USAC
FACULTAD DE INGENIERÍA
ÁREA DE ESTADÍSTICA
Coordinación

MANUAL DE ESTADÍSTICA DESCRIPTIVA

Guatemala, noviembre 2011

ÍNDICE DE CONTENIDOS

página

ESTADÍSTICA DESCRIPTIVA

DÍA 1

I. Unidad: Introducción a la Estadística

1.1 Estadística

Se denomina Estadística a la rama de las matemáticas que se ocupa de reunir, organizar, presentar, analizar e interpretar datos numéricos y que ayuda a resolver problemas como el diseño de experimentos y la toma de decisiones.

1.2 Tipos de Estadística

1.2.1 Estadística descriptiva

Se encarga de la recolección, agrupación y presentación de los datos de una manera tal que los describa fácil y rápidamente.

1.2.2 Estadística inferencial

Involucra la utilización de una muestra para sacar alguna inferencia o conclusión sobre la población de la cual procede la muestra. Puede definirse como aquellos métodos que hacen posible la estimación de una característica de una población o la toma de una decisión referente a una población, basándose sólo en los resultados de una muestra. El objetivo de la inferencia estadística es obtener información acerca de la población, partiendo de la información que contiene la muestra.

A la característica numérica de una población, como el promedio de la población, la desviación estándar de la población, etc., se le denomina **parámetro**. El parámetro es una medida de resumen que se calcula para describir una característica de toda una población.

A la característica numérica de una muestra, como el promedio de la muestra, la desviación estándar de la muestra, etc., se le denomina **Estadístico**. El estadístico es una medida de resumen que se calcula para describir una característica de una sola muestra de la población.

La **población** es el conjunto de todos los individuos. Como **individuo** se entiende cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo.

Una **muestra** es un subconjunto que seleccionamos de la población, es una parte representativa de la población que se selecciona para ser estudiada ya que la población es demasiado grande para analizar su totalidad. Al Proceso de obtener muestras se le conoce como **muestreo.**

Existen dos tipos de muestreo: **aleatorio** y **no aleatorio**. El **muestreo aleatorio** es en el cual la muestra se obtiene dando la misma oportunidad a cada elemento de la población de pertenecer a ella. Al muestro aleatorio también se le conoce como muestreo **representativo**. Mientras que en el **muestreo no aleatorio** la muestra se obtiene sin darle la misma oportunidad a cada elemento de la población de pertenecer a ella. El muestreo aleatorio puede llevarse a cabo a través de urnas o tómbolas, o bien usando números aleatorios, mientras que el muestreo no aleatorio se realiza a juicio.

Una segunda clasificación del muestreo surge en la forma en que se selecciona la muestra, así el muestreo puede ser **con reemplazo** y **sin reemplazo. El muestreo con reemplazo** es el muestreo en el cual cada miembro de una población puede seleccionarse más de una vez, cada vez que se toma un elemento la población conservará su tamaño. El **Muestreo sin reemplazo** es en el cual cada miembro de una población puede seleccionarse únicamente una vez y en este caso el tamaño de la población se va reduciendo conforme se conforma la muestra.

1.3 Tipos de variables

Una **variable** es una característica de interés sobre cada elemento individual de una población o muestra. Un **dato** es el valor de la variable asociada a un elemento de una población o muestra. Este valor puede ser un número, una palabra o un símbolo. Un **experimento** es una actividad planeada cuyos resultados producen un conjunto de datos.

Dependiendo del número de características que se analizan de la población, las variables se pueden clasificar en:

a) Variables unidimensionales: sólo recogen información sobre una característica. Ejemplo: edad de los alumnos de una clase.
b) Variables bidimensionales: recogen información sobre dos características de la población. Ejemplo: edad y altura de los alumnos de una clase.
c) Variables pluridimensionales o multidimensionales: recogen información sobre tres o más características. Ejemplo: edad, altura y peso de los alumnos de una clase.

Dependiendo del tipo de datos las variables pueden clasificarse en:

a) Variables cualitativas o atributos: no se pueden medir numéricamente. Ejemplo: religión, nacionalidad, color de la piel, sexo.
b) Variables cuantitativas: tienen valor numérico. Ejemplo: edad, longitud, precio.

