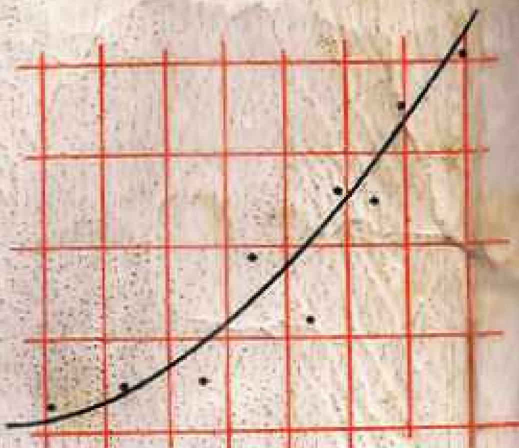
Elementos de Estadística para Preparación y Evaluación de Proyectos



mario murillo oporto

Tercera Edición Revisada

Agosto 1990

Elementos de Estadística para Preparación y Evaluación de Proyectos

mario murillo oporto

La Paz - Bolivia - 1990

Primera Edición, 1974
Segunda Edición Revisada, 1977
Tercera Edición Corregida, 1990

PROLOGO

Es posible prescindir del concurso de la Estadística en la tarea de Preparación y Evaluación de Proyectos ? ...., indudablemente que no.

Es innegable que una de las herramientas principales para la elaboración de todo trabajo de investigación, donde los factores que intervienen sufren cambios constantes que implican incertidumbre y que frente a ellos hay necesidad de hacer predicciones para tomar la acción adecuada en una política de decisiones, es la Estadística.

El presente trabajo fué realizado tomando como base las notas de las clases que me correspondió dictar en la Universidad Mayor de San Andrés, Universidad Católica Boliviana y en el primer Curso de Preparación y Evaluación de Proyectos (Programa BID-ISAP-CONEPLAN-OEA - Ciudad de La Paz, 1973). Se han incluido los temas que se consideraron más importantes, tratando de dejar claras las ideas centrales y sus aplicaciones, sin entrar a muchas demostraciones matemáticas, de acuerdo al objetivo perseguido. Se recomienda a los estudiantes ampliar la lectura de cada Capítulo, al menos con uno de los textos mencionados en la bibliografía.

Deseo resaltar la persona de Don Hugo Javier Ochoa G., Director del Curso de Preparación y Evaluación de Proyectos, citado anteriormente, de quién surgió la idea de realizar el presente documento. Para él, mis mejores palabras de elogio y agradecimiento por su colaboración desinteresada, tolerancia y estímulo constante de que fuí objeto durante el tiempo que duró mi labor.

La Paz - Bolivia
1974

PROLOGO A LA TERCERA EDICION

Al presentar esta nueva edición, deseo expresar especialmente mi más profundo sentir de agradecimiento a todos cuantos me honraron con la lectura de las dos impresiones anteriores.

La aceptación que tuvo el texto y el actual interés manifestado por contar con el, sobre todo por estudiantes de carrera y muchos egresados, han constituido motivos suficientes para haber emprendido la tarea de hacer las correcciones que eran necesarias y de añadir las soluciones de ejercicios seleccionados.

Abrigo, nuevamente, la esperanza de que esta obra en el nivel dado, pueda servir de un elemento más de apoyo para quienes están incursionando en el conocimiento de la Estadística y sus aplicaciones; entre ellos, a estudiantes de la Universidad Boliviana y de otras Instituciones donde se enseña esta ciencia.

Mario Murillo Oporto

La Paz - Bolivia
Agosto, 1990

C O N T E N I D O

CAPITULO 3

MEDIDAS DE TENDENCIA CENTRAL

CAPITULO 4

MEDIDAS DE DISPERSION

CAPITULO 5

TEORIA DE PROBABILIDADES

CAPITULO 6

VARIABLES ALEATORIAS Y FUNCIONES DE DISTRIBUCION

CAPITULO 7

MODELOS DE DISTRIBUCIONES DE PROBABILIDADES

CAPITULO 8

MUESTREO Y DISTRIBUCIONES EN EL MUESTREO

CAPITULO 12

NUMEROS INDICES

APENDICES

CAPITULO I

PROBLEMAS DE DECISION ESTADISTICA

1.1 Introducción

Intuitivamente, desde muy temprana edad, casi todos hemos utilizado medidas estadísticas sobre la base de una colección de datos obtenidos de alguna manera. Así por ejemplo, desde que estuvimos en la escuela, ya se sabía que para aprobar una materia o un curso, era necesario un *promedio* mínimo; promedio que aprendimos a calcular dividiendo la suma total de las notas parciales entre el número de ellas. Una verdulera sin ninguna instrucción escolar es capaz de predecir cuáles serán sus ventas en un día dado, sobre la base de sus experiencias pasadas en días similares. Un campesino agricultor está en condiciones de decir que en cada nidada de huevos de gallina, generalmente el cuarenta por ciento resultan gallos.

Así como los tres casos mencionados anteriormente, podemos citar miles de ejemplos en los que está implícita la idea intuitiva de observaciones, medidas y predicciones estadísticas. Como puede apreciarse, no se tratan de hechos recientes, toda una vida seguramente se ha estado frente a razonamientos iguales, los mismos que forman parte de lo que nosotros actualmente llamamos *Estadística*. No se trata de algo nuevo ni reciente, como la teoría de la relatividad, la teoría de conjuntos o la gramática estructural. Quién nos dice si no se trata de una de las disciplinas más antiguas, puesto que se tienen indicios de que en sociedades como la Egipcia, mucho antes de conocer los jeroglíficos, tenían ya alguna forma de registro de datos estadísticos; lo propio ocurre en nuestro continente con la civilización "Quechua", que antes de conocer la escritura ya utilizaban los "Quipus" para coleccionar informaciones relacionadas con la producción agrícola y otros.

La importancia que tiene, sobre todo, la Estadística moderna en el desarrollo de la vida actual, es simplemente singular. Es posible que ninguna actividad en cualquier campo de la investigación científica marche adelante sin el concurso de las estadísticas, particularmente en el estudio de mercados, instrumento de mucha utilidad para la Evaluación de Proyectos. La gran importancia de la Estadística y sus múltiples aplicaciones podrán apreciarse a medida que vayamos estudiando cada uno de los capítulos del presente trabajo.

1.2

Significación y Naturaleza de la Estadística

Como se ha dicho, en la vida moderna la influencia de la estadística es tal, que son muy pocas las personas que no han escuchado la palabra "estadística", o que están en medio de fenómeno estadísticos, o que utilizan la estadística de alguna manera. En primera instancia, casi todos creen que la estadística consiste únicamente en la recolección de un conjunto de datos, la presentación de éstos en tablas o gráficos y el cálculo de totales, promedios, porcentiles, proporciones, etc. Así por ejemplo, las estadísticas económicas para algunos consiste tan sólo en tener datos numéricos sobre empleo, ingresos, precios de ventas, importaciones, ventas, etc. Es como si la estadística tratara únicamente de presentar una información acerca de cualquier actividad expresada en números. Aún hay quienes creen ciegamente en las frases clásicas "las cifras hablan por sí mismas" o "los números no mienten". Esto es relativo, ya que es posible presentar datos malamente llamados estadísticos, que aparenten informar algo que no es real. El significado de *estadística* es mucho más amplio y va más allá de lo que se indicó anteriormente. La estadística se refiere también a un cuerpo de técnicas o metodología para la recopilación, presentación y análisis de datos cuantitativos y cualitativos, y al uso de tales datos para tomar decisiones. Pero esto no es todo, ya que se refiere también a la predicción frente a la incertidumbre de los fenómenos de la naturaleza.

En otras palabras, podemos decir que la *estadística* es un método que sirve para tomar decisiones cuando hay incertidumbre, sobre la base de datos numéricos, y calcular su riesgo. La importancia de la estadística como un procedimiento para tomar decisiones, es obvio si consideramos que estamos viviendo en un mundo donde los sucesos futuros están llenos de variación e inceritidumbre.*

---

* Quedará más clara la idea de lo que significa realmente la *estadística*, conforme se vayan viendo los temas que siguen.

Cuando hablamos de tomar decisiones, nos referimos a las *decisiones estadísticas*, que estén basadas en datos numéricos y sus probabilidades de ocurrencia.

Gracias a la estadística será posible contestar preguntas como estas: Están aumentando los costos de mano de obra por unidad producida? Qué incidencia tiene el incremento de producción de las materias primas en el Producto Interno Bruto (PIB)? En cuánto debe incrementarse la producción de cemento para cubrir la demanda de viviendas ocasionada por una explosión demográfica? Cuál será la proporción de artículos defectuosos producidos por un nuevo sistema de producción?. Una buena respuesta para éstas y muchas otras preguntas, lógicamente dependerá de cuán bien haya sido elegido el método y formulación del procedimiento estadístico a seguir.

1.3 Etapas de la Investigación Estadística

Cualquiera que sea el campo de aplicación estadística, el procedimiento a seguir básicamente es el mismo. Las etapas principales son cinco, cuyo orden lógico es el siguiente: 1) formulación o definición del problema; 2) diseño del experimento; 3) recolección de los datos; 4) clasificación, tabulación y descripción de los resultados; y 5) generalización o inferencia final.

En estadística, la totalidad de datos pertinentes que pueden ser recolectados y observados para un problema dado, se llama *población* o *universo*. En otras palabras, se trata de un *conjunto* de elementos de los cuales deseamos saber alguna característica particular. Así por ejemplo, si deseamos conocer el porcentaje de desocupados en la ciudad de La Paz, los componentes de la población serán todas las personas en edad de trabajo que viven en La Paz, de los cuales se puede obtener la información respectiva para averiguar el porcentaje de los que no trabajan.

El segundo paso consiste en determinar cómo se obtendrá la información para el estudio de alguna característica de la población. Para ello, se deberán registrar los datos correspondientes a las *unidades* o *elementos* de la población o solamente a una parte de estos. En el primer caso el procedimiento se denomina *enumeración completa* o *censo*; y el segundo, *muestreo*. El método a seguir dependerá de una serie de factores. En la práctica, por razones de costo, tiempo requerido e imposibilidad de realizar un censo, se usa el muestreo como método adecuado de estudio. Así por ejemplo, se puede tomar sólo una parte de las personas en edad de trabajo residentes en La Paz para averiguar el porcentaje de desocupados. Sobre la base de una muestra, tratamos de tomar una decisión acerca de la población de la cual se toma la muestra. Entonces, es obvio que la muestra debe representar adecuadamente a la población. La obtención de una muestra representativa es uno de los componentes más importantes de la teoría estadística, la cual implica preguntas como estas: Cuál debe ser el tamaño de la muestra?. Qué tipo de datos deben ser coleccionados?. Cuáles son los datos a coleccionarse?. Estas preguntas pertenecen al *diseño de experimentos* en estadística. En el capítulo correspondiente veremos este método.

Cuando se siga el procedimiento de tomar una muestra, ésta debe ser aleatoria. Una *muestra aleatoria* puede definirse como un experimento en el cual cada elemento de la población es elegido al azar con la misma chance. Enfatizamos ésto, debido al hecho de que sólo las muestras aleatorias pueden tratarse con probabilidades, y la consideración de probabilidades es la base para tomar inferencias válidas.

La tercera etapa es la recolección de los datos de acuerdo al experimento diseñado. El método para la obtención de la información dependerá de la magnitud y características peculiares del problema que se quiera estudiar. Esta parte corresponde a las técnicas estadísticas de recopilación de datos propiamente dicha. Sin entrar en detalles, podemos mencionar en líneas generales los más usados: 1) Obtener datos de publicaciones o de registros, y 2) Mediante encuestas (las que pueden ser por medio de entrevistas pérsonales, por teléfono o por correspondencia). Dependerá también de los recursos económicos y tiempo disponibles para la ejecución de encuestas y el procesamiento de los datos obtenidos.

La cuarta etapa consiste en organizar adecuadamente la información. La organización de los datos a su vez implica los siguientes pasos: 1) Los datos deben ser *corregidos* cuidadosamente, de tal manera que las omisiones, inconsistencias, respuestas irrelevantes y cálculos equivocados en la encuesta, no den lugar a resultados erróneos. 2) Debe hacerse una *clasificación* adecuada, y presentar la información en tablas de una manera sistemática, pues de ello dependerá una buena interpretación de los datos. 3) Consiste en la tabulación de los datos clasificados, la que se puede hacer en forma manual o mecánica mediante computadoras electrónicas. De aquí se obtendrán medidas que describan, por ejemplo: porcentajes, promedios, desviaciones estándar, etc.

Una medida derivada de una muestra se denomina *estadístico* *. Si una medida es calculada de las observaciones de la población entera, se dice que es un *parámetro*. En el ejemplo cuando tratamos de estudiar el desempleo, el resultado del porcentaje de personas sin trabajo puede obtenerse de una muestra: El proceso de esta cuarta etapa, alcanza a definir lo que se llama *estadística descriptiva*.

Cuando la muestra incluye a todos los elementos de la población, esto es, cuando una investigación se efectúa por enumeración completa, el estudio de la última etapa, se reduce a la estadística descriptiva. Entonces cuando

---

* Otros autores utilizan la denominación, *estadística*.

se tiene la información de todo el conjunto, las decisiones no tendrán ninguna dificultad de ser formuladas.

Mientras que si la información proviene sólo de una parte de la población, el estudio no termina únicamente con las medidas descriptivas. Con los resultados de la muestra se trata de conocer el verdadero estado de la población. En otras palabras, la última etapa consiste en intentar una respuesta sobre la base del estadístico muestral, respecto de la pregunta originalmente formulada, la cual está relacionada con el parámetro de la población. Así por ejemplo si suponemos que de una muestra se obtuvo que el 20% son desocupados en la ciudad de La Paz. Entonces este resultado, es un *estimador* del porcentaje total de desocupados. En el fondo, con esto se está haciendo una generalización (o inferencia) sobre los otros trabajadores no observados.

El estudio de la metodología para obtener conclusiones de generalización, se llama *estadística inductiva*, que no sólo es el modelo estadístico más valioso sino también la parte másfascinante de la estadística. El valor de la estadística inductiva descansa en el hecho de que es característicamente un método científico de investigación. También, en este caso, los conceptos se aclararán mucho más en los capítulos corr spondientes a la vez que se vaya presentando la metodología para resolver *problemas de decisión*.

1.4

Población, Medidas Elementales y Observaciones

Conviene ampliar un poco más algunos conceptos que si bien son intuitivos, en muchos casos suelen presentar ambigüedad. En estadística, el término *población* comunmente se usa para designar una colección o reunión de elementos individuales, como personas, animales, cosas, que serán observados según se tenga un problema especial de investigación. Los objetos individuales en una población se llaman *unidades elementales*. Las unidades elementales poseen ciertas *características*, algunas veces se las consideran como rasgos o propiedades, las cuales pueden ser cualitativas o cuantitativas. Por ejemplo, un problema de decisión puede consistir en observar la efectividad de una máquina laminadora de madera; aquí, la población incluye todas las unidades elementales de láminas de madera hechas por dicha máquina, y la característica de la unidad elemental es su cualidad observada, es decir si está bien o mal laminada. Otro problema puede tratarse del ingreso familiar en una ciudad dada; aquí la población incluye a todas las viviendas familiares de dicha ciudad, y la característica a observarse es la cantidad del ingreso familiar de las unidades elementales que son las viviendas familiares.

El resultado de observar una unidad elemental se llama una *observación*. Notemos que la definición de una población y la característica de la unidad elemental a observarse dependerá de la naturaleza del problema de que se trate. Aclaremos estos conceptos con los siguientes ejemplos, dada la importancia que revisten:

Ejemplo 1.1 Zapatos para niños.

a) Problema de decisión: Decidir la cantidad apropiada de producción de zapatos para niños de acuerdo al tamaño

b) Población: Todos los niños de Bolivia

c) Característica: Tamaño del pie de cada niño

Ejemplo 1.2 Pronóstico de ventas.

a) Problema de decisión: Estimar la demanda futura de los consumidores.

b) Población: Todos los almacenes de venta

c) Característica: Volumen de ventas

Ejemplo 1.3 Inversión Personal

a) Problema de decisión: Determinar la inversión de capital

b) Población: Listado de capitales invertidos en La Paz

c) Característica: Tasas de dividendos.

Ejemplo 1.4 Pronóstico de Turismo

a) Problema de decisión: Estimar la demanda de hospedaje de los turistas.

b) Población: Todas las hospederías (hoteles, alojamientos, residenciales, etc.)

c) Característica: Transito de turistas.

Las poblaciones pueden clasificarse en infinitas y finitas. Una *población finita* es aquella población que contiene un número finito de unidades elementales; mientras que *población infinita* es una población que contiene un número infinito de unidades elementales.

1.5

Medidas de las Unidades Elementales

Las características de las unidades elementales pueden ser expresadas en términos de números de modo que ellos proporcionen mediante métodos estadísticos indicadores de análisis. Dijimos que las características de las unidades elementales pueden ser cualitativas o cuantitativas. Las características cuantitativas pueden transformarse en datos numéricos directamente midiendo cada unidad elemental, tal como metros, onzas, $b., grados centigrados, etc. De esta manera los resultados de las medidas de las unidades elementales forman la totalidad de observaciones, las cuales son expresadas numéricamente en términos de la unidad de medida usada. Los valores que las observaciones cuantitativas pueden tomar, suelen llamarse variantes, y el conjunto de ellas, *variable*. Esto se debe al hecho de que generalmente de unidad a unidad elemental tomarán distintos valores; y es justamente éste el caso interesante, puesto que se trata también de medir dichas variaciones. Si una variable toma un mismo valor, se dice que es una *constante* por ejemplo, si la edad de los niños al iniciar un curso de pre-Kinder, es la misma.

Cuando los valores de una variable son arreglados en algún orden ascendente o descendente, se forma una *serie estadística*. Una serie puede ser contínua o discreta. Una *serie contínua* es una variable que puede tomar cualquier valor numérico comprendido en un determinado recorrido. Es el caso en que los valores sucesivos de la serie pueden diferir en cantidades infinitesimales. En otras palabras una serie continua es tal que la unidades elementales pueden tomar valores divididos entre dos valores contiguos. Ejemplos de series continuas, son las medidas de peso, estatura, tiempo, longitud o temperaturas.

Pero no toda unidad de medida es divisible, esto es, la unidad de medida puede definirse solamente en términos de enteros. Así por ejemplo, el número de hijos; podrán ser 1, 2, 5 o 10, pero nunca 1/5, pues no es posible admitir una fracción de niño. Las series en las cuales no es posible que hayan medidas continuas se denominan *series discretas*.

No obstante que teóricamente hay una distinción clara y precisa entre variaciones continuas y discretas, en la práctica para trabajos estadísticos se suelen realizar aproximaciones relativas, asumiendo que una serie discreta, se aproxima a una serie continua y viceversa, según sea el tratamiento estadístico a emplearse. Así, una serie continua, puede transformarse en discreta a fin de facilitar las medidas de la población; por ejemplo cuando se toman las estaturas redondeando únicamente en centímetros (pues no tiene sentido decir que una persona mide 175,405671..... cms.).

Otras veces series discretas se aproximan a contínuas con el fin de utilizar, por ejemplo, la curva normal de probabilidades. También, esta parte, se aclarará más adelante sobre todo desde el punto de vista práctico.

En el caso de tener características cualitativas no es posible medirlas directamente. Para salvar este problema se clasifican las unidades elementales en dos o más grupos con cierta cualidad o propiedad particular, las cuales se llaman *atributos*, y luego se asigna un valor a cada grupo. Así por ejemplo, si se trata de un conjunto de artículos producidos por una máquina, éstos pueden clasificarse en malos, regulares y buenos, y entonces podemos asignar el número 0 cada vez que aparece un artículo malo, el 1 si es regular y el 2 si el artículo es bueno. Este tipo de cuantificación de características cualitativas, es muy útil para poder hacer inferencias estadísticas.

Muchas de las unidades elementales tienen características tanto cuantitativas como cualitativas, por ejemplo si la unidad elemental es una persona, su estatura será una característica cuantitativa, mientras que su deseo por consumir leche o no hacerlo, es una característica cualitativa, etc. Por esta razón es muy importante definir con precisión la característica que se desea estudiar.

1.6

Distribución de Observaciones y Distribución de Frecuencias

Al conjunto de observaciones numéricas de una serie, llamaremos *datos estadísticos*. Además sabemos que esas observaciones varían de valor de elemento a elemento. Nuestro interés no es conocer valores individuales separadamente, porque la estadística se refiere esencialmente a la información de toda la población. De esta manera, nosotros usualmente intentamos resumir las observaciones individuales en una u otra forma cuando estudiamos una población. El resumen más comprensivo de una colección completa de observaciones se llama *distribución de frecuencias*. En el capítulo siguiente discutiremos detalladamente la construcción y uso de las distribuciones de frecuencias. Aquí usaremos este concepto solamente en relación al problema de aplicación a la población.

Así por ejemplo, cuando se miden las estaturas de 350 estudiantes y los resultados son ordenados en una serie en orden ascendente, primero que todo advertimos el *recorrido* de la serie, que es la diferencia entre el mayor y menor valor de la serie. Entonces podemos proceder a dividir el recorrido en un número de partes o *clases*, sobre la cual se distribuyen las informaciones. Cada clase evidentemente tienen su propia frecuencia, ya que el número de medidas que caen en cada clase posiblemente diferirán de una clase a otra. Una distribución de frecuencias puede presentarse en forma tabular, en una y

tabla de distribución de frecuencias. Sin embargo también es importante conocer la distribución de *frecuencias relativas*, que no son otra cosa que la proporción del total de unidades de una clase en relación al total de unidades observadas. En la tabla 1.1 se representan estos conceptos. En estudios

TABLA 1.1 Distribución de frecuencias de 350 estudiantes clasificados según sus estaturas.

