiza una determinada cantidad de hijas que la reemplazarán. Si es igual a 1, entonces una madre es reemplazada exactamente por una hija. Cuando es menor que 1 no se garantiza el reemplazo de una generación por otra. Si en un país la TNR se mantiene durante mucho tiempo por debajo de 1, la población comenzará a decrecer a partir de un momento dado. Si por otra parte la TNR se mantiene igual a 1 (nivel de reemplazo) se observará después de varias décadas que la población dejará de crecer y mantendrá un número fijo de habitantes, lo que se conoce como población estacionaria.

Si por el contrario la TNR estuviera por debajo de 1 solamente algunos años, no necesariamente decrecerá su monto pues los nacimientos de épocas pasadas pueden haber creado una concentración de mujeres en edades reproductivas que harán que los nacidos vivos actuales superen a las defunciones y por ende no se dé un decrecimiento de la población.

Para calcular la TNR necesitamos las tasas de fecundidad por edades y una tabla de mortalidad femenina.

Como, $TNR = n * K * \sum {}_nf_x * \frac{{}_nL_x}{n * l_0} = K * \sum {}_nf_x * \frac{{}_nL_x}{l_0}$, a partir datos de la tabla anterior podemos intentar su calculo. En esta expresión la sumatoria toma en cuenta las edades del período reproductivo y el término $\frac{{}_nL_x}{n * l_0}$ nos indica la probabilidad de sobrevivir una mujer desde el nacimiento hasta el grupo de edad cumplida (x, x+n); los restantes elementos de la fórmula son comunes a la TBR.

Veamos a continuación un ejemplo del calculo del valor de la TNR basado en información de Cuba.

Edad de la madre	${}_nf_x$ (1987)	${}_nL_x$ (1986-87)	$\sum {}_nf_x * {}_nL_x$
15-19	0.0815	486914	39683.491
20-24	0.1169	483853	56562.4157
25-29	0.0925	480138	44412.765
30-34	0.0499	475854	23745.1146
35-39	0.0183	470835	8616.2805
40-44	0.0033	464728	1533.6024
45-49	0.0011	456168	501.7848

Total	0.3635	175055.454
TBR	0.8866	

De donde, TNR = 0.4878* (175055.454/100000) = 0.8539

En el cuadro se ilustra el cálculo de la TNR usando las tasas de fecundidad del año 1987 y la tabla de vida femenina del período 1986-87. Nótese que tomamos los valores de la función ${}_nL_x$ desde el grupo 15-19 hasta 45-49 años. Luego multiplicamos los valores de ${}_nf_x$ y ${}_nL_x$ y obtenemos el total para todos los grupos del período reproductivo. Finalmente multiplicamos por K (0.4878) y dividimos por 100000 (l_0). En el propio cuadro se da el valor de la TBR, el cual es casi un 4% mayor que el de la TNR, debido a la ausencia de la mortalidad.

Resumen del capítulo:

Debe reconocerse que la fecundidad es una de las variables más importantes del crecimiento poblacional.

Deben conocerse las medidas más importantes usadas, yendo desde las más simples hasta las más refinadas, a decir, tasa bruta de natalidad, tasa de fecundidad general, tasa global de fecundidad, tasa bruta de reproducción y por último la tasa neta de reproducción. Estas dos últimas son también medidas de la reproducibilidad de la población, es decir, dan una medida de cómo una generación es remplazada por otra. Es importante conocer las consecuencias que se derivan cuando el valor de la tasa neta de reproducción es mayor, igual o menor que 1 (nivel de reemplazo).

Ejercicios propuestos:

1) ¿Qué sería necesario para calcular un indicador de fecundidad (TBN, TGF, TBR, etc) en hombres, en lugar de en mujeres?
2) Un país exhibe una tasa bruta de reproducción igual a 1.09 hijas por mujer durante varias décadas. ¿Quiere esto decir que el reemplazo generacional ha estado garantizado?
3) ¿Al aumentar el valor de la tasa neta de reproducción necesariamente debe aumentar el de la tasa bruta de reproducción?

7. Migraciones y distribución espacial

Como ya se apuntó, por lo general una población está asociada a un área geográfica determinada. Entre los diferentes países y regiones se producen movimientos de personas, aún más, dentro de un mismo país los individuos se trasladan de un lugar a otro.