Por su parte, las **variables cuantitativas** se pueden clasificar en discretas y continuas.

a) Discretas: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Estas variables representan conteos, por ejemplo: el número de alumnos en un salón de clase puede ser: 35, 60, 100, etc., nunca podrá ser 41.3.
b) Continuas: pueden tomar cualquier valor real dentro de un intervalo. Estas variables representan mediciones, por ejemplo, la altura de una persona puede ser 1.35 m, 1.68, 1.90, etc.

1.4 Niveles de medición

Los **Niveles o Escalas de medición** son las formas de clasificar los datos, pueden ser:

a) Escala Nominal: se caracteriza por datos que consisten exclusivamente en nombres, rótulos o categorías. Los datos no pueden acomodarse según esquema de ordenamiento (digamos de bajo alto). El término nominal puede asociarse con “sólo nombres”.

La escala de medida nominal, puede considerarse la escala de nivel más bajo, y consiste en la asignación, puramente arbitraria de números o símbolos a cada una de las diferentes categorías en las cuales podemos dividir el carácter que observamos, sin que puedan establecerse relaciones entre dichas categorías, a no ser el de que cada elemento pueda pertenecer a una y solo una de estas categorías.

Se trata de agrupar objetos en clases, de modo que todos los que pertenezcan a la misma sean equivalentes respecto del atributo o propiedad en estudio, después de lo cual se

asignan nombres a tales clases, y el hecho de que a veces, en lugar de denominaciones, se le atribuyan números, puede ser una de las razones por las cuales se le conoce como "medidas nominales".

Se ha de tener presente que los números asignados a cada categoría sirven única y exclusivamente para identificar la categoría y no poseen propiedades cuantitativas.

b) Escala Ordinal: implica datos que pueden acomodarse en algún orden, pero no es posible determinar diferencias entre los valores de los datos, o tales diferencias carecen de significado.

En caso de que puedan detectarse diversos grados de un atributo o propiedad de un objeto, la medida ordinal es la indicada, puesto que entonces puede recurrirse a la propiedad de "orden" de los números asignándolo a los objetos en estudio de modo que, si la cifra asignada al objeto A es mayor que la de B, puede inferirse que A posee un mayor grado de atributo que B.

La asignación de números a las distintas categorías no puede ser completamente arbitraria, debe hacerse atendiendo al orden existente entre éstas.

Los caracteres que posee una escala de medida ordinal permiten, por el hecho mismo de poder ordenar todas sus categorías, el cálculo de las medidas estadísticas de posición, como por ejemplo la mediana.

Ejemplo: Al asignar un número a los vehículos en un taller de servicio, según el orden de llegada, estamos llevando una escala ordinal, es decir que al primero en llegar le asignamos el nº 1, al siguiente el nº 2 y así sucesivamente, de esta forma, cada número representará una categoría en general, con un solo elemento y se puede establecer relaciones entre ellas, ya que los números asignados guardan la misma relación que el orden de llegada al taller.

c) Escalas por intervalos es como el nivel ordinal, con la propiedad adicional que podemos determinar magnitudes de diferencias entre los datos que tienen algún significado. Sin embargo, no hay un punto de partida o cero inherente (natural) en el que la cantidad esté totalmente ausente.

En esta escala, además de poseer las características de la escala ordinal, encontramos que la asignación de los números a los elementos es tan precisa que podemos determinar la magnitud de los intervalos (distancia) entre todos los elementos de la escala. Sin lugar a dudas, podemos decir que la escala por intervalos es la primera escala verdaderamente cuantitativa y a los caracteres que posean esta escala de medida pueden calculársele todas las medidas estadísticas a excepción del coeficiente de variación.

Ejemplos: El lapso transcurrido entre 1998-1999 es igual al que transcurrió entre 2000-2001.

Si a una hora tenemos una temperatura ambiente de 15ºC significa que hace frío, si después de varias horas la temperatura cambia a 30ºC no significa que hace el doble de frío, y si se tuviera 0ºC no significa que ya no hace frío o bien que no hay temperatura.

d) Escala de razón: Es el nivel de medida más elevado y se diferencia de las escalas de intervalos únicamente por poseer un punto cero propio como origen; es decir que el valor cero de esta escala significa ausencia de la magnitud que estamos midiendo.