Estatura (en metros)	Frecuencia Absoluta	Frecuencia Relativa
clase (1.40 - 1.44)	3 3/350	0,0086
1.45 - 1.49	11	0,0314
1.50 - 1.54	25	0,0714
1.55 - 1.59	42	0,1200
1.60 - 1.64	67	0,1914
1.65 - 1.69	84	0,2400
1.70 - 1.74	73	0,2086
1.75 - 1.79	35	0,1000
1.80 - 1.84	8	0,0229
1.85 - 1.89	2	0,0057
	350	

posteriores podremos apreciar la analogía que tienen las frecuencias relativas con el concepto de probabilidad. En síntesis, la distribución de frecuencias, a menudo se utiliza para resumir datos para posteriores análisis.

1.7 Especificación de parámetros de Decisión

Como hemos explicado al principio, toda vez que un problema puede ser resuelto con la ayuda de datos numéricos, la primera cosa es definir el problema estadísticamente; esto es, definir la población estadística y especificar el parámetro de decisión. Hasta ahora hemos discutido varios aspectos relacionados con la definición de una población estadística, ahora estudiaremos el problema de seleccionar parámetros de decisión.

Frecuentemente un parámetro se llama también *valor verdadero*. Se trata de un valor simple obtenido por métodos estadísticos que describe en forma resumida la característica en estudios de una población. Una población puede tener muchas características, por lo tanto, habrán muchos parámetros. A menu do se desea conocer, un valor mínimo, un valor máximo, un recorrido, el valor total de todos los valores individuales (o el agregado), un valor promedio de todos los valores individuales, etc. Un *parámetro de decisión* es una característica o propiedad de una población que se requiere como base para tomar una acción de decisión. Es claro que no todos los parámetros de una población son parámetros de decisión dada la situación de un problema. De la misma manera que es importante definir la población, la selección del parámetro de decisión apropiado dependerá de la naturaleza del problema en manos. Otras veces son necesarios varios parámetros de decisión, para resolver un problema de decisión estadística.

Para terminar este capítulo, definiremos a continuación algunos parámetros de decisión más usuales. La discusión detallada de todos los parámetros y sus propiedades se verán en el capítulo 3.

1.8

El Agregado, La Propoción, La Media Aritmética, La Varianza y la Desviación Estándar

Será importante para el estudio de los métodos estadísticos, fijar bien los conceptos y los símbolos que representarán a las diferentes medidas, con el fin de facilitar la relación entre ellas mediante fórmulas. Así, generalmente para los parámetros de decisión poblacional, suelen usarse letras griegas, pero por razones tipográficas, no siempres será posible recurrir al alfabeto griego, en un caso tal, recurriremos a las letras latinas. Un símbolo muy usado para la operación fundamental de *suma* es la letra griega mayúscula $\sum$ (sigma), la cual indica que deben añadirse un conjunto de observaciones elementales, previamente definidos. (Se recomienda estudiar el texto sobre Sumatorias y Productorias citado en la Bibliografía).

Ahora definamos los parámetros de decisión indicados en el título de esta sección. El *agregado* de una población es simplemente el *total* o la *suma* de los valores numéricos de las observaciones individuales en la población. Sea T la suma de los N valores de la variable, entonces

$$T = x_1 + x_2 + \ldots + x_N = \sum_{i=1}^{N} x_i \qquad (1.1)$$

donde cada x_i (i=1, 2,.., N) representa el valor individual de la iésima observación. Tomemos un ejemplo sencillo: Se trata de saber el ingreso total por mes de una familia en la que trabaja los padres y los tres hijos mayores, sabiendo que:

el ingreso mensual del padre es $= x_1 = 1.200$ \$b.
el ingreso mensual de la madre es $= x_2 = 850$ \$b.
el ingreso mensual del 1er. hijo es $= x_3 = 1.250$ \$b.
el ingreso mensual del 2do. hijo es $= x_4 = 1.300$ \$b.
el ingreso mensual del 3er. hijo es $= x_5 = 750$ \$b.

Entonces el ingreso total de la familia es :

$$T = \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$= 1.200 + 850 + 1.250 + 1.300 + 750 = 5.350$$

Antes de continuar, vale la pena hacer alguna referencia repecto del número de elementos de la población (N) , y el número de elementos de la muestra (n). Usaremos N sólo cuando se está especificando que el número de las unidades elementales observadas son todas las de la población. En cambio usaremos generalmente n, no sólamente para indicar que las observaciones provienen de una muestra, sino que eventualmente puede ser igual a N (o sea que puede coincidir el tamaño de la muestra con el tamaño de la población). Además, como haremos máshincapié en la estadística inferencial, trabajaremos particularmente con muestras. Entonces, el agregado de una muestra será

$$T = x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i \qquad (1.2)$$

es decir, que se suman los valores de las n obsevaciones disponibles. Si en el contexto se está trabajando con n observaciones, podemos simplificar la notación de la sumatoria poniendo únicamente $\sum x$.

Cuándo se usará el agregado como parámetro de decisión? A menudo, por ejemplo en el campo económico, se desea saber algunos totales, como ser: El Producto Nacional Bruto, el total de los ingresos personales de cierto sector de trabajo, el valor total de las importaciones de artículos suntuarios efectuados en un año determinado, etc. Casi en todos los campos de investigación una primera medida que se desea conocer, es precisamente el agregado o total de los valores que presentan los elementos simples de la población.

Otro parámetro de decisión es la *proporción* de una población (la designaremos con la letra mayúscula P), que es la razón del número de unidades elementales A, que poseen un cierto atributo, entre el número total de unidades elementales de la población N. Simbólicamente,

$$P = \frac{A}{N} \qquad (1.3)$$

Por ejemplo, si examinamos la conducta de los consumidores de carne, y hallamos que 90 personas de 150 personas encuestadas prefieren carne de res, entonces $A = 90$ y $N = 150$, y $P = \frac{A}{N} = \frac{90}{150} = 0,60$, ó 60%

Si bien la proporción es uno de los conceptos más simples, se trata de uno de los parámetros más importantes de decisión. Así por ejemplo, a menudo se desea saber el porcentaje de reprobados de un curso universitario; la proporción de desocupados en una ciudad; la proporción de artículos defectuosos con un cierto proceso de producción; la proporción de personas que están de acuerdo con un sistema de precios; el porcentaje de personas afectas al uso de cierta marca de cigarrillos, etc.

Otra propiedad de una población usada frecuentemente como parámetro de decisión es la medida de un "promedio". En estadística, como veremos más adelante, hay muchos tipos de promedios. El más simple como también el más importante, es la *media aritmética*, que es la suma de los valores observados de la población dividida entre el número total de observaciones. A la media aritmética poblacional, la designaremos con la letra griega μ.

Entonces la media aritmética poblacional será

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum x}{N} \tag{1.4}$$

La media aritmética a menudo es el parámetro clave para muchos problemas estadísticos. Prácticamente se usará este concepto en todo el presente trabajo, por su gran utilidad en la estadística inductiva. Este parámetro de decisión se usa en muchísimos casos de problemas estadísticos; por ejemplo, cuando se desea saber el ingreso promedio de las personas no asalariadas, el tiempo promedio de duración de un cierto artículo, la producción promedio de artículos manufacturados, el consumo promedio de carne de pescados, el promedio de personas que enferman de cáncer a los pulmones de una población con hábito de fumar, etc. Seguramente no hay una persona adulta, con o sin preparación estadística, que no hable del concepto de promedio en alguna actividad de su vida cotidiana.

En el ejemplo del ingreso familiar que se dió anteriormente, podemos medir el ingreso promedio de las cinco personas que trabajan, o sea

$$\mu = \frac{\sum x_i}{N} = \frac{5.350}{5} = 1.070 \text{ \$}$$

de manera que los cinco miembros de la familia que trabajan, ganan mensualmente en promedio 1.070 $

Hay otros parámetros de decisión también de gran importancia, como aquellos que miden el grado de variación de los valores de las unidades elementales en investigación. Dada una serie estadística, el valor de una variable generalmente difiere de otra dentro de la misma serie. Es obvio que si el valor observado de todas las unidades elementales de la población es único, no existirá ninguna variación. Para medir una variación de valores, debe hacerse con repecto a algún punto de referencia. La medida que tiene mayor importancia, es aquella que nos mide las variaciones (o desvíos, o diferencias) de cada unidad elemental respecto de la media aritmética de todas las observaciones. Además, será el promedio de las variaciones quien nos proporcione un mejor indicador de las variaciones poblacionales. La medida más usada para este fin, es la *varianza* de la población, que es simplemente el promedio de los cuadrados de las desviaciones de los valores individuales respecto de la media. A la varianza poblacional, la designaremos por σ^2 entonces

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N} = \frac{\sum (x_i - \mu)^2}{N} \qquad (1.5)$$

Como esta medida cambia el valor original de las unidades elementales (al ser elevados al cuadrado), una medida que se deduce directamente de la varianza, es la *desviación estandar*, que es la raíz cuadrada de la varianza, o sea :

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

La desviación estándar σ, es una de las medidas más importantes que sirve como indicador de la variabilidad de una población; además, se verá que es un parámetro de múltiples aplicaciones.

Si consideramos la información de nuestro ejemplo, la varianza y desviación estándar del ingreso familiar serán:

$$\sigma^2 = \frac{\sum (x_i - 1070)^2}{N} = \frac{(x_1-1070)^2 + (x_2-1070)^2 + (x_3-1070)^2 + (x_4-1070)^2 + (x_5-1070)^2}{5}$$

$$= \frac{(130)^2 + (-220)^2 + (180)^2 + (230)^2 + (-320)^2}{5} = 50.600$$

De donde

$$\sigma = \sqrt{50.600} = 224,94$$

Entonces la variación de los ingresos de las cinco personas, es de 224,94 \$.

CAPITULO 2

DISTRIBUCION DE FRECUENCIAS

Tal como habíamos adelantado en la sección 1.6 del capítulo anterior, las decisiones estadísticas están basadas sobre datos numéricos; los mismos que pueden presentarse en forma resumida en una *tabla de distribución de frecuencias*, cuya importancia es innegable, en virtud de que nos proporciona una primera aproximación respecto de la forma en que están distribuidas las series de datos estadísticos. Por otro lado, la representación gráfica de una distribución de frecuencias también es útil, puesto que ello nos dará una primera idea del tipo de distribución de que se trata.

Recordemos sin embargo, que a los datos se los necesita sólo cuando los estados de la naturaleza son desconocidos. Por otro lado, los tipos de datos que se requieren dependen sobre todo del problema de decisión a resolver. Así, muchas veces inclusive es necesario recoger varias características simultáneamente de cada elemento de la población. Cuando obtengamos un sólo valor de las unidades elementales, llamaremos datos *univariantes*, si obtenemos dos valores, diremos datos *bivariantes;* y en general, si obtenemos dos o más valores, diremos datos *multivariantes*. En el presente capítulo, ilustraremos las distribuciones de frecuencias y sus características, utilizando una sola variable.

2.1

Construcción de Distribuciones de Frecuencias

Se ha visto que los datos recogidos de las unidades elementales de la población o de la muestra, individualmente significan poco para el investigador. En cambio, prestan gran utilidad, cuando éstos se hallan ordenados o clasificados de alguna manera sistemática. Las distribuciones de frecuencias dependerán del tipo de series que son posibles de obtener para resolver un problema estadístico. Así, las series pueden ser discretas o continuas. En el caso de tener valores discretos, de acuerdo a su número, pueden ser clasificados en *clases* o grupos.

Empezaremos por ver el caso más simple de las distribuciones de frecuencias. Partiremos de la siguiente información correspondiente a las calificaciones (redondeadas) obtenidas en una prueba por 50 alumnos, el la escala del 10 al 70.

TABLA 2.1 Calificaciones obtenidas por 50 alumnos en una prueba de evaluación, en una escala del 10 al 70.

40	40	60	30	50	50	60	50	60	30
30	10	50	20	30	40	40	40	50	40
60	20	40	70	30	30	30	30	40	10
20	50	30	40	10	60	20	60	40	60
50	70	20	40	40	20	10	30	30	50

De acuerdo a los conceptos y definiciones anteriores, cada uno de los elementos de la Tabla 2.1 corresponden al valor de cada una de las 50 unidades elementales de la población, a los que simbólicamente designaremos por x_i (en este caso, i varía de 1 a 50, o sea que hay $x_1, \ldots, x_{50}$ valores). Los valores de la Tabla anterior, pueden resumirse en una Tabla de Distribución de Frecuencias, agrupando a todos aquellos valores iguales, en cada una de las categorías de calificaciones, como se ve en la Tabla 2.2

TABLA 2.2 Distribución de Frecuencias Absolutas correspondiente a las calificaciones abtenidas por 50 alumnos en una prueba de evaluación (Escala, 10 al 70).

Calificación X	Registro o cómputo de las repeticiones de los valores distintos que toma la variable	Frecuencias Absolutas $n_i = f$
$x_1 = 10$	////	$n_1 = 4$
$x_2 = 20$	///// /	$n_2 = 6$
$x_3 = 30$	///// ///// /	$n_3 = 11$
$x_4 = 40$	///// ///// //	$n_4 = 12$
$x_5 = 50$	///// ///	$n_5 = 8$
$x_6 = 60$	///// //	$n_6 = 7$
$x_7 = 70$	//	$n_7 = 2$

En la Tabla 2.1 vemos que hay varios x_i iguales. Sin embargo, en la Tabla 2.2, en la primera columna aparecen tan sólo siete valores para los x_i; esto se debe a que se han clasificado a las x_i, sólo en los distintos valores que ella toma, ya que en la columna de las frecuencias absolutas aparecen el número de veces que se repite un valor x_i particular. Algunos autores, usan otra letra para designar a los *distintos* valores que toma la variable X, por ejemplo, la letra Y; de manera que en este caso tendríamos $y_1, y_2, \ldots, y_7$

para la primera columna de la Tabla 2.2. Nosotros usaremos la misma letra para designar la variable que se esté considerando, sea que tengamos el conjunto de datos originales sin agupar, o cuando los tengamos agrupados.

Para la notación de las *frecuencias absolutas*, usaremos n_i, que nos indicará: el número de veces que aparece repetido un valor x_i; de esta manera, en nuestro ejemplo, $n_4 = 12$ significa que se han encontrado 12 alumnos con una calificación de 40. En la columna central de la Tabla 2.2 se lleva la cuenta mediante rayas, las que son marcadas a medida que aparece un valor x_i (De la misma manera que se hace un recuento de votos, en una elección cualquiera). Esto es lo que se llama también *computación* de la información original, que de acuerdo al volumen de los datos puede realizarse en forma manual o mecánica.

Por otro lado, podemos representar graficamente una distribución de frecuencias absolutas, como se ve en la Figura 2.1

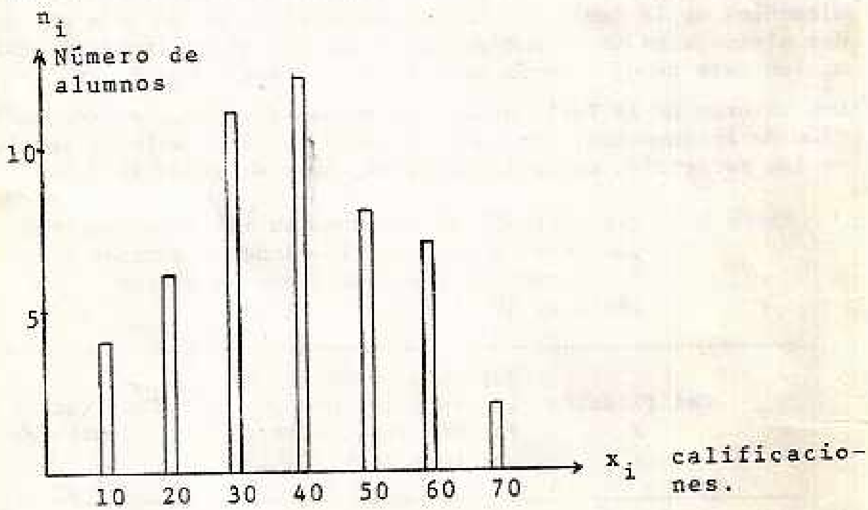


Figura 2.1 Gráfico de barras correspondiente a las calificaciones obtenidas por 50 alumnos en una prueba de evaluación.

Otro gráfico que también es útil, es el que corresponde a las frecuencias acumuladas, (Ver Figura 2.2)

Una tabla de distribución de frecuencias es más completa cuando además incluye otros conceptos en base a las frecuencias absolutas. La *frecuencia relativa* (o simplemente *frecuencia*) de un valor observado distinto, es la razón que existe entre el número de repeticiones de un x_i y el número total de elementos de la población, la que representaremos por $h_i = n_i/n$. Una tabla de distribución de frecuencias, también puede incluir las acumulaciones de las frecuencias, tanto absolutas como relativas. Así, N_i representará la *frecuencia absoluta acumulada*, y H_i representará a las *frecuencias relativas acumuladas*. La Tabla 2.3 nos ilustra estos conceptos para el ejemplo de las calificaciones de los 50 alumnos.

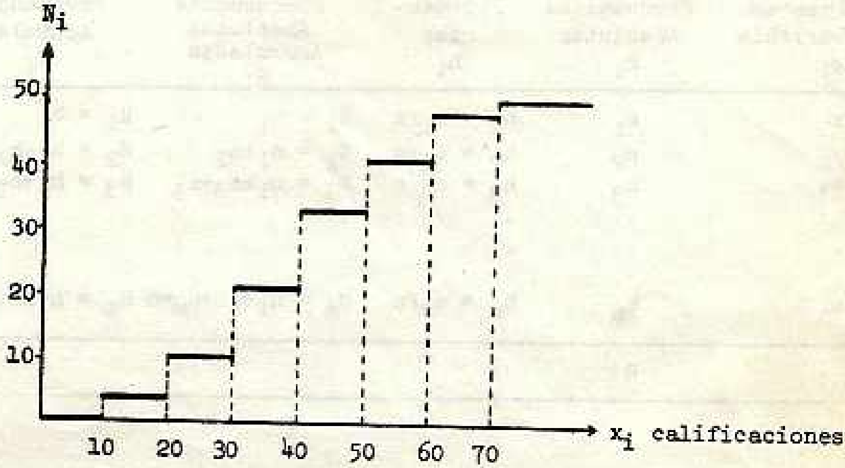


Figura 2.2 Gráfico acumulativo de las calificaciones obtenidas por 50 alumnos en una prueba de evaluación

TABLA 2.3

Calificaciones	Frecuencia Absoluta n_i	Acumulación de las Frecuencias Absolutas N_i	Frecuencia h_i	Acumulación de las Frecuencias H_i
10	4	4	0,08	0,08
20	6	10	0,12	0,20
30	11	21	0,22	0,42
40	12	33	0,24	0,66
50	8	41	0,16	0,82
60	7	48	0,14	0,96
70	2	50	0,04	1
$\sum$	$n=50$		1	

$h_i = \frac{n_i}{n}$

En general, si se tienen n valores observados, una tabla de distribución de frecuencias, simbólicamente estará dada por: (Ver Tabla 2.4)

Donde m índica el número de valores distintos que toma la variable. Cuando se dice, distribución de frecuencias, se entiende también la *forma* en que están distribuidas las repeticiones de cada valor distinto de la población.

$h_i = \frac{n_i}{n}$

TABLA 2.4 Distribución de Frecuencias y Frecuencias Acumuladas.

Valores de la variable x_i	Frecuencias Absolutas n_i	Frecuencias h_i	Frecuencias Absolutas Acumuladas N_i	Frecuencias Acumuladas H_i
x_1	n_i	$h_1 = n_1/n$	$N_1 = n_1$	$H_1 = h_1$
x_2	n_2	$h_2 = n_2/n$	$N_2 = n_1+n_2$	$H_2 = h_1+h_2$
x_3	n_3	$h_3 = n_3/n$	$N_3 = n_1+n_2+n_3$	$H_3 = h_1+h_2+h_3$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_m	n_m	$h_m = n_m/n$	$N_m = n_1+..+n_m=n$	$H_m = h_1+..+h_m=1$
$\sum$	n	1		

2.2

Propiedades y Relaciones de las Frecuencias

Vale la pena hacer algunas consideraciones sobre los elementos que constituyen una Tabla de Distribución de Frecuencias. Los valores x_i de la variable no tienen restricción alguna, puesto que pueden ser números enteros negativos o positivos, o números fraccionarios. Así por ejemplo, si la variable representa al número de hojas, de más o de menos, en un proceso mecánico de empaquetado de 1.000 hojas, las x_i pueden tomar valores como $x=-2$, $x = 0$ ó $x = 4$; lo que significa respectivamente que faltan dos hojas, que está exacto o que hay 4 hojas demás.

Las frecuencias absolutas o repeticiones n_i, y las frecuencias absolutas acumuladas N_i, son números enteros positivos. Las frecuencias h_i y las frecuencias acumuladas H_i, son números fraccionarios positivos no mayores que 1. De esta manera, se tienen las relaciones:

$$0 \leq h_i \leq 1$$

$$\sum_{i=1}^{m} h_i = 1$$

$$\sum_{i=1}^{m} n_i = n$$

Por otro lado, para un N_j y H_j cualquiera

$$N_j = \sum_{i=1}^{j} n_i \quad ; \quad H_j = \sum_{i=1}^{j} h_i$$

También

$$N_m = n \quad ; \quad H_m = 1$$

Finalmente,

$$n_1 = N_1 \leq N_2 \leq \ldots \leq N_m = n$$

$$h_1 = H_1 \leq H_2 \leq \ldots \leq H_m = 1$$

Todas estas relaciones o propiedades, pueden observarse en las Tablas 2.3 y 2.4

2.3

Distribución de Frecuencias para un Número grande de Observaciones y para Variables Contínuas

En la mayor parte en que se tratan datos contínuos, o cuando se tiene un número grande de datos discretos, no es posible asignar una clase única a todos los valores distintos de la variable (sobre todo si son muchos). La única novedad que aparece, en relación a lo estudiado en las secciones anteriores, es que los datos se agrupan en un número razonable de clases que usualmente es de 5 a 15.