Por otra parte, los países se caracterizan por subdivisiones como provincias o estados, y dentro de éstos en municipios, condados y demás. Las grandes ciudades concentran en su seno altas proporciones de habitantes, sobre todo obreros que viven en barrios cercanos a zonas industriales; las ciudades menores y pueblos son más numerosas con menor monto poblacional, luego vienen las aldeas, hasta llegar a zonas totalmente rurales, donde la población es muy escasa. Es típico ver en muchos de nuestros países cómo la población rural se concentra alrededor de volcanes o cerca de las cuencas de los ríos y lagos, en busca de tierras fértiles bien irrigadas, y pesca favorable.

Al fenómeno que consiste en el traslado de grupos poblacionales de una región a otra se le conoce como migración. A su vez, se le denomina interna, cuando ese movimiento ocurre dentro de un mismo país y externa, cuando se da entre países.

La manera en que un territorio es ocupado por sus habitantes se le denomina distribución espacial de la población. Es decir, la forma en que están situados los diferentes asentamientos humanos en determinada área geográfica.

Tanto la migración como la distribución espacial de la población tienen enorme importancia en la salud pública. Por ejemplo, los problemas de salud que emergen en las grandes zonas industriales de las ciudades debidos a la contaminación, el stress, etc (infarto, dolencias respiratorias), por lo general no son los mismos que se observan en regiones rurales donde predominan las actividades agrícolas.

De igual manera, cuando personas de otros países ingresan a uno nuevo pueden trasladar problemas de salud (enfermedades y epidemias) que no existen o que están controlados en el país de destino y así poner en tensión el sistema de salud local.

7.1 Algunas medidas que caracterizan la migración.

Antes de cualquier intento por medir el impacto de la migración sobre una comunidad, es necesario hacer algunas precisiones. En este sentido, es preciso definir lo que se entenderá por migrante, y dentro de esto, diferenciar entre distintas categorías como pueden ser: emigrante (inmigrante) temporal y definitivo. Generalmente, las legislaciones de los países tienen definido con rigor estas categorías.

Partiendo de estas definiciones y de un registro de migraciones, se pueden calcular diversas medidas del fenómeno.

$$\text{Tasa de emigración} = \frac{\#\text{ de emigrantes en un período de tiempo (un año)}}{\text{Población media del período}} \text{ x } 1000$$

Mide la frecuencia de personas que emigran durante un período de tiempo, que generalmente se toma como un año, con respecto a la población del país de origen, es decir, la población a riesgo de emigrar.

$$\text{Tasa de inmigración} = \frac{\#\text{ de inmigrantes en un período de tiempo (un año)}}{\text{Población media del período}} \text{ x } 1000$$

Mide la frecuencia de personas que entran al país respecto a la población de destino. En algunas ocasiones esto ha sido tomado como una deficiencia metodológica de este indicador, pues el denominador de esta tasa no es precisamente la población a riesgo de ser inmigrante.

$$\text{Tasa de saldo migratorio} = \frac{\text{Saldo migratorio (SM)}}{\text{Población media}} \text{ x } 1000$$

Mide la frecuencia del saldo migratorio respecto a la población del país o región, en un año. Su valor puede ser negativo en caso que el saldo lo sea. Su valor es la diferencia entre la tasa de inmigración y emigración.

Estos índices pueden calcularse por edad, sexo, provincia, municipio; para inmigrante (emigrante) temporal o permanente, interno o internacional, etc.

Ejemplo:

La provincia de Pinar del Río en el año 1998 tenía una población 729 109. El número de inmigrantes y emigrantes en dicho año fue de 2900 y 3239 respectivamente.

$$\text{Tasa de inmigración} = \frac{2900}{729109} \cdot 1000 = 3.9$$

$$\text{Tasa de emigración} = \frac{3239}{729109} \cdot 1000 = 4.4$$

$$\text{Tasa de saldo migratorio} = \frac{2900 - 3239}{729109} = -0.5$$

Durante ese año entraron casi 4 personas por cada mil y salieron del territorio algo más de 4, lo que resumido significa que la provincia es emisora de población con una pérdida de 5 individuos por cada 10 000 personas.

7.2 Algunas medidas que caracterizan la distribución espacial de la población.

En cuanto a la distribución espacial de la población, los indicadores más comúnmente usados para describirla son:

- Densidad de la población

Mide cuántos habitantes hay, en promedio, por cada kilómetro cuadrado de superficie, en determinado momento. Puede calcularse por: cada zona de residencia (urbana y rural), provincia, ciudades, municipios, etc.

$$\text{Densidad de población} = \frac{\text{Total de población}}{\text{Area del territorio donde está asentada la población (en km}^2\text{)}}$$

Ejemplo:

La provincia de Pinar del Río tiene una extensión de 10 924.6 km^2, el 30 de junio de 1998 su densidad era de

$$\frac{729109}{10924.6} = 66.7 \text{ hab./km}^2.$$

- Distribución porcentual de la población según zona de residencia urbana y rural.