Si se observa una carencia total de propiedad, se dispone de una unidad de medida para el efecto. A iguales diferencias entre los números asignados corresponden iguales diferencias

en el grado de atributo presente en el objeto de estudio. Además, siendo que cero ya no es arbitrario, sino un valor absoluto, podemos decir que A tiene dos, tres o cuatro veces la magnitud de la propiedad presente en B.

1.5 Recopilación de datos

1.5.1 Fuentes para obtener datos

Los datos pueden obtenerse de dos tipos de fuentes:

a) **Fuentes internas:** cuando los datos son parte de la propia actividad del ente que los recopila, se dice que el dato es interno y la fuente es interna.
b) **Fuentes externas:** cuando se tiene que otras empresas, instituciones, poblaciones, etc., fuera del ente recopilador.

1.5.2 Técnicas para recopilar datos

Para obtener la información existen varias técnicas, entre estas: encuesta, entrevista, cuestionario y observación.

a) **Encuesta:** Conjunto de preguntas tipificadas dirigidas a una muestra representativa, para averiguar estados de opinión o diversas cuestiones de hecho. A diferencia de un censo, donde todos los miembros de la población son estudiados, las encuestas recogen información de una porción de la población de interés, dependiendo el tamaño de la muestra en el propósito del estudio.

b) **Entrevista:** Las entrevistas se utilizan para recabar información en forma verbal, a través de preguntas que propone el analista. Quienes responden pueden ser gerentes o empleados, los cuales son usuarios actuales del sistema existente, usuarios potenciales del sistema propuesto o aquellos que proporcionarán datos o serán afectados por la aplicación propuesta. El analista puede entrevistar al personal en forma individual o en grupos. Sin embargo, las entrevistas no siempre son la mejor fuente de datos de aplicación. En otras palabras, la entrevista es un intercambio de información que se efectúa cara a cara. Es un canal de comunicación entre el analista y la organización; sirve para obtener información acerca de las necesidades y la manera de satisfacerlas, así como concejo y comprensión por parte del usuario para toda idea o método nuevos. Por otra parte, la entrevista ofrece al analista una excelente oportunidad para establecer una corriente de simpatía con el personal usuario, lo cual es fundamental en transcurso del estudio.

c) **Cuestionario:** se entiende por cuestionario a la lista de preguntas que se proponen por cualquier fin, el cuestionario proporcionan una alternativa muy útil para la entrevista; si embargo, existen ciertas características que pueden ser apropiada en algunas situaciones e inapropiadas en otra. Al igual que la entrevistas, deben diseñarse cuidadosamente para una máxima efectividad.

d) **Observación:** Otra técnica útil para el analista en su progreso de investigación, consiste en observar a las personas cuando efectúan su trabajo. Como técnica de investigación, la observación tiene amplia aceptación científica. Los sociólogos, sicólogos e ingenieros industriales utilizan extensamente ésta técnica con el fin de estudiar a las personas en sus actividades de grupo y como miembros de la organización. El propósito de la organización es múltiple: permite al analista determinar que se está haciendo, como se está haciendo, quien lo hace, cuando se lleva a cabo, cuánto tiempo toma, dónde se hace y por qué se hace.

DÍA 2

II. Unidad: Presentación de datos de una sola variable

2.1 Distribución de frecuencias

La información estadística puede constar de un gran número de observaciones y mientras mayor sea el número, mayor puede ser la conveniencia y necesidad de presentarla en forma resumida, la cual puede permitir algunos detalles pero en cambio puede revelar la naturaleza general de la información. Un resumen de tal distribución se denomina **Distribución de Frecuencias**. Puede decirse también, que una distribución es el patrón de variabilidad mostrado por los datos de una variable. La distribución muestra la frecuencia de cada valor de la variable.

La tabla No.1 resume las edades de 1,763,000 varones que constituían la fuerza laboral masculina de cierto país:

Tabla No. 1
Fuerza laboral masculina de un país
(clasificación por edades)

Edad	Número de varones
14 a19 años	218,000
20 a 24 años	313,000
25 a 55 años	977,000
Más de 55 años	255,000
Total	**1,763,000**

La tabla No. 1 presenta datos de una variable cuantitativa, por lo cual se trata de una Distribución de Frecuencia Cuantitativa, por otra parte, la tabla No.2 es una Distribución de Frecuencias Cualitativas, debido a que el campo de especialización del alumno no puede ser medidos sino sólo puede ser descrito.