ıa *clase* es un intervalo dentro del :ual situamos todos los valores
ndidos entre los *límites de la clase*. La amplitud de una clase se lla
tervalo de clase, la cual es la diferencia entre los límites superior e
rior, de la clase. El número de valores de una serie que caen en una
e, se llama *frecuencia de clase*. Para fines de cálculos, como veremos
adelante, se usará el punto medio de la clase (que algunos autores llaman
rca de clase), que es el promedio de los límites de una clase. Así, en
s cálculos basados en datos agrupados, se asigna el valor del punto medio
e la clase a todas las observaciones contenidas en ésta. En este caso el
ratamiento para los datos contínuos o discretos es la misma.

Consideremos la información de la Tabla 2.5 para tratar todos estos conceptos.

TABLA 2.5 Salarios semanales de 100 obreros, que trabajan en distintas empresas constructoras de viviendas

280	405	300	330	295
290	420	305	320	365
360	370	405	330	375
355	355	400	335	380
375	305	380	355	390
410	420	375	370	300
400	390	385	385	315
390	385	395	380	340
310	380	360	370	365
290	360	355	320	375
360	350	355	330	385
370	350	310	315	330
420	400	325	380	315
285	380	335	345	320
370	370	410	360	340
390	350	390	345	350
415	345	360	340	345
345	430	310	395	365
350	415	325	375	355
410	380	320	395	335

Para determinar el número aproximado de clases, el estudiante puede usar la *Regla de Sturges* como una guía, la que esta dada por

$$k = 1 + 3,3 \log n$$

donde

k = número aproximado de clases
n = número total de observaciones en la muestra
log = logaritmo de base 10

Por ejemplo para la información de la Tabla 2.5 tendremos

$$k = 1 + 3,3 \log 100 = 1 + (3,3) \,.\, 2 \doteq 8$$

Entonces, el número de clases para los salarios semanales de los obreros será 8. Esto no es rígido, puesto que se trata solo de una aproximación, por tanto, si es necesario se pueden tomar 7, 9 o 10 clases.

La selección del número de clases puede establecerse también tomando las siguientes dos consideraciones: 1) procurar una distribución tal que haya una frecuencia mayor que las demás, 2) que el punto medio de la clase tome un valor fácil de trabajar en cómputos. Otro aspecto importante, es propender a que todas las clases en que se divida un conjunto de observaciones, tengan el mismo intervalo. Finalmente, si llamamos c al tamaño del intervalo de una clase, los límites de la primera clase o intervalo se determinan a partir del menor valor observado más la amplitud. En nuestro ejemplo, el tamaño de la clase es

$$c = \frac{V_S - V_I}{k} = \frac{430 - 280}{8} = 18{,}5 \quad (18{,}75?)$$

donde V_S es el valor superior y V_I el valor inferior, que toman las observaciones. Luego, el primer intervalo de clase será de 280 a 298,5; el segundo de 298,5 a 317; y así sucesivamente.

Por otro lado, el punto medio de la primera clase es

$$P_m = \frac{L_I + L_S}{2} = \frac{280 + 298{,}5}{2} = 289{,}25$$

donde L_I y L_S son los límites inferior y superior, respectivamente del intervalo de clase. Como el valor medio de la clase en este caso es un número no muy maniobrable, podríamos cambiar el número de clases, por ejemplo a 10 en lugar de 8. En tal caso el tamaño del intervalo de clase es:

$$c = \frac{430 - 280}{10} = 15$$

De modo que el primer intervalo será de 280 a 295, y el punto medio 287,5.

De manera que la tabla de distribución de frecuencias de datos agrupados para nuestro ejemplo será:

TABLA 2.6 Distribución de Frecuencias de los salarios semanales de 100 obreros.

Intervalos de clase Ingresos x_i	Punto medio (o marca de clase) x_i	Número de Obreros n_i	Acumulac. de número de obreros N_i	Frecuencia del número de obreros h_i	Acumulac. de frecuencias H_i
280 - 295	287,5	4	4	0,04	0,04
295 - 310	302,5	5	9	0,05	0,09
310 - 325	317,5	10	19	0,10	0,19
325 - 340	332,5	9	28	0,09	0,28
340 - 355	347,5	13	41	0,13	0,41
355 - 370	362,5	15	56	0,15	0,56
370 - 385	377,5	18	74	0,18	0,74
385 - 400	392,5	12	86	0,12	0,86
400 - 415	407,5	8	94	0,08	0,94
415 - 430	422,5	6	100	0,06	1
	3550	100		1	

Para calcular algunas medidas de los datos observados, los puntos medios de clase equivalen a los x_i. Este concepto se usará bastante en el proximo capítulo.

De Tabla 2.6 pueden obtenerse muchas informaciones respecto de los valores observados de la población; como por ejemplo, el número y porcentaje de los obreros que ganan entre 340 y 355 \$, el porcentaje de obreros con ingreso inferior a 370 \$, etc.

Hay una observación importante que debe hacerse cuando se confecciona una distribución de frecuencias por intervalos de clase para una serie continua. En nuestro ejemplo el límite superior de un intervalo, coincide con el límite inferior del intervalo inmediato que sigue, lo cual da lugar a la siguiente pregunta: Si un obrero gana, digamos 295 \$, será incluido en el primer intervalo o en el segundo?. Nosotros convencionalmente lo incluiremos en el intervalo inmediato superior. Algunos autores para salvar este tipo de dificultad trivial, clasifican los intervalos, de la siguiente manera:

Intervalo de Clase
280 a 294,99
295 a 309,99
310 a
....

Como puede apreciarse, con la salvedad del caso, podemos evitar el uso de números como 294,99, etc.

2.4 Representación Gráfica de Distribuciones de Frecuencias

No sólo una Tabla de distribución de frecuencias presta una información adecuada para un primer análisis del conjunto de observaciones de una población, también la representación gráfica de la distribución de frecuencias, permite visualizar la forma en que se distribuye un conjunto de observaciones. Los gráficos más usuales son los llamados: *histograma de frecuencias y polígono de frecuencias*. Hay una serie de técnicas para ésto, sobre todo cuando los intervalos de clase son diferentes. Nosotros veremos únicamente el caso en que los intervalos de clase son iguales, puesto que es el más frecuente y además, siempre es posible condicionar cualquier conjunto de observaciones en intervalos iguales. La representación gráfica, se hace básicamente utilizando los ejes coordenados en un plano cartesiano, tomando una escala adecuada.

El histograma de frecuencias para la Tabla 2.6 estará dada por

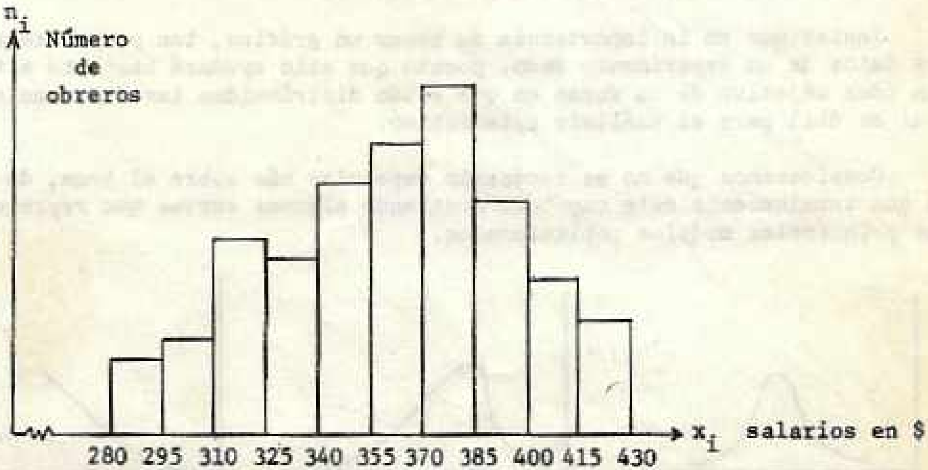


Figura 2.3 Histograma de frecuencias absolutas.

Mientras que el polígono de frecuencias será

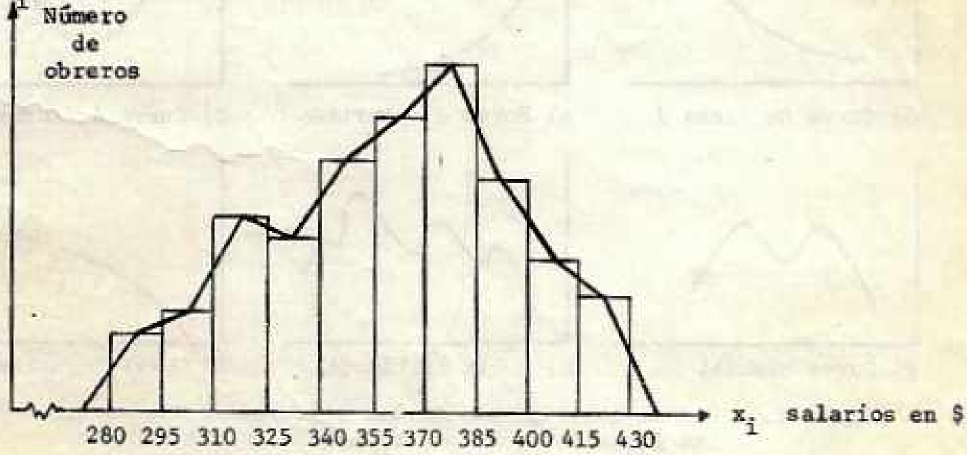


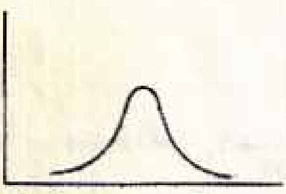
Figura 2.4 Polígono de frecuencias absolutas

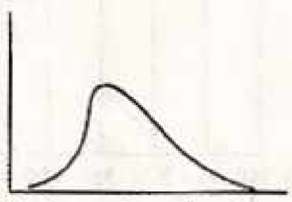
El polígono se traza uniendo los puntos medios de las marcas de clase respectivas. Además, para cerrar el polígono, se toman los puntos medios anterior y posterior al primer y último intervalos.

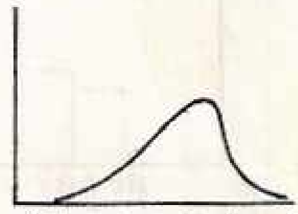
Si se tomaran los valores de las frecuencias relativas, su histograma y su polígono de frecuencias, tendrían la misma forma de las Figuras 2.3 y 2.4 por la proporcionalidad entre todas las frecuencias absolutas y relativas (Tabla 2.6), además que realizando un cambio de escala adecuado en el eje de las ordenadas, los gráficos serán iguales. También es posible graficar fácilmente las frecuencias acumuladas.

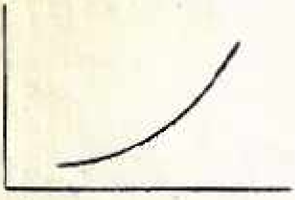
Insistimos en la importancia de hacer un gráfico, tan pronto se tengan los datos de un experimento dado, puesto que ello ayudará bastante a tener una idea objetiva de la forma en que están distribuidas las frecuencias, lo cual es útil para el análisis estadístico.

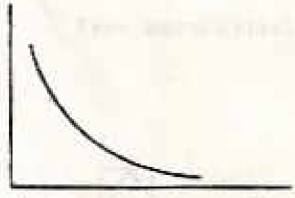
Consideramos que no es necesario especular más sobre el tema, de manera que terminaremos este capítulo mostrando algunas curvas que representan los principales modelos poblacionales.

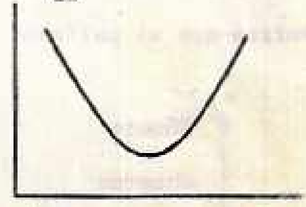

a) Curva normal

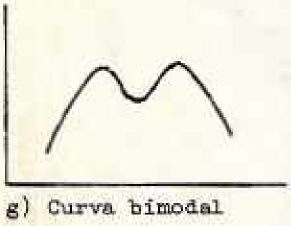

b) Sesgada a la derecha


c) Sesgada a la izquierda


d) Curva de forma J


e) Forma J invertida


f) Curva de forma U


g) Curva bimodal


h) Curva multimodal

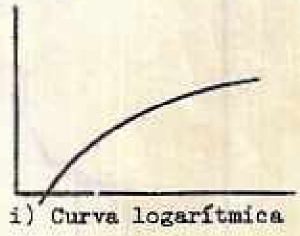

i) Curva logarítmica

Figura 2.5 Curvas que representan las principales formas que toman las distribuciones de frecuencias poblacionales.

2.5

Ejercicios

1. Cuándo es aconsejable usar distribuciones de frecuencias?
2. Cuáles son las ventajas y desventajas del ordenamiento de datos? Es el orden necesario para la construcción de una Tabla de distribución de frecuencias?
3. Qué rol importante juega el punto medio en una distribución numérica?
4. Indique 4 ejemplos de distribuciones unidimensionales, 3 de bidimensionales y 2 de multidimensionales, relacionados con la economía del país.
5. Determine si las siguientes relaciones son posibles y justifique su respuesta.

5.1	$H_4 = 0{,}30$	$n = 10$	$n_3 = 31$
5.2	$h_1 = 4$	$h_3 = 12$	$h_4 = 15$
5.3	$h_2 = 0{,}40$	$n = 50$	$n_1 = 20$
5.4	$h_1+h_2+h_3+h_4 = 1$	$n_6 = 3$	$H_5 = 1$
5.5	$H_4 = 0{,}20$	$H_6 = 0{,}12$	$h_5 = -0{,}08$
5.6	$m = 6$	$H_2 = 0{,}6$	$h_1 = 0{,}20$
	$H_3+H_4 = 1{,}90$	$h_4 = 0{,}2$	

6. Comente las siguientes afirmaciones:

   6.1 No tiene mayor importancia el criterio que se tome para determinar a cuál intervalo pertenece un elemento cuyo valor coincide con el límite de la clase.
   6.2 La suma de las frecuencias absolutas es siempre igual a 1.
   6.3 N_7 indica la frecuencia relativa de valores de la variable mayores que x_7.
   6.4 Mientras mayor es el número de intervalos elegidos para la formación de una distribución de frecuencias menor es la exactitud de los estadísticos que se calculan.
   6.5 $H_i > h_i$ (para todo i)
   6.6 Es preferible trabajar siempre con pocos intervalos.
   6.7 El número de intervalos en que se clasifican un conjunto de datos, depende de la cantidad de éstos.

7. Indicar cuáles de las siguientes variables pertenecen a series continuas y cuáles a discretas: a) tiempo; b) sueldos en una empresa pública; c) magnitudes de grados; d) edad; e) datos de censo; f) distancia recorrida por un móvil; g) puntajes de futbol; h) peso; i) edades mentales; j) ingreso per cápita.

8. Reducir las distribuciones A y B que siguen, a frecuencias de porcentajes y trazar los polígonos de frecuencia sobre los mismos ejes.

Qué consecuencias puede concluir al respecto?

Puntajes	Grupo A	Grupo B
16 - 19	4	0
20 - 23	3	5
24 - 27	2	15
28 - 31	8	10
32 - 35	12	22
36 - 39	20	40
40 - 43	10	58
44 - 47	5	12
48 - 51	0	5
52 - 55	1	8

9. Completar los datos que faltan en la siguiente distribución:

X_i	n_i	N_i	h_i
1	4	4	0,08
2	4	8	0,08
3	8	16	0,16
4	7	23	0,14
5	5	28	0,10
6	10	38	0,20
7	7	45	0,14
8	5	50	0,10
	50		

$h_i = \frac{n_i}{n}$

10. La duración en horas de una muestra de lámparas incandescentes de 40 watts y 110 voltios en pruebas forzadas de control de calidad fueron las siguientes:

Duración de los elementos.

1,10	1,26	0,83	1,30	1,25	1,49	1,19
1,30	1,23	0,87	1,36	0,94	1,21	1,23
1,24	1,10	1,23	1,32	0,99	1,39	0,95
1,20	0,93	1,18	1,34	0,92	1,02	0,77
1,20	1,05	1,23	1,26	1,20	1,08	1,11
0,82	0,98	1,02	1,03	0,82	0,91	1,16
0,76	1,24	1,00	0,61	0,96	0,83	1,22
1,40	1,33	1,12	1,23	1,25	1,42	1,38
0,94	1,17	0,98	1,69	1,40	0,97	1,30
1,34	0,99	1,08	1,09	1,24	0,87	1,11

10.1 Construir una distribución de frecuencias sobre la duración de las lámparas en la muestra.

1,69 − 0,61 = 1,08

c = 1,08 / ... = 0,108

10.2 Cuántas clases usaría usted para la distribución de frecuencias? Justifique el número de clases que haya seleccionado.

10.3 Usó usted intervalos de clase iguales o desiguales? Por qué?

10.4 Se le presentó algún problema especial al seleccionar los límites de clase?

10.5 Represente gráficamente el histograma y polígono de frecuencias.

10.6 Represente gráficamente la distribución de frecuencias acumulativa.

10.7 En los puntos 10.5 y 10.6, trabajó usted con absolutas o relativas? Explique su elección.

10.8 Analice sus gráficos y escriba un breve comentario para presentar sus descubrimientos.

11. Tome 15 monedas y después de mezclarlas, eche sobre una superficie plana y anote el número de caras resultantes. Repita este experimento 100 veces y construya con los resultados una distribución de frecuencias.

12. Cuarenta familias escogidas al azar, que viven en un campamento minero, incluían los siguientes números de miembros por familia:

6, 4, 5, 4, 4, 3, 5, 3, 6, 5, 4, 5, 5, 6, 3, 4, 3, 2, 6, 3,
4, 3, 9, 3, 5, 3, 2, 4, 5, 3, 5, 3, 5, 3, 11, 3, 4, 4, 3, 5.

12.1 Hacer un histograma de frecuencias, que muestre las frecuencias de las familias según el número de miembros.

12.2 Construir un polígono de frecuencias basado en los datos anteriores.

12.3 Qué da mayor imagen de los distintos tamaños de las familias, el histograma o el polígono de frecuencias? Por qué?

12.4 Qué tipo de datos se han tabulado?

CAPITULO 3

MEDIDAS DE TENDENCIA CENTRAL

3.1

Tipos de Promedio

La distribución de frecuencias, tal como hemos estudiado en el capítulo anterior, no sólo es un resumen de los datos observados, también ella nos muestra la forma en que se distribuye la población; pero es más, cada uno de los valores incluidos en la Tabla de Distribución de Frecuencias, proporciona una información estadística. De manera que se tiene un conjunto de datos estadísticos descriptivos, ya que cada uno de ellos nos describen la densidad de observaciones que caen en una clase o varias clases. Sin embargo, frecuentemente se necesita tener una sola medida que describa la naturaleza de los datos en su conjunto, es decir, un número simple que a su vez sea "representativo" de todas las observaciones. Ser representativo significa, que el número debe reflejar la tendencia de los valores individuales que están distribuidos alrededor de cierto valor central. Es obvio que el valor más representativo para un conjunto de números normalmente no es el valor más pequeño ni el más grande, sino que es un número cuyo valor está en algún punto intermedio del grupo. Por esta razón, un número representativo es aquél que indica una *medida de tendencia central*, conocido comunmente como *promedio*.

El promedio se emplea con frecuencia como mecanismo para resumir un conjunto de cantidades o números, sobre todo si es grande, a fin de describir los datos estadísticos. Así por ejemplo, la edad promedio de los estudiantes de una universidad, el salario mensual promedio de los artesanos en una ciudad; el promedio de quintales de papa producida por hectárea, el promedio de gastos en consumo de los trabajadores mineros, etc. Los promedios también se utilizan para comparar una población con otra; por ejemplo, el promedio general de calificaciones obtenidas en un colegio comparado con el promedio de otro colegio; comparar los rendimientos promedios de unidades producidas por distintas fábricas, etc.

Un promedio se dice también que es una *medida de posición*, en razón de que las formas en que están distribuidas las series estadísticas pueden ser iguales, diferenciándose únicamente en el valor promedio, como se verá en las secciones que siguen.

Las medidas de tendencia central más utilizadas son: la Media Aritmética, la Media Geométrica, la Media Armónica y la Media Cuadrática. Se incluye también en esta categoría a la Mediana y la Moda, que tienen su característica particular a diferencia de las tres primeras.

3.2 Media Aritmética

La medida de tendencia central más útil y la más usada, es la "Media Aritmética", que algunas veces se la llama simplemente "media" o "promedio".

DEFINICION 3.1 La *media aritmética* es la suma de todos los valores observados dividido por el número de observaciones. Si $x_1, x_2, \ldots x_n$ son los valores observados de una muestra, entonces la media aritmética (designada por $\bar{x}$) es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Algunas veces utilizaremos un "operador" que indique la operación de calcular una medida. Así, para la media aritmética, se usará la letra M, de manera que M(X) nos representará la media aritmética de la variable X (que es una característica cualquiera de la población); entonces,

$$M(X) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

Si se tomaran a todos los valores de la población, tendríamos

$$M(X) = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (3.2)$$

Las fórmulas (3.1) y (3.2) se utilizan para calcular la media aritmética, cuando se tiene la información original, es decir, sin necesidad de haberlos agrupado en una tabla de distribución de frecuencias.

Ejemplo 3.1 Supongamos que se tiene el siguiente conjunto de calificaciones que obtuvieron 10 alumnos en una prueba de matemáticas: 3, 4, 5, 4, 6, 5, 4, 6, 7, 3. Entonces, si X es la variable que indica las calificaciones, tendremos

$$M(X) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{10} = \frac{x_i + x_2 + \ldots + x_{10}}{10}$$

$$= \frac{3+4+5+4+6+5+4+6+7+3}{10} = \frac{47}{10} = 4{,}7$$

Luego 4,7 es el promedio de las calificaciones de los 10 alumnos.

En el caso de que se tenga una distribución de frecuencias absolutas, la media aritmética se calculará, como sigue:

$$M(X) = \bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n}$$

donde m indica el número de clases o valores distintos que toma la variable y n_i las veces que se repite un valor x_i distinto.

Ejemplo 3.2 Tomemos la información del ejemplo 3.1, pero esta vez se tiene a los datos agrupados en una tabla, tal como sigue:

TABLA 3.1 Frecuencias de calificaciones de 10 alumnos.