Se clasifica a la población según resida en la zona urbana o en la rural. Se puede dar para todo el país o según provincia, municipio, etc.

Ejemplo:

En la siguiente tabla se representa la distribución de la población según provincia y zona de residencia.

Provincia	% de pob. urbana	% de pob. rural
Pinar del Río	63,9	36,1
La Habana	78,4	21,6
C de La Habana	100,0	-
Matanzas	80,3	19,7
Villa Clara	77,5	22,5

Cienfuegos	80,7	19,3
Sancti Spiritus	69,7	30,3
Ciego de Avila	74,6	25,4
Camagüey	75,0	25,0
Las Tunas	58,8	41,2
Holguín	59,0	41,0
Granma	57,6	42,4
Santiago de Cuba	70,2	29,8
Guantánamo	59,6	40,4
Isla de la Juventud	87,5	12,5
Cuba	75,3	24,7

- Distribución porcentual de la población según el tamaño de los asentamientos.

La población de un país se clasifica según el tamaño de los asentamientos poblacionales, también puede inscribirse la cantidad de los mismos. Para ello es necesario que los asentamientos estén definidos administrativamente y cartográficamente con vistas a hacer una clasificación correcta. Generalmente esta información está disponible a partir de los censos de población.

Ejemplo:

Mediante una tabla similar a la siguiente, se puede ofrecer una valiosa información:

Tamaño del asentamiento	% de población (número de asentamientos)
Un millón de habitantes o más	16.7 (1)
500 000 – 999 999	1.6 (1)
100 000 – 499 999	3.9 (2)
50 000- 99 999	5.6 (3)
20 000- 49 999	15.3 (5)
5 000- 19 999	24.8 (23)
Menos de 5 000	32.1 (53)
Total	100.0 (88)

Estos datos son hipotéticos, no obstante se percibe una situación que presumiblemente corresponde a un país con alto grado de ruralidad, con una población concentrada mayoritariamente en asentamientos pequeños.

Resumen del capítulo:

El estudiante debe tener claros los conceptos de inmigrante y emigrante, las formas de medir el impacto de la migración mediante las tasas, así como las diferentes formas de describir la distribución espacial de la población.

Ejercicio propuesto:

A su juicio, habrá diferencias en cuanto a los problemas de salud que puedan generarse en dos países, uno de los cuales presenta una distribución espacial donde predominan los pequeños asentamientos y por el contrario, en el otro hay una alta concentración de la población en grandes ciudades.

Bibliografía Consultada

1. Lotka, A. Teoría de las asociaciones biológicas. Celade, Santiago de Chile,1976.
2. La población de Cuba. Centro de estudios demográficos. Editorial de Ciencias Sociales, La Habana, 1976.
3. Elizaga, Juan. C. Métodos Demográficos para el estudio de la mortalidad. Centro Latinoamericano de Demografía. Santiago. de Chile, 1972
4. Haupt, A and Kane, T. Guía Rápida de población. Population Reference Bureau, Inc. Washington, D. C. 1980.
5. Spiegelman, M. Introducción a la Demografía. Fondo de Cultura Económica. Mexico, 1972.
6. Ortega, A. Tablas de Mortalidad. Centro Latinoamericano de Demografía. San José, Costa Rica, 1987.
7. Ministerio de Salud Pública. Anuario Estadístico 1996. La Habana, 1997.
8. Ministerio de Salud Pública. Informe Anual 1989. La Habana, 1990.
9. Jenicek, M. Epidemiología: Principios, técnicas, aplicaciones.Salvat Editores, S.A. Barcelona, 1988.
10. Keyfitz, N. Applied Mathematical Demography. John wiley& Son. 1977.
11. World Health. Organization. Manual of Mortality Analysis. Geneva,1977.
12. Jaspers, D y Ortega, A. Mortalidad (Selección de artículos). Facultad de Salud Pública. La Habana, 1986.
13. Keyfitz, N y Flieger, W. Demografía: métodos estadísticos.Ediciones Marymar. Buenos Aires, 1975.