Calificación X	Número de alumnos n_i	Frecuencia h_i
$x_1 = 3$	$n_1 = 2$	$h_1 = 0{,}20$
$x_2 = 4$	$n_2 = 3$	$h_2 = 0{,}30$
$x_3 = 5$	$n_3 = 2$	$h_3 = 0{,}20$
$x_4 = 6$	$n_4 = 2$	$h_4 = 0{,}20$
$x_5 = 7$	$n_5 = 1$	$h_5 = 0{,}10$
$\sum$	10	1

Entonces, la media aritmética es

$$M(X) = \bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{10} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_5 n_5}{10}$$

$$= \frac{3 \times 2 + 4 \times 3 + 5 \times 2 + 6 \times 2 + 7 \times 1}{10}$$

$$= \frac{6+12+10+12+7}{10} = \frac{47}{10} = 4{,}7$$

Cuando se tiene la distribución de frecuencias relativas, el cálculo de la media, será

$$M(X) = \bar{x} = \sum_{i=1}^{m} x_i h_i \qquad (3.4)$$

donde

$$h_i = \frac{n_i}{n}$$

Ejemplo 3.3 Consideremos nuevamente la información del Ejemplo 3.1 y calculemos la media con las frecuencias de la Tabla 3.1

$$M(X) = \bar{x} = \sum_{i=1}^{5} x_i h_i = x_1h_1 + x_2h_2 + \ldots + x_5h_5$$

$$= 3(0,2)+4(0,3)+5(0,2)+6(0,2)+7(0,1)$$

$$= 0,6+1,2+1,0+1,2+0,7 = 4,7$$

Obsérvese que por los tres métodos el resultado es el mismo.

La expresión (3,3), puede también escribirse

$$M(X) = \bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{m} n_i}$$

expresión que recibe el nombre de media aritmética *ponderada*, cuyos pesos o ponderaciones son las frecuencias absolutas n_i. La media aritmética expresada en (3.4) también resulta ser una media ponderada con sus frecuencias relativas h_i.

Hasta aquí, vimos el cálculo de la media, tanto para datos sin agrupar como para datos agrupados. En este último caso sólo para cuando las clases son únicas. Nos preguntamos ahora, cómo se calculará si se tiene una distribución con intervalos de clase. No hay ningún problema, pues, para ello se toman los puntos medios de clase como valores de la variable, y se calcula de manera idéntica a los ejemplos anteriores.

Ejemplo 3.4 Calculemos la media aritmética tomando en cuenta la información de las columnas 2 y 3 de la Tabla 2.6. Como los valores del punto medio de clase, juegan el mismo papel de la variable, entonces

$$M(X) = \frac{\sum x_i n_i}{100} = \frac{(287,5)4+\ldots+(422,5)6}{100} = \frac{36.085}{100} = 360,85 \text{ \$}$$

Existen varios métodos abreviados de cálculo, que en la actualidad pierden algo de importancia, dado que casi siempre es posible contar con una calculadora sea mecánica o electrónica. De manera que le prestaremos más atención a las propiedades de la media y sus consecuencias. Es increible cómo ayuda el tener dominio de todo lo relacionado con la media aritmética, en especial para estudios posteriores.

3.3 Propiedades de la Media Aritmética

La media aritmética es un valor representativo en el sentido de que es el centro de gravedad puesto que balancea a todos los valores de las observaciones que están a ambos lados de esta medida.

Una primera observación, es que conociendo la media aritmética y sabiendo cuál es el número de observaciones, fácilmente se obtiene el agregado o total de los valores, es decir

$$\text{Como} \quad \bar{x} = \frac{\sum x}{n} \Rightarrow \sum x = n\bar{x} \tag{3.5}$$

Por ejemplo, si se sabe que en promedio un vendedor de loterías gana una comisión de 35 \$ diarios, entonces, el distribuidor de billetes de loterías, pagará a sus 120 vendedores, 35 x 120 = 4.200 \$

$\bar{x} \cdot n = \sum x$

Antes de ver otras propiedades, definamos lo que se entenderá por desvíos (o desviaciones). Un *desvío* es la diferencia de un valor observado respecto de algún otro valor. Los dos tipos de desvíos más importantes son aquellos en que el punto fijo de referencia es el origen (o punto cero) y aquellos en que el punto de referencia es la media.

<u>Desviaciones respecto de la media</u>: Si x_i (i=1, 2, ..., n) son los n valores observados, entonces los desvíos respecto de su media, se define

$$d_i = x_i - \bar{x} \quad \text{para cada } i=1, 2, \ldots, n \tag{3.6}$$

1) La propiedad más importante de estos desvíos, es que la suma de todos ellos es nula. Es decir, para datos sin agrupar se tiene

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{3.7}$$

y para datos agrupados

$$\sum_{i=1}^{m} d_i \, n_i = \sum_{i=1}^{m} (x_i - \bar{x}) \, n_i = 0$$

Se demuestran fácilmente aplicando las propiedades de sumatorias, por ejemplo para (3.7), se tiene

$$\sum d_i = \sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

Como una consecuencia lógica de esta última parte tenemos que la media aritmética de los desvíos es nula, o sea

$$M\,(d) = \frac{\sum d_i}{n} = \frac{0}{n} = 0 \qquad (3.8)$$

Podemos construir una tabla de distribución de frecuencias, incluyendo una columna para los desvíos respecto de la media. Así, para la Tabla 3.1 tendremos.

TABLA 3.2 Desvíos respecto de la media

$\bar{x} = 4{,}7$

Observaciones x_i	n_i	Desvíos $(x_i - \bar{x})$
3 4.7	2	- 1,7
4	3	- 0,7
5	2	0,3
6	2	1,3
7	1	2,3

2) Si todos los valores observados son iguales, naturalmente que su media será el valor común. Si K representa el valor observado de todas las unidades elementales, entonces

$$M\,(K) = \frac{\sum K}{n} = \frac{nK}{n} = K \qquad (3.9)$$

Esto es lógico. Por ejemplo, si todos los empleados de una oficina ganan 2.000 \$, entonces la media de lo que ganan es 2.000 \$.

3) Si existe un factor constante en todos los valores observados, entonces,

$$M\,(KX) = K \cdot M\,(X) \qquad (3.10)$$

4) La media de la suma de dos variables es equivalente a la suma de sus medias individuales, de esta manera,

$$M(X+Y) = M(X) + M(Y) \qquad (3.11)$$

Si se tienen r submuestras y se calculan sus respectivas muestras, entonces, la media general de la muestra se obtiene promediando las medias submuestrales ponderadas con sus tamaños correspondientes, es decir:

$$M(X) = \frac{n_{(1)} M\left[X_{(1)}\right] + \dots + n_{(r)} M\left[X_{(r)}\right]}{n_{(1)} + \dots + n_{(r)}} \qquad (3.12)$$

donde $M\left[X_{(j)}\right]$ es la media de la submuestra j-ésima

$n_{(j)}$ es el tamaño de la submuestra j-ésima

y $\sum_{j=1}^{r} n_{(j)} = n$ tamaño de la muestra.

Las identidades (3.10), (3.11) y (3.12), se demuestran fácilmente aplicando la definición de media aritmética y las propiedades de sumatorias

Ejemplo 3.5 La media de los salarios en una fábrica A que cuenta con 125 obreros es 875 \$; y la media de los salarios de otra fábrica similar B, que tiene 53 obreros es 1.050 \$. Cuál es el salario promedio de los trabajadores de ambas fábricas?. La solución es sencilla. Tomando la relación (3.12) se tiene

$$M(X) = \frac{125 \times 875 + 53 \times 1050}{125 + 53} \quad \frac{165.025}{178} = 927{,}1 \text{ \$}$$

3.4 Media Geométrica

DEFINICION 3.2 La *media geométrica* se define como la raíz enésima del producto de los n valores observados, o sea,

$$G(X) = \sqrt[n]{x_1 \, x_2 \dots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad (3.13)$$

cuando los datos no están agrupados.

Si se tiene una distribución de frecuencias, entonces

$$G(X) = \sqrt[n]{\prod_{i=1}^{m} x_i^{n_i}} \qquad (3.14)$$

donde m es el número de valores distintos que toma la variable.

Cuando el número de observaciones es mayor que dos, se pueden transformar las fórmulas anteriores mediante la aplicación de logaritmos decimales, o sea:

$$\log G(X) = \frac{1}{n} \sum_{i=1}^{n} \log x_i \qquad (3.15)$$

y

$$\log G(X) = \frac{1}{n} \sum_{i=1}^{m} n_i \log x_i \qquad (3.16)$$

que sirven para calcular la media geométrica para datos sin agrupar y para datos agrupados, respectivamente.

Ejemplo 3.6 Determinemos la media geométrica para la información del Ejemplo 3.1. Como no ofrece dificultad calcular G (X) directamente con los datos observados, veamos cómo computaríamos si se tienen los datos agrupados (Tabla 3.3)

TABLA 3.3 Cálculo de la media geométrica para datos agrupados.

x_i	n_i	$\log x_i$	$n_i \log x_i$
3	2	0,47712	0,95424
4	3	0,60206	1,80618
5	2	0,69897	1,39794
6	2	0,77815	1,55630
7	1	0,84510	0,84510
	10		6,55976

de manera que $\log G(X) = \frac{1}{10} (6,55976) = 0,655976$

Luego $G(X) \doteq 4,53$

La aplicación más útil de la media geométrica, es para promediar razones de cambio.

<u>Ejemplo 3.7</u> Supongamos que la población de cierta ciudad tuvo un incremento de 100.000 a 250.000 habitantes durante el período 1920 - 1970. Cuál es la razón promedio de cambio por década?. Si calculamos mediante la media aritmética, tendremos 150 % $\div$ 5 = 30%; medida que no es correcta, puesto que la población crece con razón compuesta. La siguiente Tabla muestra cómo se obtiene la razón promedio de cambio correcta:

TABLA 3.4 Cálculo de la media geométrica de cambio de la población de cierta ciudad para las décadas 1920 - 1970

Años	Población	Cambios relativos de la población (x)	Logaritmo de los cambios relativos
1920	100.000		
1930	115.000	1,150	0,06070
1940	140.000	1,217	0,08529
1950	175.000	1.250	0,09691
1960	215.000	1.229	0,08955
1970	250.000	1,163	0,06558
Total			0,39803

$$\text{Luego } \log G(X) = \frac{\sum \log x}{n} = \frac{0{,}39803}{5} = 0{,}0796$$

de donde

$$G(X) = 1{,}201, \text{ o sea alrededor de } 20{,}1\% \text{ por década.}$$

La razón promedio de crecimiento de la población para la última década es

$$\log G(X) = \frac{\sum \log x}{n} = \frac{0{,}06558}{10} = 0{,}006558$$

o sea $G(X) = 1{,}015$, o aproximadamente 1,5 % por año.

Estos cálculos de las razones de cambio promedio, se basan bajo el supuesto de que la razón de cambio es constante. Cuando el cálculo implica un número considerable de años, generalmente se usa la fórmula

$$P_n = P_o (1 + r)^n$$

donde r es la razón de cambio.

3.5

Media Armónica

DEFINICION 3.3 La *media armónica* de una serie de valores $x_1, x_2, \ldots, x_n$, se define como el recíproco de la media aritmética de los recíprocos de los valores observados. Así, para datos sin agrupar, se tiene

$$H(X) = \frac{1}{\dfrac{1/x_1 + 1/x_2 + \ldots + 1/x_n}{n}} = \frac{1}{\dfrac{\sum_{i=1}^{n} 1/x_i}{n}} = \frac{n}{\sum 1/x_i} \qquad (3.17)$$

y para datos agrupados

$$H(X) = \frac{1}{\dfrac{\sum_{i=1}^{m} \dfrac{1}{x_i} n_i}{n}} = \frac{n}{\sum_{i=1}^{m} n_i/x_i} \qquad (3.18)$$

Ejemplo 3.8 Nuevamente tomando la información del Ejemplo 3.2, tendremos:

$$H(X) = \frac{10}{\frac{2}{3} + \frac{3}{4} + \frac{2}{5} + \frac{2}{6} + \frac{1}{7}} = \frac{10}{2{,}293} = 4{,}36$$

De la misma manera que la media aritmética y la media geométrica, la media armónica es también afectada por los valores de cada elemento observado. La media aritmética es la más afectada por los valores extremos que la media geométrica, y ésta a su vez más afectada que la media armónica. Las magnitudes de las tres diferentes medidas para los mismos datos, tienen la siguiente relación:

$$M(X) \geqslant G(X) \geqslant H(X)$$

Si observamos los resultados de los ejemplos 3.2, 3.6 y 3.8, efectivamente,

$$M(X) = 4{,}7 > G(X) = 4{,}53 > H(X) = 4{,}36$$

GEOMETRICA

La media armónica es útil para promediar principalmente razones. Es el caso en que se trata de promediar velocidades y tiempos promedio de trabajo.

Ejemplo 3.9 Una fábrica de muebles de madera, ha asignado a cinco de sus trabajadores para completar una orden de 200 sillas de un cierto tipo. Las razones de productividad de los cinco trabajadores están dados como sigue:

TABLA 3.5

Trabajador	Tiempo de Producción
A	5 horas por silla
B	8 horas por silla
C	6 horas por silla
D	12 horas por silla
E	4 horas por silla

Se trata de encontrar el promedio de horas por silla para el grupo de trabajadores.

Utilizando como medida la media armónica, tenemos

$$H(X) = \frac{5}{\frac{1}{5}+\frac{1}{8}+\frac{1}{6}+\frac{1}{12}+\frac{1}{4}} = \frac{600}{24+15+20+10+30} = 6{,}0606 \text{ horas por silla}$$

Este resultado es correcto solamente si suponemos que cada uno de los obreros ha trabajado el mismo número de horas, o sea,

$$6{,}0606 \times 200 = 1.212{,}12 \text{ horas (requeridas para completar la orden)}$$

luego, $\frac{1.212{,}12}{5} = 242{,}42$ horas por trabajador

En la práctica, es posible que cada obrero de una fábrica trabaje la misma cantidad de tiempo pero produzca diferente número de sillas. Es justamente en este caso que la media armónica es más realista que la media aritmética, ya que

$$M(X) = \frac{5+8+6+12+4}{5} = 7 \text{ horas por silla}$$

supondría que a cada uno de los trabajadores se le asignó el mismo número de sillas para cumplir la orden.

3.6 Media Cuadrática

DEFINICION 3.4 La *media cuadrática* de n valores se define como la raíz cuadrada de la media aritmética de los cuadrados de las observaciones.

Para datos sin agrupar, se tiene

$$M_c\ (X) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{3.20}$$

y para datos agrupados

$$M_c\ (X) = \sqrt{\frac{\sum_{i=1}^{m} x_i^2\, n_i}{n}} \tag{3.21}$$

Ejemplo 3.10 Tomando la información de la Tabla 3.1, se tiene que

$$M_c = \sqrt{\frac{9 \times 2 + 16 \times 3 + 25 \times 2 + 36 \times 2 + 49 \times 1}{10}} = 4,87$$

La aplicación de la media cuadrática se verá en el proximo capítulo, cuando estemos tratando los cálculos de las medidas de variación.

3.7 Mediana

Hemos visto que las medidas anteriores están influenciadas particularmente por los *valores* de las observaciones. De esta manera, cuando hay valores extremos, generalmente la media no es una buena medida, así por ejemplo, si la producción diaria de un obrero es normal durante 4 días de la semana, y el 5o. día tiene un rendimiento casi nulo, su rendimiento medio desciende considerablemente. Esta influencia profunda de los valores extremos sobre la media aritmética, implica que este promedio frecuentemente no proporcione una medida significativa de la tendencia central, es decir, que indique un punto cercano a aquél en que la mayor parte de los elementos están localizados, si la distribución es marcadamente oblicua. En estos casos, a menudo se utiliza otro tipo de medida que no está influenciada por los valores extremos, o sea que es un valor que está situado en el centro del número de datos. Esta medida recibe el nombre de mediana.

DEFINICION 3.5 La *mediana* (M_e) es un valor de las observaciones que divide en dos partes iguales al número total de observaciones, cuando éstos están ordenados de acuerdo a sus valores.

Es decir que cuando la serie de observaciones está *ordenada* según sus valores, la mediana es un valor observado, tal que antes y después de éste valor, hay el mismo número de observaciones. Así por ejemplo, supongamos que se tienen los siguientes valores observados: 4,5,8,2,7,2,3. Previamente se los ordena, o sea, 2,2,3,4,5,7,8; luego, la mediana es 4. Ya que antes del 4 hay 3 observaciones, igualmente hay 3 observaciones por encima del 4.

Cuando se tiene un número par de observaciones ordenados, por ejemplo, 3,4,6,7,7,8, no hay un valor observado que satisfaga plenamente la definición de mediana. En este caso, podría ser cualquier valor comprendido entre 6 y 7; pero en la práctica, se usa el punto medio de entre los dos valores centrales, por tanto, la mediana será $M_e = \frac{6+7}{2} = 6,5$

La característica típica de la mediana es que divide al conjunto de observaciones en dos partes iguales; es decir que el 50% de las observaciones tienen valores menores que la mediana, y el otro 50% de las observaciones tienen valores mayores. En cierto sentido, la mediana es también un punto de equilibrio, puesto que balancea al número de elementos de la serie estadística.

Una propiedad interesante que tiene la mediana es que ,"la suma de los desvíos respecto de la mediana en valor absoluto, es menor que la suma de los desvíos respecto de cualquier otro punto en la distribución." Es decir que $\sum \left| x_i - M_e \right|$ es un mínimo. Una aplicación simple de esta propiedad, es la localización óptima de un punto.

Ejemplo 3.11 Supongamos que en las inmediaciones de una carretera hay siete poblaciones, cuyas distancias en kilómetros son la siguientes:

Poblaciones:		P1	P2	P3	P4	P5	P6	P7
Kilómetros:	0	10	42	72	94	130	136	150

Cada población debe surtirse diariamente de un frigorífico distribuidor de carne. Para minimizar el total de las distancias de las siete poblaciones al frigorífico, ¿dónde debiera estar localizado éste? Natural -

mente que no serán las poblaciones extremas, como P1 o P7. Recordando que la suma de los desvíos absolutos respecto de la mediana es un mínimo, es fácil decidir que el frigorífico será localizado en la poblción P4 (o en su vecindad). La mediana es 94 kilómetros, ya que el camino total a recorrerse hasta este punto de las otras poblaciones es de 292 Km. Si se localiza, el frigorífico en cualquier otro punto, la suma de las distancias será mayor. Por ejemplo, para la población P3, la suma es 314 Kms. La siguiente Tabla nos muestra dichos cómputos:

TABLA 3.6 Cálculo de la distancia mediana

Población	Localización	Distancia de P3	Distancia de P4
P1	10	62	84
P2	42	30	52
P3	72	0	22
P4	94	22	0
P5	130	58	36
P6	136	64	42
P7	150	78	56
Total Kms.		314	292

Si se probara localizar en cualquier otra población la distancia total a recorrerse, será mayor que si el frigorífico está localizado en la "pobla - ción 4".

Hasta aquí, unicamente nos hemos referido a la determinación de la mediana, cuando los datos no están agrupados. Ahora nos toca ver cómo se calculará en el caso de tener una distribución de frecuencias, cuando la clase es única y cuando se tienen intervalos de clase.

1) <u>Cuando la clase es única</u>: Sea N_j la j-ésima frecuencia absoluta acumulada inmediata superior a $\frac{n}{2}$, entonces:

a) Si $\frac{n}{2} > N_{j-1} \Longrightarrow M_e = x_j$

b) Si $\frac{n}{2} = N_{j-1} \Longrightarrow M_e = \frac{x_{j-1} + x_j}{2}$ (3.22)

donde x_j y x_{j-1} son los valores j-ésimo y (j-1)-ésimo de las N_j y N_{j-1} respectivamente. Esto mismo, en términos gráficos (Ver figura (3.1)

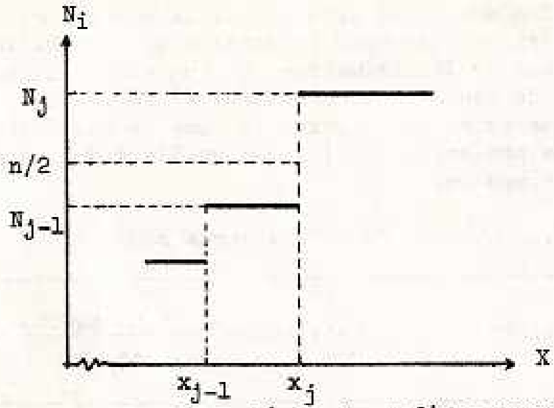


Figura 3.1 Localización de la mediana para una distribución de frecuencias como el de las Tablas 2.3 y 2.4

Entonces, para calcular la mediana se debe determinar previamente $\frac{n}{2}$, y luego ver si éste valor es menor o igual que una j-ésima frecuencia acumulada, y luego seguir la regla (3.22)

<u>Ejemplo 3.12</u> Utilicemos la información de la Tabla 2.3, que es:

TABLA 3.7

Calificación X	Frecuencias Absolutas n_i	Frecuencias Absolutas Acumuladas
10	4	4
20	6	10
30	11	21 25
40	12	33
50	8	41
60	7	48
70	2	50

Vemos que $\frac{n}{2} = 25$ es un valor que está comprendido entre 21 y 33 de las frecuencias acumuladas, es decir $21 < 25 < 33$ (lo que equivale en símbolos a $(N_{j-1} < \frac{n}{2} < N_j)$.

Entonces, como $25 > 21$, o sea $\frac{n}{2} > N_{j-1}$, por a) de (3.22), se tiene $M_e = x_j$ en consecuencia, $M_e = 40$

2) <u>Cuando se tienen intervalos de clase</u>: Bajo los mismos supuestos del caso (1) la mediana se obtiene como sigue:

$$\text{a)}\quad \text{Si } \frac{n}{2} > N_{j-1} \Longrightarrow M_e = x_{j-1} + c_j \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}} \qquad (3.23)$$

$$\text{b)}\quad \text{Si } \frac{n}{2} = N_{j-1} \Longrightarrow M_e = x_{j-1}$$

donde c_j es el tamaño del intervalo de clase. Todo esto, se deduce de la Figura 3.2

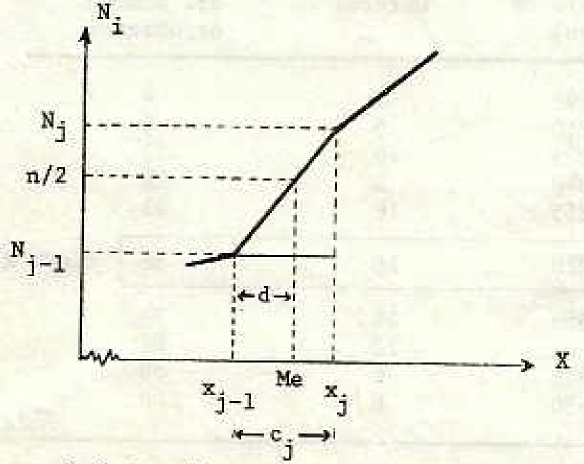


Figura 3.2 Localización de la mediana cuando se tiene una variable continua o una distribución con intervalos de clase.

Vemos que la mediana será

$$M_e = x_{j-1} + d$$

El cálculo de <u>d</u>, es como sigue:

$$\frac{N_j - N_{j-1}}{c_j} = \frac{\frac{n}{2} - N_{j-1}}{d}$$

de donde

$$d = c_j \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}}$$

Entonces, para calcular la mediana cuando se tiene una tabla de distribución de frecuencias con intervalos de clase, es necesario utilizar la columna de frecuencias absolutas acumuladas, luego encontrar la clase que contiene la mediana y finalmente interpolar la mediana a partir de los valores incluidos en dicha clase. Este procedimiento lo ilustraremos en el siguiente ejemplo:

<u>Ejemplo 3.13</u> Tomemos la Tabla 2.6 que corresponde a la distribución de frecuencias de salarios semanales de 100 obreros.

TABLA 3.8

Ingresos (intervalo de clase)	Número de Obreros	Acumulación del número de obreros	
280 - 295	4	4	
295 - 310	5	9	
310 - 325	10	19	
325 - 340	9	28	
340 - 355 X_{j-1}	13	41 N_{j-1}	
355 - 370	15	56 N_j	Clase mediana
X_{j-1} 370 - 385 X_j	18.	74	
385 - 400	12	86	
400 - 415	8	94	
415 - 430	6	100	

Hay 100 obreros en la distribución. La mediana debe ser al final del 50 avo elemento en la distribución. Entonces $N_j = 56$ y $N_{j-1} = 41$, luego $M_e = 355 + d$.

Como $c_j = 15$, entonces, aplicando el cálculo de <u>d</u>, tenemos

$$d = 15 \cdot \frac{50 - 41}{56 - 41} = 9$$

En consecuencia, $M_e = 355 + 9 = 364$ \$

3.8 Cuartiles, Deciles, Percentiles

Como hemos visto, cuando se trata de encontrar un valor que divida al número total de observaciones en dos partes iguales, la medida adecuada para este fin es la mediana. Pero otras veces puede uno estar interesado en hallar un valor que esté situado, ya no al centro del número de observaciones, sino por ejemplo, en la primera cuarta parte o en la tercera cuarta parte. Las medidas para este fin, se denominan *cuartiles*. El método para hallar un cuartil, es completamente análogo al caso de la mediana; o sea utilizando $\frac{n}{4}$ para el primer cuartil, y $\frac{3n}{4}$ para el tercer cuartil. Podemos observar, de inmediato, que el segundo cuartil, lógicamente, coincidirá con la mediana, puesto que $\frac{2n}{4}=\frac{n}{2}$.

Si los porcentajes de división fueran del 10%, entonces las medidas que separen el número de observaciones en décimas de parte, se llaman *deciles*. En este caso también, se calcula en forma idéntica que para la M_e, con tal que se use $\frac{n}{10}$ para el primer decil, $\frac{2n}{10}$ para el segundo decil, etc.

Finalmente, los *centiles* se interpretan en forma análoga, utilizando $\frac{n}{100}$ Esto es lo que se llama también percentiles o porcentiles.

3.9 Moda

Otra medida de tendencia central, es la moda o (modo o valor modal), que se define como sigue:

DEFINICION 3.6 La *moda* es un valor distinto observado de una serie estadística que se repite más veces.

De esta manera, el valor modal es el valor más frecuente en una serie de datos. Es evidente que la moda no siempre estará en el centro, sino que, como ocurre a menudo, puede ser un valor extremo.

La moda para datos no agrupados de unos pocos valores puede obtenerse por simple inspección, así por ejemplo, si tenemos las series:

2, 2, 5, 7, 9, 9, 9, 9, 11, 12 :   la moda es $M_o = 9$

3, 5, 8, 10, 12, 15, 16 :   no existe la moda

2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9:   tiene dos modas, o sea $M_o = 4$ o 7

Cuando hay dos modas, se dice *bimodal*, y en el caso de existir más de dos se conoce con el nombre de *multimodal*.

Si se tienen datos agrupados, tal que los valores de la clase son únicos, entonces la moda, es obvio que sea la que tiene mayor frecuencia, así en el ejemplo 3.2, la moda es 4.

Cuando se tiene una distribución de frecuencias con intervalos de clase, la moda se determina por interpolación, es decir

$$M_o = x_{j-1} + \left(\frac{d_1}{d_1 + d_2} \right) c \qquad (3.24)$$

donde x_{j-1} = límite inferior de la clase modal (o sea el intervalo que tiene mayor frecuencia)

d_1 = diferencia entre la frecuencia de la clase modal y la frecuencia premodal

d_2 = diferencia entre la frecuencia de la clase modal y la frecuencia de la clase post modal.

c = tamaño del intervalo de la clase modal.

Esto mismo, gráficamente, se tiene:

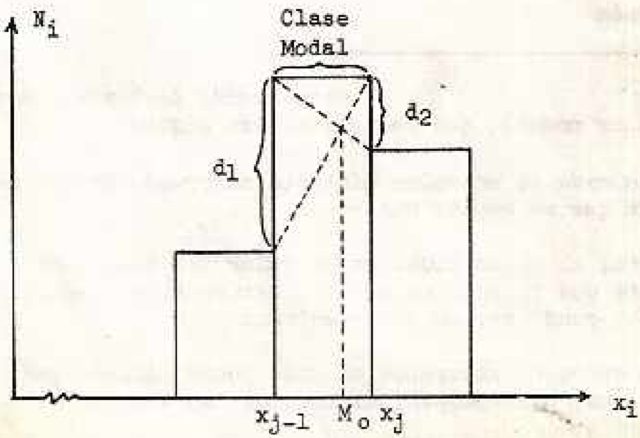


Figura 3.3 Interpolación geométrica de la M_o, para una distribución con intervalos de clase.

<u>Ejemplo 3.14</u> Determinemos la Moda para la información del ejemplo 3.13

Vemos que en la Tabla 2.6 (la misma del ejemplo 3.13), la clase modal es 370-385, ya que en este intervalo hay el mayor número de observaciones. Entonces

$$x_{j-1} = 370$$
$$d_1 = 18 - 15 = 3$$
$$d_2 = 18 - 12 = 6$$
$$c = 15$$

Ahora, aplicando la fórmula (3,24), tenemos

$$M_o = 370 + \frac{3}{3+6} \cdot 15 = 375 \ \$$$

Hagamos una observación importante: si se tomara directamente la información original (Tabla 2.5), hay valores de los ingresos que se repiten más veces que 375, así por ejemplo, el 380 se repite 7 veces, mientras que 375 solamente 5. Dejamos al estudiante la interpretación y análisis de este hecho.

3.10 Relaciones entre Media, Mediana y Moda

De acuerdo a la simetría o asimetría de los datos observados, podemos ver la relación empírica entre la media, mediana y moda (Figura 3.4).

En general, cuando la distribución es casi simétrica, la mediana se localiza aproximadamente a un tercio de la distancia entre la media y la moda. La fórmula que describe esta relación es

$$\text{Moda} = \text{Media} - 3\ (\text{Media} - \text{Mediana}) \qquad (3.25)$$

Para el ejemplo de los salarios semanales de los 100 obreros, la moda aproximada será

$$M_o = 360{,}85 - 3\ (360{,}85 - 364) = 370{,}3$$

Es importante notar que la moda obtenida por este método aproximado, no necesariamente coincide con la moda calculada mediante la interpolación.

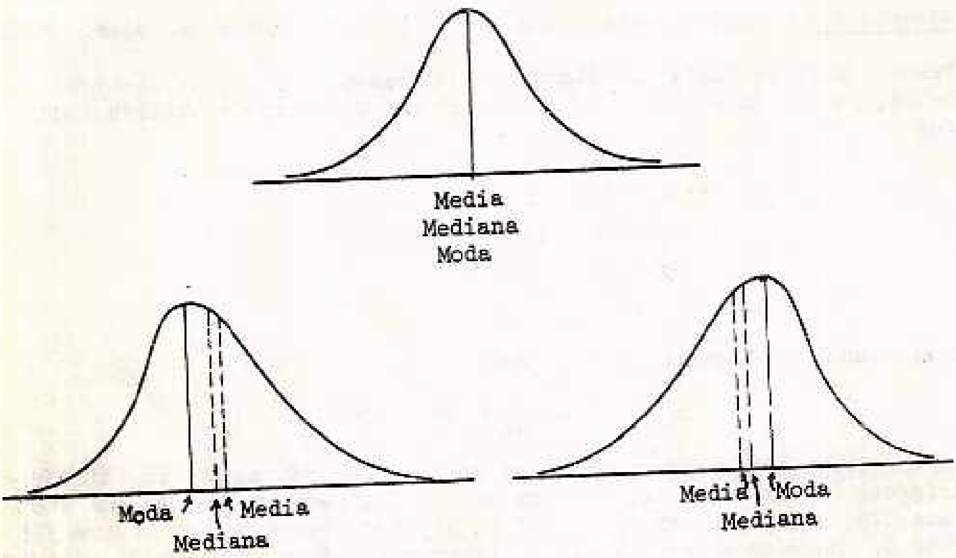


Figura 3.4 Ejemplos de distribuciones de frecuencias: simétrica, asimétrica a la derecha y asimétrica a la izquierda.

3.11

Ejercicios

1. Calcular la media aritmética, mediana y la moda, tomando la información de los ejercicios 8, 10, 11 y 12 del Capítulo 2.

2. Calcular la media, la mediana y la moda de las siguientes distribuciones de frecuencias:

2.1

Puntajes		No. de alumnos	
40 - 44	42	2	2
45 - 49	47	0	0
50 - 54	52	5	7
55 - 59	57	7	14
60 - 64	62	9	23
65 - 69	67	11	34
70 - 74	72	6	40
75 - 79	77	8	48
80 - 84	82	4	52
85 - 89	87	2	54
90 - 94	92	2	56
		56	

2.2

Salarios $	No.de obreros
60	6
70	1
80	3
132	1
150	1
500	1

Cuál es su opinión sobre los resultados?

3. Señale y justifique la respuesta correcta de las siguientes aseveraciones:

   3.1 La media aritmética debe coincidir con algunos valores que toma la variable, si ésta es discreta.

   3.2 El promedio de las notas de un curso fué de 58. Las mujeres obtuvieron un promedio de 38 y los hombres un promedio de 64. Luego, el curso tiene más de 90 alumnos.

   3.3 Mientras menos intervalos se empleen, más precisa es la media aritmética calculada.

   3.4 El ingreso real medio de los obreros agrícolas es \$ 350, el de los mineros \$ 540 y el de los industriales \$ 640; es decir, el ingreso medio del conjunto de obreros es de \$ 510. V

   3.5 La suma de los desvíos respecto de la mediana es cero.

   3.6 Para calcular la media geométrica, es preciso que todos los intervalos sean iguales.

3.7 El 75% de los empleados públicos son hombres, mientras que en el sector privado es el 81%. Esto es, hay un 22% de mujeres empleadas en ambos sectores.

3.8 La media armónica de una constante es igual al recíproco de la constante.

3.9 Los siguientes datos son consistentes:

a) $h_2 = 0{,}4$; $h_1 = 0{,}2$; $H_3 = 0{,}8$; $n = 50$; $n_4 = 5$;

$\overline{X} = 25$; $M_d = 30$

b) $m = 5$; $\sum_{i=1}^{5} X_i^2 n_i = 400$; $\overline{X} = 5$; $h_5 = 0{,}1$

$H_3 = 0{,}5$; $n_4 = 8$

4. En una oficina de empleados públicos hay 35 hombres con una edad media de 27,5 años y 15 mujeres las que, en promedio, son un 22% más jóvenes. Cuál es la edad media de los empleados?

5. Un auto sube a La Cumbre a 30 Kms por hora y baja a 60 Kms por hora. Cuál es la velocidad media?

6. Las distribuciones de los ingresos de dos ciudades son las siguientes:

CIUDAD "A"		CIUDAD "B"	
Ingresos Anuales por habitante (en miles de $)	Población Remunerada	Ingresos Anuales por habitante (en miles de $)	Población Remunerada
8 - 10	3.000	6 - 9	1.000
10 - 12	8.000	9 - 12	2.000
12 - 14	4.000	12 - 15	5.000
14 - 16	1.000	15 - 18	2.000
16 - 20	400	18 - 21	1.500
20 - 22	100	21 - 24	1.000
		24 - 27	400

6.1 Calcule el ingreso promedio del 40% de la población de mayores ingresos en cada ciudad. Compare los resultados y discuta sus conclusiones. Realice el mismo cálculo para el 50% de la población de menores ingresos.

6.2 Cuál es el ingreso promedio de las dos ciudades en conjunto?

6.3 Suponiendo que la ciudad B devalúa su moneda en 30% y la ciudad A en 5%, calcule los nuevos ingresos promedio en ambos países.

7. Al revisar algunos archivos de datos estadísticos, se halló la siguiente información:

Tipos de Alquiler $b.		Número de Familias
100 - 300	200	20
300 - 500	400	120
500 - 700	600	160
700 - 900	800	
900 - 1.100	1000	30
1.100 - 1.300	1200	10

Se trata de averiguar el número de familias que pagan alquileres entre $b. 700 y $b. 900, sabiendo que el promedio de alquileres es de $b. 670.

8. Un Club de tenis tiene socios en las villas A(12%), B(15%), C(13%), D(25%), E(9%), F(11%), G(7%) y H(8%), situadas a lo largo de una carretera. Las distancias entre las villas son: 6 Km. de A a B; 3 Km. de B a C; 8 Km. de C a D; 7 Km. de D a E; 6 Km. de E a F; 15 Km. de F a G y 9 Km. de G a H. Dónde debe situarse el campo deportivo de manera que minimice los gastos de transporte si todos los socios concurrirán habitualmente. Suponga que el costo de transporte por cada Km. es constante en toda la región.

C A P I T U L O 4

MEDIDAS DE DISPERSION

En el presente capítulo estudiaremos una característica importante de las distribuciones de frecuencias que es el grado de variación o de variabilidad, que algunas veces se llama *dispersión* de los valores observados. Una medida de dispersión es importante desde dos puntos de vista: en primer término, puede ser usada para mostrar el grado de variación entre los valores de los datos observados. Por ejemplo, una pequeña dispersión de los salarios mensuales de un grupo de trabajadores en una fábrica, indicará que a los obreros se les paga salarios aproximadamente iguales; por otro lado, una dispersión grande, dará la idea de que los salarios son muy diferentes. En segundo lugar, puede ser usada para complementar un promedio,para describir un conjunto de datos o para comparar una serie de informaciones con otra. Cuando la dispersión es baja, el valor promedio se vuelve altamente significativo, o sea que la media es un valor realmente representativo; en cambio, si la dispersión es alta, la media se vuelve poco o nada representativa.

Si por ejemplo, se tuvieran los números 1, 3 y 11, la media de ellos es 5; mientras que la media de los números 3, 6 y 6, también es 5. En el primer caso, el valor de la media no está próximo a ningún número del grupo, por tanto, se espera una alta dispersión; en cambio en el segundo caso, la media está próximo a cada número del grupo, por lo que se espera una menor dispersión. El hecho de que la medida de dispersión del primer grupo de números es más alta que el de la segunda, da una mejor comprensión en la comparación de las medidas de los dos grupos de datos (el ejemplo 4.8 aclara más esta idea).

Cuando se habla de dispersión, es natural que se trate de medir las diferencias o comparaciones respecto de algún valor, tal que nos indique en promedio las variaciones que existen entre los valores observados comparados con dicho valor. Generalmente para medir las variaciones se toma como referencia un punto central de los valores observados, tal como la media. Una de las medidas de dispersión más importantes es la desviación estándar, que es una medida que considera qué tan lejos de la media están localizados cada uno de los valores de la serie estadística en observación. De esta manera, parece lógico pensar que mientras menor sea la desviación estándar de una distribución de frecuencias, más significativo será generalmente un promedio como medida de la tendencia central para dicha distribución. El uso de esta medida es grande debido a que nos permite tomar en cuenta los elementos de variación para extraer conclusiones sólidas y tomar decisiones apropiadas de la información disponible. En los capítulos subsiguientes, se verán muchas aplicaciones de esta medida.

Veremos a continuación con cierto detalle, cada una de las distintas medidas de dispersión y sus propiedades.

4.1 Recorrido

La medida más simple de la dispersión de un conjunto de observaciones es el *recorrido* (o rango), que es la diferencia entre el mayor y menor valores observados.

DEFINICION 4.1 Si denominamos R al recorrido, x_{max} al valor máximo observado y x_{min} al menor valor observado, entonces

$$R = x_{max} - x_{min} \tag{4.1}$$

Ejemplo 4.1 De los valores dados en la Tabla 2.5 se observan que,

$$x_{max} = 430 \qquad y \qquad x_{min} = 280$$

luego el recorrido es

$$R = 430 - 280 = 150$$

Esta medida no está afectada por los valores comprendidos entre el valor más alto y más bajo, por tanto, el recorrido no es un estimador muy relevante.

El recorrido es la medida más simple que sirve como complemento de la media, por ejemplo: el promedio de producción de un grupo de trabajadores en una fábrica A, puede ser 50 unidades diarias con un recorrido de 20 a 65 unidades; y el promedio en una fábrica B, puede ser también 50 unidades, pero con recorrido de 35 a 60 unidades. Considerando los dos diferentes recorridos, podemos concluir que el promedio es más representativo de las unidades producidas por los trabajadores de la fábrica B comparado con la producción promedio de la fábrica A.

4.2 Desviación Media

Una primera medida basada en todos los elementos observados y diseñada para medir la dispersión *alrededor de un promedio*, es la llamada *desviación media* o *desviación promedio*. La medida de referencia puede ser la media o la mediana. Por ser de mayor aplicación, veremos la desviación media (DM) respecto de la media.

DEFINICION 4.2 La desviación media es simplemente la media aritmética de los desvíos de los valores individuales con respecto al promedio de ellos mismos.

Como teóricamente la suma de los desvíos respecto de la media es nula (Sección 3.3 del capítulo 3), es decir $\sum d = \sum (x - \bar{x}) = 0$, para el cálculo de la desviación media, se usan los valores absolutos de los desvíos; de esta manera, si se tiene un conjunto de valores sin ordenar, la DM será:

$$DM = M(|d|) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad (4.2)$$

y para datos agrupados,

$$DM = M(|d|) = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| n_i}{n} \qquad (4.3)$$

En ambos casos, se supone que se está trabajando con n observaciones. Como puede apreciarse por (4.2), el cálculo de la DM es muy simple. Primero se calcula la media aritmética de los valores observados; luego los desvíos de las observaciones respecto de la media y, finalmente se determina la media aritmética de los valores absolutos de los desvíos. Ilustraremos este proceso con los siguientes ejemplos:

Ejemplo 4.2 Sea la información del ejemplo 3.1 del capítulo anterior (Ver Tabla 4.1).

TABLA 4.1 Cálculo de la desviación media, para datos sin agrupar

x	$d = x - \overline{x}$	$\lvert d \rvert = \lvert x - \overline{x} \rvert$
3	- 1,7	1,7
4	- 0,7	0,7
5	0,3	0,3
4	- 0,7	0,7
6	1,3	1,3
5	0,3	0,3
4	- 0,7	0,7
6	1,3	1,3
7	2,3	2,3
3	- 1,7	1,7
$\sum$	0	11

De manera que la desviación media es

$$DM = \frac{\sum |d|}{n} = \frac{11}{10} = 1,1$$

Desde el punto de vista práctico, puede obviarse la segunda columna.

Ejemplo 4.3 Sea ahora la información del Ejemplo 3.2 y la Tabla 3.2

TABLA 4.2 Cálculo de la desviación media para datos agrupados

X_i	n_i	$\lvert d_i \rvert$	$\lvert d_i \rvert\, n_i$
3	2	1,7	3,4
4	3	0,7	2,1
5	2	0,3	0,6
6	2	1,3	2,6
7	1	2,3	2,3
$\sum$	10		11

Luego, $$DM = \frac{\sum |d_i|\, n_i}{n} = \frac{11}{10} = 1,1$$

Como se ha visto, la DM considera a todos los valores observados para su cálculo, por lo tanto, da una mejor descripción de la dispersión. De manera que nos mide la dispersión alrededor del valor promedio. Sin embargo, esta medida es un poco débil puesto que se ignora el signo positivo o negativo de las desviaciones.

Una medida más fina o más adecuada de la dispersión, es la que recibe el nombre de *desviación estándar.*

4.3

Desviación Estándar y Varianza

La medida del grado de variación de una distribución de frecuencias más comunmente usada en el análisis estadístico, es la *desviación estandar* *. Es la más importante de entre todas las medidas de dispersión, de ahí que, le prestaremos la mayor atención posible, puesto que de ello dependerá nuestro éxito cuando se hagan aplicaciones y se puedan tomar conclusiones sólidas de decisión. Una medida que en cierto modo es previa a la desviación estandar, es la *varianza*, que es una medida de dispersión de excepcional importancia. La ventaja de estas dos medidas, a diferencia de la DM, es que consideran a los desvíos con sus respectivos signos.

Cuando se toman a *todos* los elementos de una población en estudio, se usan los símbolos σ^2 y σ, respectivamente, para indicar a la varianza poblacional y desviación poblacional. En cambio, si los valores provienen de una muestra de n observaciones, se usará s^2 y s, para indicar a la varianza muestral y desviación estandar muestral, respectivamente. En general, usaremos el operador V(X) para indicar la varianza de una variable X que indica una característica cualquiera de la población en estudio.

DEFINICION 4.3 La *varianza* de n observaciones de una variable X, está definida como sigue:

$$s^2 = V\ (X) = M(d^2) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = M\left[(X - M\ (X))^2\right] \tag{4.4}$$

y la desviación estándar, se define como la raíz cuadrada de la varianza, o sea

$$s = \sqrt{V\ (X)} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}} \tag{4.5}$$

Las expresiones (4.4) y (4.5), se utilizan para calcular la varianza y desviación estándar de datos sin agrupar, es decir cuando se tiene el conjunto de observaciones originales. En el caso de tener una distribución de frecuencias, es decir datos agrupados, entonces la varianza y la desviación es-

* Algunos autores llaman también *desviación típica o tipificada.*

tándar se calculan según las fórmulas

$$s^2 = V(X) = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_i}{n} \tag{4.6}$$

y

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{n}}$$

Los dos ejemplos siguientes nos muestran el procedimiento general de cálculo.

<u>Ejemplo 4.4</u> Tomemos la información del ejemplo 4.2

TABLA 4.3 Cálculo de la varianza para datos sin agrupar

x_i	$d_i = x_i - \bar{x}$	$d_i^2 = (x_i - \bar{x})^2$
3	- 1,7	2,89
4	- 0,7	0,49
5	0,3	0,09
4	- 0,7	0,49
6	1,3	1,69
5	0,3	0,09
4	- 0,7	0,49
6	1,3	1,69
7	2,3	5,29
3	- 1,7	2,89
$\sum$	0	16,10

De manera que $V(X) = M\,(d^2) = \frac{16,1}{10} = 1,61$

y la desviación estándar será

$$s = \sqrt{1,61} = 1,2689$$

<u>Ejemplo 4.5</u> Sea el ejemplo 4.3

TABLA 4.4 Cálculo de la varianza para datos agrupados

x_i	n_i	$d = x_i - \overline{x}$	$d^2 = (x_i - \overline{x})^2$	$(x_i - \overline{x})^2\, n_i$
3	2	- 1,7	2,89	5,78
4	3	- 0,7	0,49	1,47
5	2	0,3	0,09	0,18
6	2	1,3	1,69	3,38
7	1	2,3	5,29	5,29
$\sum$	10			16,1

Luego, $V(X) = \dfrac{\sum (x_i - \overline{x})^2\, n_i}{n} = \dfrac{16,1}{10} = 1,61$

y $\qquad s = \sqrt{V(X)} = \sqrt{1,61} = 1,2689$

A continuación veremos las propiedades más importantes de la varianza, que obviamente implican las propiedades de la desviación estándar.

<u>PROPIEDADES</u>:

En lo que sigue, supondremos que se tienen n valores observados x_1, x_2, ..., x_n de una característica particular X de la población.

<u>TEOREMA 4.1</u> $\qquad V(X) \geqslant 0 \qquad$ (4.7)

<u>Demostración</u>: Por definición, $V(X) = M(d^2)$, y como $d_i^2 = (x_i - \overline{x})^2 \geqslant 0$ cualquiera que sea el valor de d_i, entonces $M(d^2) \geqslant 0$ $\blacksquare$

<u>TEOREMA 4.2</u> Si todos los valores de X son iguales a una constante K, entonces

$$V(K) = 0 \qquad (4.8)$$

<u>Demostración</u>: Por definición, $V(K) = M\left[(K - M(K))^2\right]$, pero como

$M(K) = K$ (por 3.9), entonces

$M(K - K)^2 = M(0) = 0$ $\blacksquare$

<u>TEOREMA 4.3</u> Si K es una constante cualquiera, entonces

$$V(X + K) = V(X) \qquad (4.9)$$

Demostración: $$V(X+K) = M\left\{\left[(X+K) - M(X+K)\right]^2\right\}$$
$$= M\left\{\left[X + K - M(X) - M(K)\right]^2\right\}$$
$$= M\left\{\left[X - M(X)\right]^2\right\} = V(X) \qquad \blacksquare$$

TEOREMA 4.4 Si K es una constante cualquiera, entonces

$$V(KX) = K^2\, V(X) \tag{4.10}$$

Demostración: Por definición

$$V(KX) = M\left[(KX - M(KX))^2\right] = M\left[(KX - KM(X))^2\right]$$
$$= M\left\{\left[K(X - M(X))\right]^2\right\} = M\left[K^2(X - M(X))^2\right]$$
$$= K^2 M\left[(X - M(X))^2\right] = K^2\, V(X) \qquad \blacksquare$$

TEOREMA 4.5 $$V(X) = M(X^2) - (M(X))^2 \tag{4.11}$$

Demostración: $$V(X) = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum (x_i^2 - 2\bar{x}x_i + (\bar{x})^2)}{n}$$
$$= \frac{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}{n}$$
$$= \frac{\sum x_i^2}{n} - 2\bar{x}\,\bar{x} + (\bar{x})^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2$$
$$= M(x^2) - (M(X))^2 \qquad \blacksquare$$

Esta última propiedad, que es una forma equivalente de escribir la varianza, es muy útil para trabajos de computación. La gran ventaja, radica en que no es necesario el cálculo de los desvíos. Entonces, para fines de cálculo recomendamos usar las siguientes fórmulas:

Para datos sin agrupar
$$V(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \tag{4.12}$$

y para datos agrupados

$$V(X) = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - \left(\frac{\sum_{i=1}^{m} x_i n_i}{n}\right)^2 \tag{4.13}$$

Ejemplo 4.6 La siguiente Tabla nos muestra ejemplos prácticos para estas últimas tres propiedades.

TABLA 4.5 Cálculos para la propiedad (4.11) aplicados a los teoremas 4.3 y 4.4

Datos Originales		Cada valor de X es incrementado en 5		Cada valor de X es multiplicado por 3	
Valor X	X^2	X + 5	$(X + 5)^2$	3X	$(3X)^2$
1	1	6	36	3	9
2	4	7	49	6	36
4	16	9	81	12	144
5	25	10	100	15	225
$\sum$ 12	46	32	266	36	414

$M(X) = \frac{12}{4} = 3$ $\quad$ $M(X+5) = \frac{32}{4} = 8$ $\quad$ $M(3X) = \frac{36}{4} = 9$

$M(X^2) = \frac{46}{4} = 11,5$ $\quad$ $M(X+5)^2 = \frac{266}{4} = 66,5$ $\quad$ $M(3X)^2 = \frac{414}{4} = 103,5$

$V(X) = 11,5 - 9 = 2,5$ $\quad$ $V(X+5) = 66,5 - 64 = 2,5$ $\quad$ $V(X) = 103,5 - 81 = 22,5$

$= 3^2\,(2,5)$

Nota: En el caso de tener datos agrupados, los cálculos son idénticos, teniendo únicamente cuidado no olvidar de multiplicar los valores por sus frecuencias (n_i).

Ejemplo 4.7 Calculemos las varianzas para los Ejemplos 4.4 y 4.5, aplicando la propiedad 4.11.

TABLA 4.6

X_i	X_i^2
3	9
4	16
5	25
4	16
6	36
5	25
4	16
6	36
7	49
3	9
$\sum$ 47	237

TABLA 4.7

X	n_i	x_i^2	$x_i^2\, n_i$
3	2	9	18
4	3	16	48
5	2	25	50
6	2	36	72
7	1	49	49
$\sum$	10		237

Entonces para la primera Tabla (datos sin agrupar), tenemos

$$V(X) = \frac{237}{10} - (4,7)^2 = 1,61$$

y para la segunda

$$V(X) = \frac{237}{10} - (4,7)^2 = 1,61$$

Ejemplo 4.8 Los salarios percibidos por un grupo de 25 empleados en dos empresas privadas se da en la siguiente Tabla.

TABLA 4.8 Comparación de salarios de los empleados de dos empresas privadas

SALARIOS (En cientos de $)	x_i	EMPRESA A No. de Empleados	EMPRESA A $x_i^2 n_i$	EMPRESA B No. de Empleados	EMPRESA B $x_i^2 n_i$
5 a 10	7,5	1	56,25	5	281,25
10 a 15	12,5	3	468,75	1	156,25
15 a 20	17,5	8	2450,00	3	918,75
20 a 25	22,5	5	2531,25	7	3543,75
25 a 30	27,5	6	4537,50	5	3781,25
30 a 35	32,5	2	2112,50	4	4225,00
$\sum$		25	12156,25	25	12906,25

$$\bar{x}_A = \frac{527,50}{25} = 21,1 \qquad \bar{x}_A = 2.110\ \$$$

$$\bar{x}_B = \frac{527,50}{25} = 21,1 \qquad \bar{x}_B = 2.110\ \$$$

$$S_A^2 = \frac{12.156,25}{25} - (21,1)^2 = 486,25 - 445,21 = 41,04$$

$$S_A = \sqrt{41,04} = 6,4063 \Longrightarrow 640,63\ \$$$

$$S_B^2 = \frac{12.906,25}{25} - 445,21 = 516,25 - 445,21 = 71,04$$

$$S_B = \sqrt{71,04} = 8,4285 \Longrightarrow 842,85\ \$$$

Como puede apreciarse, si bien el ingreso promedio de los empleados en ambas empresas son iguales, vemos que hay mayor dispersión en salarios que perciben en la empresa B.

Es importante notar que la desviación estándar, como una medida descriptiva, es esencialmente una media; y como se trata de una media de desviaciones, una desviación estandar pequeña indica un alto grado de uniformidad de las observaciones de la serie estadística; y si la desviación estándar es un valor grande, nos indica poca uniformidad en los datos observados. De esta manera, si se comparan dos o más series, que tienen aproximadamente la misma media, el más representativo será aquél que tenga menor desviación estándar. (Ver ejemplo 4.8).

4.4 Dispersión Relativa

Las medidas de variabilidad absolutas, como las que acabamos de ver, no siempre son posibles de utilizar, por ejemplo. para comparar dos conjuntos de valores, sobre todo si éstos tienen distintas unidades de medida. Por ésto, en muchos problemas, una medida de variabilidad relativa para distribuciones de frecuencias suele ser más significativa que la variabilidad absoluta.

Si dos conjuntos de valores se están comparando, los valores absolutos son convenientes solamente cuando los promedios de los dos conjuntos son aproximadamente del mismo tamaño y las unidades de medida son idénticos. Efectivamente, la comparación de dos diferentes unidades de medida, tales como número de libros comparados con el número de horas de viaje, no tiene sentido. Incluso cuando las unidades son las mismas, la comparación del grado de dispersión basados en los valores absolutos de los diferentes conjuntos suele ser aún difícil. Así por ejemplo, supongamos que se tiene la siguiente información: El ingreso medio mensual de cierto grupo de trabajadores adultos es $\bar{x}$ = 1.875 \$ con una desviación estandar s= 285 \$; en tanto que el ingreso medio mensual para un grupo del mismo tamaño de voceadores, es 315 \$ con una desviación estándar de 80 \$. No se puede concluir que la desviación estándar más alta (285 \$)de el más alto grado de dispersión (en los ingresos de los trabajadores adultos). La desviación estándar es significativa solamente en relación con la media respecto a la cual se calcula.

De esta manera es que se requiere de una medida de dispersión que exprese en términos relativos, para comparaciones como los casos indicados anteriormente. En general, una dispersión relativa es el cociente de una medida dada de dispersión dividida por el promedio con respecto al cual las desviaciones fueron medidas.

La medida de dispersión relativa más usada es el coeficiente de variación (CV)

DEFINICION 4.4 El *coeficiente de variación* está dada por

$$CV = \frac{s}{\bar{x}} \tag{4.14}$$

donde s es la desviación estándar y $\bar{x}$ la media aritmética, de un mismo conjunto de observaciones. O también $\frac{s}{\bar{x}} \cdot 100\%$ *(en términos de porcentaje).

Así, en nuestro ejemplo hipotético, para los trabajadores adultos, se tiene $CV = \frac{285}{1875} = 0{,}152$ o sea 15,2 %, mientras que el coeficiente de variación de los ingresos de los voceadores es $CV = \frac{80}{315} = 0{,}254$ o sea 25,4 %. De esta manera, es posible hacer comparación de la variabilidad de los ingresos promedio de ambas poblaciones. Como puede apreciarse, la variabilidad de los ingresos de los voceadores es mayor que la de los trabajadores adultos.

Otro coeficiente también usado con bastante frecuencia, es el *coeficiente de desviación media*, que está definida por:

$$CDM = \frac{DM}{\bar{x}} \tag{4.15}$$

Se interpreta en forma análoga al caso anterior.

Vale la pena mencionar algunos otros tipos de medidas, pero que no son de dispersión propiamente dicha. Así por ejemplo, tenemos el *coeficiente de asimetría* (CA_s)

$$CA_s = \frac{M(X) - M_o}{s} \tag{4.16}$$

que nos mide el grado de asimetría. En general se presentan los tres casos de la Figura 4.1.

* Fué Pearson quién usó por primera vez este coeficiente.

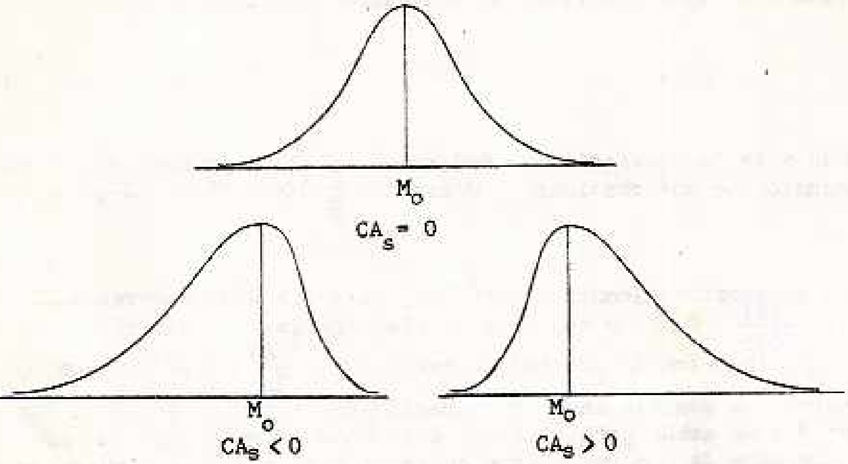


Figura 4.1 Ejemplos de asimetría: El primer caso muestra simetría perfecta. El segundo caso se dice que tiene asimetría negativa; y el tercer caso indica asimetría positiva.

Una medida de la mayor o menor concentración alrededor de la media, es el *coeficiente de apuntamiento*, que es

$$C_{Ap} = \frac{M\ (d^4)}{\left(V\ (X)\right)^2} \tag{4.17}$$

que cumple las siguientes relaciones:

Si $C_{Ap} = 3$ indica distribución normal

Si $C_{Ap} > 3$ indica apuntamiento hacia arriba

Si $C_{Ap} < 3$ indica apuntamiento hacia abajo

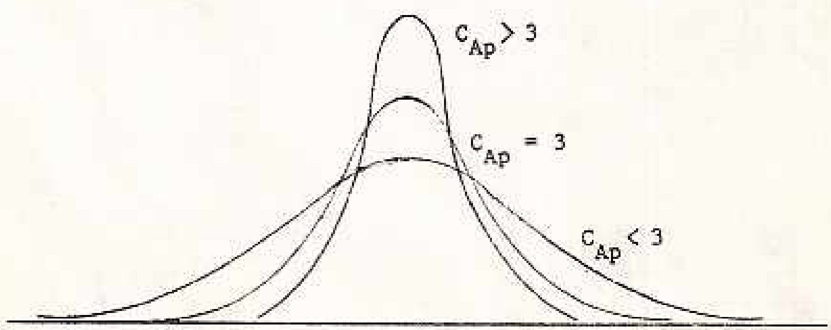


Figura 4.2 Ilustración aproximada del mayor o menor apuntamiento relativo

Ejercicios

1. Determinar el recorrido, la desviación media, la desviación estándar y el coeficiente de variación para los Ejercicios 8, 10, 11 y 12 propuestos en el Capítulo 2.

2. Si las calificaciones numéricas obtenidas en los exámenes finales de un curso, cuyo promedio aritmético es 5,3 con $S^2 = 0{,}2$ fuesen modificadas:

   a) por adición de un punto a cada una de ellas, ó
   b) por elevación general de un cinco por ciento.

   Qué efecto producirían estas modificaciones en la nota media y en la desviación estándar de dichas notas?

3. En el país A el ingreso per-cápita es de \$ 2.000 y S = 500 \$. En el país B se tiene el mismo ingreso por persona, pero la desviación estándar es \$ 1.000. En qué país es más uniforme la distribución del ingreso?

4. La varianza de dos números es 1 y su media aritmética 8. Calcule los números.

5. Las notas de 50 alumnos se clasifican en una tabla de frecuencias con cuatro intervalos de igual magnitud. Se pide calcular la varianza sabiendo que:

   $x_2 = 50$; $n_1 = 4$; $N_2 = 20$; $n_3 = 25$ y $\overline{x} = 62$

6. En un estudio se obtuvo el resultado del cuadro siguiente:

x_i	n_i		$(x_i - \overline{x})^2$	$\cdot n_i$
1	1	1	4	4
2	6	12	1	6
3	7	21	0	0
4		4x	1	x
5		5y	4	4y

   Si se sabe que la media es 3 y la varianza 0,95, se pide calcular las frecuencias absolutas que faltan.

7. El coeficiente de variación de los sueldos de 200 empleados de una empresa es 57%. Después de reajustar todos los sueldos en \$ 110,

el coeficiente vale 50%. Además, la gerencia fija un sueldo mínimo de \$ 710, lo que veneficia a 35 empleados, que antes del reajuste ganaban menos de \$ 600, con un sueldo medio de \$ 400. Se desea saber cuál será el pago mensual total de la empresa, después del reajuste.

8. Determine si las siguientes relaciones son posibles:

   8.1 $\sum x_i^2 n_i = 3.000$; $n = 50$; $\bar{x} = 62$

   8.2 $DM = 47$; $S = 38$; $n = 17$

   8.3 $M\left(X^2\right) = 300$ $\quad S = 74$

   8.4 $\sum |x_i - M_e| = 3.000$; $n = 15$; $DM = 220$; $S = 210$

9. Determine si las siguientes afirmaciones son correctas:

   9.1 Si todos los valores de la variable de una distribución, cuya desviación estándar es 30, se aumentan en 30%, la nueva varianza es 1521.

   9.2 Si los ingresos de la población activa en Bolivia se expresan en dólares, la varianza antes de la devaluación es mayor que después de ella.

   9.3 Si cierta distribución tiene una varianza igual a 144 y otra una desviación estándar igual a 11,5 puede afirmarse que la primera tiene mayor dispersión.

   9.4 Cuando la variable se multiplica por una constante, todas las siguientes medidas cambian de valor: media aritmética, mediana, coeficiente de variación y moda.

   9.5 La varianza es siempre mayor que la desviación estándar.

10. Sea X una variable con media $\bar{x}$ y desviación estándar S_x. Si definimos la nueva variable Y, tal que

$$Y = \frac{X - \bar{X}}{S_x}$$

Demostrar que la media y la varianza de Y son iguales a cero y a uno, respectivamente.

11. Determinar el coeficiente de asimetría y el coeficiente de apuntamiento para los ejercicios 8 y 12 propuestos en el Capítulo 2.

C A P I T U L O 5

TEORIA DE PROBABILIDADES

Una de las herramientas fundamentales para el estudio de la estadística moderna, es la *teoría de probabilidades*. a diferencia de la estadística descriptiva que está relacionada con el resumen de datos y la descripción de éstos mediante algunas medidas, la *estadística inferencial*, es una generalización hecha sobre la base de datos muestrales respecto del comportamiento de alguna característica particular de la población. En vista de que los datos obtenidos de una muestra nos dan una información incompleta de todo el conjunto de elementos de una población, tomar decisiones de generalización para la población, implica el riesgo de hacer una inducción falsa; la medida para esta incertidumbre es la probabilidad. Generalmente la incertidumbre se debe al carácter *aleatorio** de que no se sabe con certeza cuál será el resultado que se obtendrá de una muestra, y además si el resultado coincide o nó con el que se obtendría tomando en cuenta a todos los elementos de la población.

Más adelante, en los capítulos respectivos veremos con detalle la aplicación de las probabilidades a la estadística inferencial, concretamente en lo relacionado a la estimación estadística y la docimasia de hipótesis.

Para el estudio de la teoría de probabilidades y de la teoría estadística, un instrumento matemático muy importante, es la Teoría de Conjuntos. Entonces, antes de entrar al estudio de la probabilidad, esbozaremos en forma sintética los conceptos y propiedades de los conjuntos, que nos permitirán comprender con mayor facilidad el tratamiento de las probabilidades.

---

* El término aleatorio, que se usará en el contexto del presente trabajo, debe tomarse como sinónimo de azar, es decir que del conjunto de resultados posibles de un experimento, ocurre uno de ellos, pero no se sabe cuál. Así por ejemplo, decir que se *seleccionan al azar* dos personas de diez (todas seleccionables), para que formen parte de un tribunal calificador, significa que no se sabe a priori cuáles serán las dos seleccionadas; esto mismo se quiere decir con la palabra aleatorio en lugar de azar.

5.1

Noción de Conjunto

En esta sección veremos de la teoría de conjuntos solamente la parte necesaria para comprender el concepto de las probabilidades modernas, o sea, desde un punto de vista abstracto, que consiste en partir de un conjunto de axiomas o postulados, de los que se deducen sus propiedades y aplicaciones. Este *método axiomático* será el que utilizaremos, sin descuidar desde luego los otros métodos, como ser: el método de la *frecuencia relativa de un suceso* y el *método combinatorio*.

La noción de conjunto es, tal vez, una de las más difíciles de explicar y comprender de todas las matemáticas. Sin embargo, se puede decir también que es lo más sencillo del mundo, ya que todos sabemos intuitivamente, que se entiende por la palabra *conjunto*. Así, cuando decimos que un conjunto de palabras forma una oración, que un conjunto de cabellos de un hombre constituye su cabellera o que una máquina está formada por un conjunto de piezas , etc., son expresiones que no ofrecen dificultad alguna de comprensión para nadie. Es precisamente esta noción intuitiva que aprovechó la matemática moderna para establecer una de las teorías más generales y fecundas de gran aplicación en las probabilidades y en la estadística, que es la *teoría de los conjuntos*.

Una definición simple de conjunto será entonces:

DEFINICION 5.1 Conjunto es la reunión (o colección) de objetos bien determinados y capaces de diferenciarse unos de otros.

Un objeto que pertenece a un conjunto determinado recibe el nombre de *elemento* (o componente) del conjunto . Es evidente que los elementos de un conjunto no han de ser necesariamente objetos concretos, perceptibles por alguno de nuestros cinco sentidos; así, se puede hablar de conjuntos abstractos como el conjunto de los puntos de una recta o el conjunto de los dioses de la mitología griega. Estos son conjuntos cuyos objetos se deben a nuestra intuición o nuestro pensamiento. De esta manera, los objetos que pertenecen a un conjunto determinado, pueden ser reales o ideales; por ejemplo: letras, personas, plantas, fábricas, números, ideas, etc.

En general, usaremos letras mayúsculas (A, B, C, X, Y,...) para designar conjuntos, y letras minúsculas (a, b, c, x, y, ...) para indicar a los elementos del conjunto. Para especificar que ciertos objetos pertenecen a un conjunto dado, se usan corchetes $\{\ \}$. Existen dos formas clásicas de representar un conjunto, según sea posible describir sus elementos; estos son: por *extensión* y por *comprensión*. En el primer caso se trata de un listado de todos los elementos componentes de un conjunto; mientras que en el segundo caso, se trata de una especie de regla que nos indica cuáles son los componentes, sin necesidad de describirlos a todos. Por ésto, todo conjunto debe estar bien definido en relación a sus elementos. Así por ejemplo, si de -

seamos escribir el conjunto A cuyos elementos son a, b, c y d, y el conjunto B cuyos elementos son los primeros cinco números enteros positivos, tendremos:

$$A = \{a, b, c, d\}$$

$$B = \{1, 2, 3, 4, 5\}$$

que es la forma extensiva. Si tuviésemos un conjunto C, cuyos elementos son los primeros 100 números naturales, es más conveniente expresarlo en forma comprensiva, o sea:

$$C = \{x : x = 1, 2, .., 100\}$$

que se lee, "C es un conjunto de elementos x, tal que x = 1, ó x = 2, ó x = 3, y así sucesivamente, ó x = 100".

No todo conjunto puede ser descrito en forma extensiva, así por ejemplo, un conjunto D cuyos elementos son todos los números reales comprendidos entre 0 y 1, sólo es posible expresarlo en forma comprensiva, o sea

$$D = \{x : 0 < x < 1\}$$

que se interpreta: "D es un conjunto de elementos x, tal que x es mayor que cero y menor que uno". El estudiante, con un poco de práctica, podrá fácilmente decidir cómo expresar y escribir un conjunto.

Usaremos los símbolos $\in$ y $\notin$ para indicar, "pertenece a" y "no pertenece a", respectivamente. Entonces, en relación a los conjuntos A, B, C y D definidos anteriormente, podemos escribir:

$$a \in A, \quad 5 \in B, \quad 57 \in C, \quad 1/4 \in D,$$

$$57 \notin A, \; 1/4 \notin C, \; -2 \notin D, \text{ etc.}$$

Cuando un conjunto contiene un número finito de elementos se dice que es *finito*; en caso contrario, se dice que es *infinito*. Así por ejemplo, el número de trabajadores de una fábrica es finito, el número de piezas de una máquina es finito, el conjunto de números naturales es infinito, el número de los puntos de un segmento de línea es infinito, etc.

Se dice que un conjunto es *enumerable* * cuando se pueden "enumerar" sus elementos; así, el conjunto E de los números positivos múltiplos de 2:

$$E = \{0, 2, 4, 6, 8, \ldots\}$$

---

* Otras veces se dice también, *numerable*.

es numerable. Un conjunto es *no enumerable,* en caso contrario.

Antes de seguir, hagamos referencia al *signo de implicación* ($\Longrightarrow$). Este símbolo se emplea para expresar la consecuencia de un hecho, es decir que se trata de una "inferencia lógica", que puede leerse "implica", "lleva a", "tiene por consecuencia", "si ... también", o "si... entonces". Por ejemplo, si $(x = y) \Rightarrow (3x = 3y)$; Si $A = \{a, b, c\} \Rightarrow b \in A$, etc. El símbolo ($\Longleftrightarrow$) indica una equivalencia lógica (se trata de una doble implicación) o también una condición necesaria y suficiente, que puede interpretarse como "si y solamente si" (o abreviadamente "sii").

DEFINICION 5.2 Dos conjuntos A y B son *iguales* si cada elemento que pertenece a A, también pertenece a B, y cada elemento que pertenece a B también pertenece a A.

Es decir, que dos conjuntos A y B son idénticos o iguales si están compuestos de los mismos elementos. En símbolos, $A = B \Longleftrightarrow x \in A \Rightarrow x \in B$ y si $x \in B \Rightarrow x \in A$. Esto, sirve para demostrar que dos conjuntos son iguales. Cuando dos conjuntos son desiguales se representa:

$$A \neq B.$$

Notamos que el orden del listado de elementos de un conjunto no tiene importancia, así

$$A = \{1, 2, a, 3\} \quad y \quad B = \{a, 3, 1, 2\}$$

son iguales. También el número de veces que un mismo elemento es listado en un conjunto, no tiene importancia, así,

$$E = \{1, 2, 2, 1, 3, 1\} \qquad y \quad F = \{1, 2, 3\}$$

son iguales. Como puede apreciarse, el conjunto E comete redundancia en cuanto a la descripción de sus elementos. De manera que si consideramos un conjunto con <u>n</u> elementos, estos en general deben ser distintos y los elementos iguales no deben aparecer más de una vez. Esto es también importante para determinar el *número* de elementos que tiene un conjunto; por ejemplo en el caso anterior, erróneamente podría decirse que el conjunto E tiene 6 elementos, siendo lo correcto decir que tiene 3 elementos.

DEFINICION 5.3 Sean los conjuntos A y B. A es un *subconjunto* de B (representado por $A \subset B$) si y sólo si cada elemento que pertenece a A pertenece también a B.

En otras palabras, A es un subconjunto de B si B contiene a A (representado por $B \supset A$). Para demostrar que un conjurto A está contenido en un conjunto B, basta comprobar la implicación siguiente para un elemento <u>a</u> cualquiera:

$$a \in A \Rightarrow a \in B$$

El conjunto $\{a\}$ es un subconjunto de $\{a, b\}$, y el conjunto $\{a, b\}$ es un subconjunto de $\{a, b, c\}$. El conjunto de todas las personas residentes en La Paz es un subconjunto de todos los residentes de Bolivia; el conjunto de todos los trabajadores mineros es un subconjunto de todos los trabajadores.

Cuando un conjunto A no está contenido en el conjunto B, se denota por $A \not\subset B$.

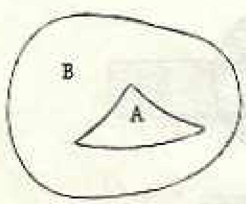
Tomando en cuenta el concepto de subconjunto, la igualdad de dos conjuntos A y B, también se define: "Dos conjuntos son iguales si cada uno de ellos está contenido en el otro, es decir, $A = B \Leftrightarrow A \subset B$ y $B \subset A$".

Debe tomarse en cuenta la diferencia importante que existe entre los conceptos, <u>pertenencia</u> e <u>inclusión</u>. Si por ejemplo tenemos los conjuntos

$$A = \{1, 2, 3, 4\}, \quad B = \{1, 2, 4\} \text{ y } C = \{1\},$$

entonces es correcto decir $B \subset A$ y $C \subset A$, ya que cada elemento de B pertenece a A, y cada elemento de C pertenece a A; mientras que es incorrecto decir que $A \subset C$ o que $B \subset C$, ya que A y B no están contenidos en C. También es correcto decir que $3 \in A$ y $1 \in C$, y es incorrecto expresar que $3 \subset C$ y $1 \subset C$, ya que 3 y 1 no son subconjuntos. Así, podemos decir que el conjunto de todos los estudiantes de economía es un subconjunto de todos los estudiantes, pero no podemos decir que el conjunto de todos los estudiantes de economía pertenecen al conjunto de todos los estudiantes. Un estudiante particular de economía pertenece al conjunto de todos los estudiantes de economía, pero dicho estudiante no es un subconjunto de ellos.

A veces se suelen utilizar los *diagramas de Euler* (que en algunas partes se llama diagramas de Venn), para visualizar gráficamente algunas características de los conjuntos. Así por ejemplo, si un conjunto A está contenido en el conjunto B, se tiene



$A \subset B$

Figura 5.1

De las definiciones anteriores, es fácil ver que cada conjunto es igual a sí mismo, y que cada conjunto es un subconjunto de sí mismo, es decir que dado un conjunto A, se tiene:

$$A = A \quad \forall A$$

$$A \subset A \quad \forall A$$

El símbolo ($\forall$), indica "para todo" ó "cualquiera que sea"

Un conjunto que contiene todos los elementos de un cierto tipo, que sirven de base para una discusión particular, recibe el nombre de *conjunto universal* (llamado también *conjunto fundamental*). A este conjunto lo designaremos con la letra omega mayúscula Ω, y a sus elementos con omega minúscula ω, de manera que $\omega \in \Omega$. Así por ejemplo, si se está discutiendo la producción de zapatos, el conjunto universal estará compuesto de todos los zapatos producidos por las distintas zapaterías de cierta localidad.

Por otro lado, un conjunto que no tiene elemento alguno, se llama *conjunto vacío* (o también *conjunto nulo*), y se lo representa con la notación ϕ, o sea $\phi = \{\ \}$

A continuación estudiaremos un poco del *álgebra de conjuntos* (llamadas también operaciones booleanas). Al igual que en el álgebra de los números reales que está relacionada con operaciones realizadas con números y sus consecuencias, el álgebra de conjuntos está relacionada con conjuntos y sus implicaciones [4]

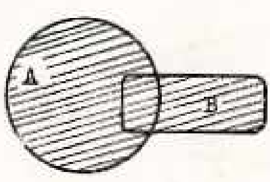
DEFINICION 5.4 Se llama *unión* (o también, reunión) de dos conjuntos A y B al conjunto formado por los elementos que pertenecen al menos a uno de los dos conjuntos A y B.

La notación de unión es $A \cup B$, donde el símbolo $\cup$ significa "unión". Para fines de demostraciones, la unión de dos conjuntos A y B se define:

$$A \cup B = \{x : x \in A \vee x \in B\}$$

donde $\vee$ indica "o bien" o "y/o".

La representación gráfica para la unión de los conjuntos A y B, será



$A \cup B$

Figura 5.2

En relación a la unión, se tienen las siguientes propiedades lógicas:

$$A \cup \phi = A$$
$$A \cup A = A$$
$$A \cup E = E, \text{ si } A \subset E$$

Veamos el siguiente ejemplo que nos aclarará la operación de unión:

Si $A = \{1, 2, 3\}$, $B = \{1, 3\}$ y $C = \{0\}$, entonces $A \cup B = \{1, 2, 3\}$, $A \cup C = \{0, 1, 2, 3\}$ y $B \cup C = \{0, 1, 3\}$. Además $A \cup B \cup C = \{0,1,2,3\}$.

Una propiedad inmediata, que se ve de la definición, es la *conmutatividad* de la operación de unión, es decir que $A \cup B = B \cup A$. Dado que el orden del listado de los elementos no tienen importancia. Además, la unión de conjuntos es *asociativa*, puesto que

$$(A \cup B) \cup C = A \cup (B \cup C)$$

Una segunda operación, es la llamada intersección de dos conjuntos.

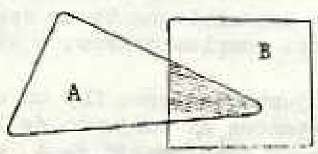
DEFINICION 5.5 Se denomina *intersección* de dos conjuntos A y B al conjunto formado por todos los elementos comunes de A y B.

La notación de intersección es $A \cap B$, donde el símbolo $\cap$ significa "intersección". Para fines de demostraciones, se define equivalentemente:

$$A \cap B = \{x : x \in A \wedge x \in B\}$$

donde el símbolo $\wedge$, indica "y".

Para este caso, el diagrama euleriano, será



$A \cap B$

Figura 5.3

En este caso, se tienen las siguientes propiedades lógicas:

$$A \cap \phi = \phi$$
$$A \cap A = A$$
$$A \cap E = A, \text{ si } A \subset E$$

Si por ejemplo, $A = \{1, 2, 3\}$, $B = \{0, 2\}$ y $C = \{1, 3\}$, entonces $A \cap B = \{2\}$, $A \cap C = \{1, 3\}$, $B \cap C = \phi$. La operación intersección al igual que la unión, es también conmutativa y asociativa, ésto es

$$A \cap B = B \cap A \quad \text{y} \quad A \cap (B \cap C) = (A \cap B) \cap C$$

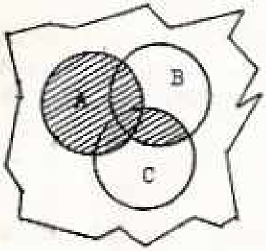
por otro lado, vemos que

$$A \cap B \subset B \quad \text{y que} \quad A \cap B \subset A$$

Una propiedad relacionada con las operaciones de unión e intersección, es la *distributiva*, ésto es

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Esto mismo en términos del diagrama de Euler, se tiene:



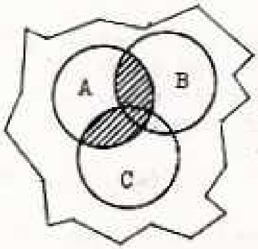


Figura 5.4

donde las partes achuradas corresponden a las operaciones indicadas.

Si la intersección de dos conjuntos es vacía, es decir, si

$$A \cap B = \emptyset$$

entonces A y B son *conjuntos disjuntos*, (ó incompatibles, o excluyentes entre sí). Por ejemplo, si $A = \{1, 2\}$ y $B = \{3, 4\}$, se tiene $A \cap B = \emptyset$, luego A y B son disjuntos.

Si consideramos un conjunto universo particular Ω, entonces un conjunto A incluído en Ω, especifica automáticamente un segundo conjunto distinto de A, al que se lo llama conjunto complementario, o simplemente complemento.

DEFINICION 5.6 Dado un conjunto universo Ω y un conjunto A tal que $A \subset \Omega$, entonces el conjunto de los elementos de Ω que no pertenecen a A se denomina *complemento* de A, y se escribe A^c. Es decir,

$$A^c = \{x : x \notin A \text{ para todo } A \subset \Omega\}$$

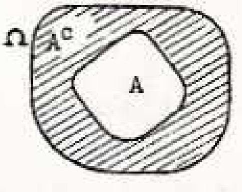


Figura 5.5

Por ejemplo; si $\Omega = \{a, b, c, d, e, f, g\}$, y $A = \{b, d, g\}$, entonces $A^c = \{a, c, e, f\}$. Si el conjunto universal Ω comprende todos los números naturales y A el conjunto de los números pares, entonces A^c es el conjunto de todos los números impares.

Un subconjunto y su complementario tienen las siguientes propiedades lógicas:

Si $A = \Omega$ entonces $A^c = \emptyset$
Si $B = A^c$, entonces $A = B^c = (A^c)^c$

$$A \cup A^c = \Omega$$
$$A \cap A^c = \emptyset$$

También puede observarse en relación al conjunto vació, que

$$A \cup \emptyset = A$$
$$A \cap \emptyset = \emptyset$$

De la definición de subconjunto, podemos admitir que $\emptyset \subset A$ para todo A, ya que cada elemento perteneciente a $\emptyset$ (ninguno) también pertenece a A. Desafortunadamente, el conjunto vacío $\emptyset$ no se puede ilustrar en un diagrama.

DEFINICION 5.7 . Se denomina *diferencia* *entre dos conjuntos A y B, al conjunto formado por los elementos de A y no de B, y se representa por $A \setminus B$.

Simbólicamente se tiene:

$$A \setminus B = \{x : x \in A,\ x \notin B\}$$

y graficamente

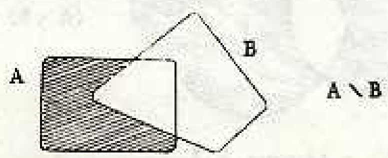


Figura 5.6

De acuerdo con la definición se cumple:

$$A \setminus B = A \setminus A \cap B \quad \text{(Fig. 5.6)}$$

También

$$A \setminus B = A \cap B^c$$

Veamos algunos ejemplos: Si $A = \{a,b,d, e, f, h\}$ y $B = \{b,c,d,f,r\}$ entonces $A \setminus B = \{a,e,h\}$ y $B \setminus A = \{c, r\}$

Se observa también que

$$A \setminus B \neq B \setminus A \quad \text{(excepto si } A = B)$$

---

* Algunos autores llaman, complemento relativo.

En efecto,

$$A \setminus B = A \setminus A \cap B$$
$$B \setminus A = B \setminus A \cap B$$

y

$$(A \setminus B) \cap (B \setminus A) = \emptyset \qquad \text{(figura 5.7)}$$

de donde se desprende:

$$\text{Si } A = B \Longrightarrow A \setminus B = B \setminus A = \emptyset$$

$$\text{Si } A \setminus B = \emptyset \text{ y } B \setminus A \neq \emptyset \Longrightarrow A \subset B$$

$$\text{Si } A \subseteq B \Longrightarrow A \setminus B = \emptyset$$

$$\text{Si } A \setminus B = A \Longrightarrow B = \emptyset \text{ y/o } A \cap B = \emptyset$$

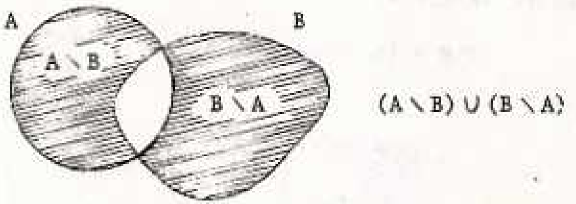


Figura 5.7

DEFINICION 5.8

1) Cuando los conjuntos $A_1, A_2, .., A_n$ son tales que su unión $A_1 \cup A_2 \cup \ldots$ $\ldots \cup A_n = \Omega$, se dice que forman un *recubrimiento* de Ω.

2) Cuando además se añade la condición de ser disjuntos dos a dos, es decir $A_i \cap A_j = \emptyset$ para todo $i \neq j$, donde i y j varían de 1 a n, el recubrimiento se llama *partición* de Ω.

Un recubrimiento y una partición de Ω, suponiendo que sólo hay cuatro conjuntos A_i, será

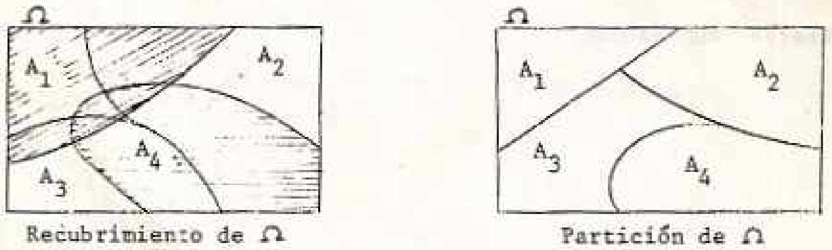


Figura 5.8

Por ejemplo si $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, entonces, los conjuntos $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6, 7\}$, $C = \{4, 5, 6, 7, 8\}$ y $D = \{9\}$, forman un recubrimiento, mientras que los conjuntos $A = \{1, 2, 7\}$ $B = \{3, 4, 5\}$ $C = \{8, 9\}$ y $D = \{6\}$ forman una partición. Las divisiones del territorio boliviano en departamentos es una partición del suelo de Bolivia.

Cuando tengamos una unión sucesiva de n conjuntos, simbolizaremos

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots\ldots \cup A_n$$

y para la intersección sucesiva

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \ldots\ldots \cap A_n$$

Para finalizar esta primera parte de la teoría de conjuntos, haremos un resumen de las operaciones que forman parte del álgebra booleana.

5.2

Operaciones Booleanas

Consideremos un conjunto referencial Ω y todas sus partes. Las operaciones *intersección, unión y complemento* efectuadas con las partes de Ω , se denominan *operaciones booleanas*.

Las principales propiedades algebraicas de las operaciones booleanas se examinarán como sigue, considerando los conjuntos A, B y C como partes de Ω.

$$\left.\begin{array}{l} A \cap B = B \cap A \\ A \cup B = B \cup A \end{array}\right\} \text{conmutativa}$$

$$\left.\begin{array}{l} (A \cap B) \cap C = A \cap (B \cap C) \\ (A \cup B) \cup C = A \cup (B \cup C) \end{array}\right\} \text{asociativa}$$

$$\left.\begin{array}{l} A \cap A = A \\ A \cup A = A \end{array}\right\} \text{idempotencia}$$

$$\left.\begin{array}{l} A\cap(B\cup C) = (A\cap B)\cup(A\cap C) \\ \\ A\cup(B\cap C) = (A\cup B)\cap(A\cup C) \end{array}\right\}$$ Distributiva (de la intersección con relación a la unión y de la unión con relación a la intersección).

$$A\cap A^c = \phi$$
$$A\cup A^c = \Omega$$
$$A\cap\phi = \phi$$
$$A\cup\phi = A$$
$$A\cap\Omega = A$$
$$A\cup\Omega = \Omega$$
$$(A^c)^c = A \qquad \text{involución}$$

$$\left.\begin{array}{l} (A\cap B)^c = A^c\cup B^c \\ \\ (A\cup B)^c = A^c\cap B^c \end{array}\right\} \qquad \text{teoremas de Morgan}$$

Las leyes de Morgan en forma general se expresan

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c$$

$$\left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$$

Las demostraciones de todas las identidades anteriores, se dejan como ejercicio.

DEFINICION 5.9. Una *clase (o familia)*, es un conjunto, tal que sus elementos son conjuntos.

Así, los conjuntos

$$\mathcal{H} = \{\{1\}, \{2\}\}$$
$$\mathcal{K} = \{\phi, \{1\}, \{2\}, \{1,2\}\}$$
$$\mathcal{A} = \{\{a\}, \{b\}, \{a,b\}\}$$

son clases. Un conjunto $\mathcal{B}$ que tiene como elementos todos los subconjuntos de $B = \{1,2,3\}$, o sea

$$\mathcal{B} = \{\phi, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, B\}$$

es una clase

DEFINICION 5.10 Una clase $\mathcal{F}$ es *cerrada* con respecto a las uniones e intersecciones si y sólo sí $A\cup B\in\mathcal{F}$ y $A\cap B\in\mathcal{F}$ para todo conjunto A y B que pertenece a la clase $\mathcal{F}$.

Por ejemplo, en relación a los ejemplos de la definición anterior, la clase $\mathcal{H}$ no es cerrada porque $\{1\}\cup\{2\} = \{1,2\}\notin\mathcal{H}$. La clase $\mathcal{K}$ es cerrada, puesto que la union o intersección de dos cualesquiera de sus elementos, también pertenecen a $\mathcal{B}$. La clase $\mathcal{A}$ es cerrada con respecto a la unión, pero no es cerrada con respecto a la intersección, ya que $\{a\} \cap \{b\} = \phi \notin \mathcal{A}$, de manera que no se puede decir que la clase $\mathcal{A}$ es cerrada. En cuanto al conjunto $\mathcal{B}$, se vé claramente que es una clase cerrada.

Algunos autores llaman *conjunto potencia*, a una clase cerrada. Es muy importante tener claro el concepto de clase (o familia) cerrada, en razón de que será de gran utilidad para el cálculo de "probabilidades" y para el concepto de "variable aleatoria" que es muy importante en el estudio de la teoría estadística.

Supongamos se tiene un conjunto definido como $A = \{a, b\}$, entonces la clase cerrada $\mathcal{A}$ de A, que también se simboliza $\mathcal{P}(A)$, será

$$\mathcal{A} = \mathcal{P}(A) = \{\phi, \{a\}, \{b\}, \{a,b\}\}$$

El *número* de elementos de una familia cerrada $\mathcal{A}$ de un conjunto A, es simplemente

$$n(\mathcal{A}) = 2^{n(A)}$$

donde $n(\mathcal{A})$ y $n(A)$ indican el número de elementos que tiene $\mathcal{A}$ y A, respectivamente.

En relación al número de elementos de un conjunto, existen *conjuntos finitos y conjuntos infinitos*. Un conjunto es finito se es vacío o si consiste exactamente de n elementos, donde n es un número entero positivo; y un conjunto es infinito en caso contrario. Así por ejemplo, el conjunto de todos los estudiantes de una universidad, es finito; mientras que el conjunto de puntos de un segmento de recta, es infinito.

Por otro lado, se dice que un conjunto es *enumerable* cuando se puede "enumerar" sus elementos. Todo conjunto finito es enumerable, mientras que no todo conjunto infinito se puede numerar, así por ejemplo, el conjunto de los números reales no es enumerable.

5.3

Relaciones entre Conjuntos

Ahora veremos algunos conceptos básicos respecto de la relación entre conjuntos. En primer lugar definamos n-tuplos ordenados.

DEFINICION 5.11 Un n-tuplo es un vector fila ordenado de n elementos, que se escribe

$$(x_1, x_2, \ldots, x_n)$$

Así por ejemplo, (1,2), (a,b), (0,40), (x_1, x_2) son 2-tuplos, que corrientemente se llama *par ordenado*; los vectores (1, 2, 3), (a, b, c), (1, 1, 1), (8, 1, 0), son 3-tuplos. Dos n-tuplos son diferentes inclusive si tienen los mismos elementos pero están en distinto orden. Por otro lado, un n-tuplo y un (n+1)-tuplo son también diferentes, ya que tienen diferente número de elementos. De esta manera, diremos que dos n-tuplos son iguales, o sea

$$(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n)$$

si y sólo si $x_1 = y_1,\ x_2 = y_2, \ldots, y_n = x_n$. Entonces,

$$(1,2) \neq (1,2,3)$$

$$(4,5) \neq (5,4)$$

$$(1, 1, 3) \neq (3, 1, 1)$$

$$(3, 4, 5) = (3, 4, 5).$$

Es frecuente el uso de pares ordenados; así por ejemplo, si deseamos observar el conjunto de matrimonios que viven en la ciudad de La Paz, cada pareja es un elemento del conjunto, entonces nosotros estamos en presencia de pares ordenados, en la que el primer elemento puede ser el nombre del esposo y el segundo elemento el nombre de la esposa. Notemos que un esposo individual que vive en La Paz no pertenece al conjunto, es un componente del par ordenado que pertenece al conjunto. Un ejemplo clásico, es el conjunto de puntos situados en un cuadrante del plano Cartesiano, ya que cada punto está representado por el par (x, y), donde x es la coordenada horizontal e y la coordenada vertical.

Una operación de interés entre conjuntos, es el producto Cartesiano, definido como sigue.

DEFINICION 5.12 Si A y B son dos conjuntos, se llama *producto cartesiano* de A y B, que se indica A x B, al conjunto de todos los posibles pares ordenados (a, b) cuyo primer elemento a pertenece a A y el segundo elemento b pertenece a B.

Simbólicamente esta definición puede escribirse

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Veamos algunos ejemplos, bajo el supuesto de que se tienen los conjuntos $A = \{1, 2, 3\}$, $B = \{4, 5\}$ y $C = \{0\}$, entonces:

$$A \times B = \{(1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5)\}$$

$$A \times C = \{(1, 0), (2, 0), (3,0)\}$$

$$C \times A = \{(0, 1), (0, 2), (0, 3)\}$$

$$B \times C = \{(4, 0), (5, 0)\}$$

$$B \times B = \{(4, 4), (4, 5), (5, 4), (5, 5)\}$$

$$C \times C = \{(0, 0)\}$$

Si $\mathbb{R}^+$ es el conjunto de número positivo, entonces $\mathbb{R}^+ \times \mathbb{R}^+$ es el conjunto de puntos en el primer cuadrante del plano cartesiano.

En general A x B y B x A son distintos: Entonces, el producto cartesiano no es conmutativo.

Es fácil generalizar el producto cartesiano para cualquier número de conjuntos. Así A x B x C es un conjunto de 3-tuplos, y $A_1 \times A_2 \ldots \times A_n$ es un conjunto de n-tuplos. Por ejemplo, tomando los conjuntos, definidos anteriormente, tendremos

$$A \times B \times C = \{(1,4,0), (1,5,0,) , (2,4,0), (2,5,0), (3,4,0), (3,5,0)\}$$

$$B \times C \times A = \{(4,0,1,), (4,0,2), (4,0,3),(5,0,1), (5,0,2), (5,0,3)\}$$

DEFINICION 5.13 Una *función elemento* f, (real-valorada), definida sobre un conjunto Ω es una regla que asocia un número (real) a cada elemento del conjunto. El número asociado a un elemento particular se llama el valor de la función para ese elemento y es denotado por f (w) para todo $w \in \Omega$

Pueden definirse muchas funciones para los elementos de un mismo conjunto (ésto es, pueden usarse distintas reglas). Nótese que puede asociarse un mismo número a más de un elemento.

Por ejemplo, si consideramos el conjunto de personas que residen en un cierto lugar, se pueden usar distintas reglas para asociar un número real a cada elemento del conjunto; tales como la edad de las personas, el peso de las personas, la estatura; los años de estudio, los ingresos, etc. Cada una de estas reglas, edad, peso, estatura, etc., se llama función elemento definida sobre el conjunto de esas personas.

Supongamos ahora que tenemos un conjunto de pares ordenados

$$\Omega = \{(x_1, x_2) : x_1 = 1, 2, 3; \ x_2 = 2, 3, 4\}$$

La regla $f(x_1, x_2) = x_1 + x_2$ para $(x_1, x_2) \in \Omega$, asocia la suma de los dos elementos de cada par del conjunto Ω. Entonces,

$f\ (1,2) = 1 + 2 = 3;\ f(2, 3) = 2 + 3 = 5;\ f(3, 2) = 5$, etc.

Una segunda función elemento definida sobre el mismo conjunto, puede ser

$$g\ (x_1, x_2) = x_1 \cdot x_2 \text{ para } (x_1, x_2) \in \Omega$$

Entonces

$$g\ (1, 2) = 1.2 = 2;\ g\ (3,3) = 3.3 = 9, \text{ etc.}$$

Dos conjuntos que son de interés para cualquier función, son: su dominio de definición y su recorrido.

DEFINICION 5.14 El *dominio* para una función elemento definida sobre un conjunto Ω es simplemente el mismo conjunto Ω; el *recorrido* de una función elemento definida sobre un conjunto Ω es el conjunto de número reales asociados con los elementos de Ω.

Respecto del recorrido, podemos decir que es el conjunto de valores de la función.

De esta manera, en relación al primer ejemplo dado anteriormente, el dominio de cada una de las funciones dadas, es el conjunto de personas que viven en dicho lugar. El recorrido de la función edad, es el conjunto de edades de los residentes en ese lugar; el recorrido de la función peso, es el conjunto de los pesos; el recorrido de la función años de estudio, es el conjunto del número de años estudiados; el recorrido de la función ingreso, es el conjunto de las cantidades monetarias obtenidas; etc. En el segundo ejemplo, el dominio de definición para ambas funciones f y g, es el conjunto Ω; el recorrido de f es el conjunto $\{3, 4, 5, 6, 7\}$ y el recorrido de g es el conjunto $\{2, 3, 4, 6, 8, 9, 12\}$.

Otra función importante, es la función conjunto, que está relacionada con la clase cerrada $\mathcal{A}$.

DEFINICION 5.15 Una regla f que asocia un número real a cada $A \in \mathcal{A}$ (denotado por f(A), se llama una *función conjunto* (real valorada) sobre $\mathcal{A}$.

Sea la clase $\mathcal{A}$ de subconjuntos de $A = \{1, 2, 3\}$, entonces

$$\mathcal{A} = \{\phi, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

Como en el caso de las funciones elemento, hay muchas diferentes funciones conjunto, que pueden definirse sobre $\mathcal{A}$. Por ejemplo,

$$\begin{aligned} f(A) &= 0 \quad \text{si } 1 \in A \\ &= 1 \quad \text{si } 1 \notin A \end{aligned}$$

es una regla que puede ser usada para asociar un número con cada conjunto $A \in \mathcal{A}$. Entonces f se llama una función conjunto definida sobre $\mathcal{A}$. En forma similar,

$$g(A) = \text{número de lementos de } A, \quad A \in \mathcal{A}$$

$$h(A) = \text{cuadrado del número de elementos pertenecientes a } A, \ A \in \mathcal{A}$$

son funciones conjunto definidas sobre $\mathcal{A}$. En este caso, el dominio de definición en cada caso es entonces $\mathcal{A}$. El recorrido de f es el conjunto $\{0,1\}$, el recorrido de g es el conjunto $\{0, 1, 2, 3\}$, y el recorrido de <u>h</u> es el conjunto $\{0, 1, 4, 9\}$.

Ejercicios 5.1

---

1. Si $A = \{x: x = 1,0,1\}$, $B = \{-1, 0,1\}$, $C = \{0,1,-1\}$ y $D = \{-1, 1, -1\}$

   Marque cuáles de las siguientes proposiciones son verdaderas y cuáles falsas:

   1.1 $A = B$ F
   1.2 $A = C$ F
   1.3 $B = C$ V
   1.4 $A \subset B$ F
   1.5 $A \in B$ F
   1.6 $B \subseteq C$ V
   1.7 $C \subset A$ F
   1.8 $C \subseteq B$ V
   1.9 $0 \in B$ F
   1.10 $0 \subset C$ F
   1.11 $A = D$ F
   1.12 $D \subset A$ F
   1.13 $B \subset D$ F
   1.14 $D \in C$ V
   1.15 $B \neq C$ F

2. Si $E = \{x: 0 \leq x \leq 1\}$, $F = \{y: 0 \leq y \leq 1\}$, $G = \{x: 0 < x < 1\}$ y $H = \{1\}$,

   Marque cuáles de las siguientes proposiciones son verdaderas y cuáles falsas:

   F 2.1 $E = F$ F
   F 2.2 $F = G$ F
   F 2.3 $F \subset G$ F
   F 2.4 $G \subset F$ F
   F 2.5 $H \subset G$
   V 2.6 $H \subset E$ V
   V 2.7 $H \in F$ V
   F 2.8 $E \subseteq F$ F
   V 2.9 $H \notin E$ F

3. Es el conjunto de todos los estudiantes un subconjunto del conjunto de toda la población de personas? Es éste un subconjunto de toda población de personas menores de 36 años de edad? Debajo de 50 años de edad?

4. Demostrar que $A \cap B \subseteq A \subset A \cup B$ y que $A \cap B \subset B \subseteq A \cup B$

5. Si $A = \{1,0\}$, $B = \{x: 0 < x < 1\}$, $C = \{1/2\}$, computar

   $A \cup B$, $A \cup C$, $B \cup C$ y $A \cap B$, $A \cap C$, $B \cap C$,

6. Verificar que $B \cup B = B \cap B = B$ para todo B

7. Qué implican las siguientes igualdades:

   7.1 $E \cap F = F$? 7.2 $E \cup F = E$?

8. Definidos los conjuntos:

$$A = \{x: x = 1,2,3,\dots,10\}, \quad B = \{x: 1 \leq x \leq 10\}$$

$$C = \{x: x = 0,1,2,3,4,5,6\} \text{ y } D = \{0, 10,20,30\}$$

Calcular:

8.1 $A \cup B$	8.7 $B \cup C$	8.13 $A \cup B \cup C$
8.2 $A \cap B$	8.8 $B \cap C$	8.14 $A \cap (B \cup C)$
8.3 $A \cup C$	8.9 $B \cup D$	8.15 $A \cup (B \cap C)$
8.4 $A \cap C$	8.10 $B \cap D$	8.16 $A \cap B \cap C$
8.5 $A \cup D$	8.11 $C \cup D$	8.17 $C \cup (A \cap D)$
8.6 $A \cap D$	8.12 $C \cap D$	8.18 $(A \cup B) \cap (C \cup B)$

9. Dado el conjunto universal $\Omega = \{1,2,3,\dots,n\}$ y los conjuntos $A = \{1,2,3\}$ y $B = \{2,3,n\}$,

   Calcular A^c, B^c y $A^c \cup B^c$. Compare este último conjunto con $(A \cap B)^c$

10. Podemos definir un conjunto A, tal que $A = A^c$? Tal que $A \subset A^c$?

11. Ilustrar con diagramas de Venn los conjuntos $A \cap B^c$, $A^c \cap B^c$, $A \cup B \cup C$ $A \cap B \cap C$, $(A \cup B) \cap C$, $(A \cap C) \cup (B \cap C)$

12. Dados los conjuntos $A = \{1,2\}$, $B = \{2,1\}$ y $C = \{10,12\}$,

    formar los productos Cartesianos AxB, AxC, BxC, BxA, CxA, CxB, AxA, BxB, CxC, AxBxC, CxBxA, CxAxB.

13. Cuáles son las condiciones para que AxB = BxA?

14. Si A es el conjunto de hombres casados en La Paz y B el conjunto de mujeres casadas que viven en La Paz, es AxB el conjunto de parejas casadas que viven en La Paz?

15. Sea A el conjunto de personas de mi familia y definida la función $f(w)$ = edad de w para $w \in A$. Especificar el recorrido de f.

16. Sean $B = \{1,2,3,4,5,6\}$ y $C = BxB$

    Si se define la función elemento

$$g\left[(x_1, x_2)\right] = x_1 + x_2 \quad \text{para } (x_1, x_2) \in C,$$

Cuál es el rango de g?

17. Si $C = \{(x_1, x_2): \ x_1 = 1, 2, 3, 4, 5, 6; \ x_2 = 1, 2, 3, 4, 5, 6\}$ y dada la función $g[(x_1, x_2)] = x_1 + x_2$, para $(x_1, x_2) \in C$.

    Especificar

$$A_1 = \{(x_1, x_2): g[(x_1, x_2)] = 7\}$$

$$A_2 = \{(x_1, x_2): g[(x_1, x_2)] = 3\}$$

$$A_3 = \{(x_1, x_2): g[(x_1, x_2)] = 10\}$$

18. Sea $\mathcal{F}$ la clase de todos los subconjuntos de $S = \{1, 2, \ldots, k\}$. Definido $n(A)$ = número de elementos en A para $A \in \mathcal{F}$. Cuál es el recorrido de $\underline{n}$?

19. Sean S y n definidos como en el problema anterior. Mostrar que

    19.1 $n(A) \leq n(B)$, si $A \subset B$

    19.2 $n(A \cup B) = n(A) + n(B)$, si $A \cap B = \emptyset$

5.4 Probabilidad

Hemos mencionado que existen tres métodos para llegar al concepto de probabilidad: La primera forma, y la más antigua, consiste en repetir muchas veces un experimento o juego en las mismas condiciones, y calcular la frecuencia relativa con que ocurre un resultado (evento o suceso); es decir, dividir el número de veces que ocurre un suceso deseado entre el número de casos posibles. Esto es lo que suele llamarse "casos favorables entre casos posibles" (medida que recibe el nombre de *frecuencia relativa* del suceso). Este método históricamente comenzó con el estudio de los juegos de azar, tal como los dados, las cartas, la ruleta, etc. Sin embargo aún sigue siendo de gran utilidad, particularmente en los casos en que se asigna la misma posibilidad de ocurrencia a todos los resultados posibles de un experimento.

Otro método para llegar a la noción de probabilidad, sobre todo desde el punto de vista computacional, es el *método combinatorio*, muy útil pero de aplicación limitada, dado que este modelo parte del principio de que todos los sucesos son "igualmente posibles"; pero en la práctica, no siempre ocurre este hecho.

En el presente trabajo, dadas las limitaciones de los dos métodos anteriores, trataremos de llegar al concepto de probabilidad, mediante el *método axiomático*. Es decir, que se parte de un conjunto de axiomas que nos permiten deducir las propiedades y aplicaciones de las probabilidades.

La probabilidad es una herramienta indispensable para toda clase de investigaciones que implican incertidumbre. Si se está frente a experimentos cuyos resultados están completamente determinados, es decir que de antemano se sabe qué suceso ocurrirá, entonces desaparece el problema de predicción y por lo tanto no hay necesidad de recurrir al cálculo de probabilidades. Sin embargo, como hay una infinidad de fenómenos en la naturaleza los cuales implican insertidumbre, la importancia de considerar la teoría de probabilidades en nuestro estudio, es realmente relevante.

La teoría de probabilidades se emplea en el tratamiento y solución de muchos problemas de peso, particularmente en casi todos los campos de la investigación científica.

La Inferencia Estadística, básicamente recurre de este modelo matemático. Las probabilidades están relacionadas íntimamente, por ejemplo, con problemas de toma de decisiones, tanto a nivel empresarial como a nivel gubernamental. Sirve para la formulación de modelos econométricos, para fines de planificación económica, evaluación de proyectos, estudios de futuros mercados, controles de producción, estimaciones de accidentes en general, para determinar disturbios en mecanismos eléctricos, etc.,etc.

En la sección que sigue, introduciremos algunos conceptos previos al desarrollo del cálculo de probabilidades, y en los capítulos siguientes discutiremos las aplicaciones que serán de nuestro mayor interés.

5.5

Espacio Muestral y Sucesos

Como se ha mencionado en la primera parte de este capítulo, la teoría de conjuntos es un auxiliar ideal para un estudio eficiente de la teoría de probabilidades. Después de las siguientes definiciones básicas, veremos de qué manera se utiliza la noción de conjuntos.

Para nosotros, un *experimento*, será cualquier operación cuyos resultados no son posibles de predecir con certeza.

En otras palabras, supongamos que estamos interesados en saber qué ocurrirá si se realiza un cierto acto. Se procede a realizar el acto una o más veces bajo las mismas condiciones. Un proceso tal, se llama un proceso aleatorio. Cada ejecución, se llama *prueba*. Todas las pruebas efectuadas bajo las mismas condiciones forman un experimento. De manera que diferentes tipos de pruebas dan lugar a diferentes experimentos. Vemos que un experimento puede consistir de una sola prueba. El acontecimiento de una prueba se llama un *resultado*, un *punto muestral*, o un *suceso elemental*. La colección, o el conjunto, de todos los resultados posibles de un experimento (ésto es, el conjunto de puntosmuestrales) constituirá un *espacio muestral*.

Para fines de especulación teórica y práctica, los experimentos pueden ser reales o conceptuales. Un experimento es *real*, cuando ejecutamos una operación con objetos tangibles desde el punto de vista físico; así por ejemplo: tomar una moneda y lanzarla, ingerir un medicamento, inyectar un virus en un animal para ver su reacción, preguntar a algunos de los obreros de una fábrica lo que ganan mensualmente, medir el tiempo de duración de un artefacto eléctrico, medir los pesos de un conjunto de naranjas, etc., etc. Por otro lado, un experimento será *conceptual*, si es posible realizarlo sin necesidad de tomar los objetos físicamente; así por ejemplo, imaginar que se lanza un dado, y hallar todas las conclusiones tal como si se realizara el experimento físicamente.

En algunos casos, hay objetos que se destruyen cuando son sometidos a un experimento determinado, así por ejemplo, cuando se prueba la resistencia de duración de ciertos artefactos eléctricos. Por esta razón, es aconsejable probar solamente con algunos artículos, cuando se desea saber la bondad de un cierto proceso de producción. Más adelante, veremos cómo se obtienen inferencias si se sigue este método de investigación.