Tabla No. 2
Campos de Especialización
de los alumnos de áreas técnicas

Campo de especialización	Número de alumnos
Construcción	42
Electrónica	88
Eléctrica	50
Mecánica	34
Total	**214**

2.1.1 Intervalos, marcas de clase y frecuencias

La dificultad de resumir un conjunto de datos, puede ser superada agrupando los diversos valores en un número reducido de clases llamados **intervalos de clase.** Cada una de las clases tiene un extremo o límite superior y uno inferior; el extremo inferior es el menor valor que puede caer en esta clase y el superior el mayor valor.

El punto medio entre el límite superior de una clase y el límite inferior de la siguiente clase es la **frontera superior o límite real superior** de la clase y la **frontera inferior o límite real inferior** de la siguiente clase. En una clase dada todos los valores deben ser mayores a la frontera inferior y menores a la frontera superior. Para evitar ambigüedades, las fronteras se expresan con una cifra decimal más que los extremos.

La diferencia entre las fronteras superior e inferior de una clase se denomina **amplitud de clase**.

El punto medio entre los dos extremos (o las dos fronteras) de una clase se denomina **marca de clase**.

El número de datos incluidos en un intervalo de clase se denomina **frecuencia de la clase**.

2.1.2 Construcción de una distribución de frecuencias para datos cuantitativos

Al construir una distribución de frecuencias para datos cuantitativos es necesario primeramente decidir cuál va a ser el número de clases. En general, este número depende fundamentalmente de la naturaleza de los datos a resumir y del objetivo que se persiga con ese resumen. Sin embargo, es posible dar ciertas guías generales que pueden ser de utilidad en la determinación del número de clases. En primer lugar, el número de clases no debe ser ni muy grande ni muy pequeño; un número pequeño de clases puede ocultar la naturaleza general de los datos y uno muy grande puede ser demasiado detallado como para relevar alguna información útil. Como regla general, se recomienda que **el número de clases esté entre 5 y 20.** La llamada regla de Sturges puede dar una aproximación razonable para el número de clases, siendo esta:

K = número de clases

$$K = 1 + 3.3 \log N$$

Una vez determinado el número de clases, debe decidirse la amplitud de estas. Tomando la misma amplitud para todas las clases, este valor queda dado por:

A = amplitud

$$A = \frac{\text{dato mayor} - \text{dato menor}}{K}$$

Una vez obtenida la amplitud de las clases se procede a calcular los intervalos y a realizar el conteo de valores para determinar la frecuencia de cada uno.

Ejemplo:

Las velocidades, en millas por hora, de los conductores de 55 automóviles fueron medidas con un radar en una calle de cierta ciudad:

27	23	22	38	43	24	35	26	28	20	18
25	23	22	63	31	30	41	45	29	43	27
29	28	27	25	29	28	24	37	28	18	29
26	33	25	27	25	34	32	36	22	33	32
21	23	24	18	48	23	15	38	26	23	21

Construya una distribución de frecuencias para las velocidades de estos 55 automóviles.

Solución:

Primero.

Para una mayor facilidad se recomienda ordenar los datos de menor a mayor.

15	21	23	24	25	27	28	29	32	36	43
18	21	23	24	25	27	28	29	33	37	43
18	22	23	24	26	27	28	30	33	38	45
18	22	23	25	26	27	29	31	34	38	48
20	22	23	25	26	28	29	32	35	41	63

Segundo

Calcular el número de clases.

$$K = 1 + 3.3 \text{ Log } (55)$$

$$K = 6.743$$

De acuerdo a la regla de Sturges, deberíamos tener 6 ó 7 clases. Para efectos de cálculos el valor de K se aproxima el entero más próximo.

K = 7

Tercero.

Calcular la amplitud.

Para esto previamente identificamos el dato mayor y el menor, en nuestro caso tales datos son 15 y 63

$$A = \frac{63 - 15}{7}$$

$$A = 6.857$$

La amplitud debe aproximarse al entero más cercano.

A = 7

Cuarto.

Una vez determinado el número de de clases y la amplitud, debe elegirse el extremo inferior de la primera clase. Dado que aquí el valor mínimo es 15, el extremo inferior puede ser 15 o menos; por consiguiente tomaremos como criterio usar el número 15.

Quinto.

Establecido el extremo inferior, se sumará la amplitud a éste para obtener el valor del límite inferior de la siguiente clase y así sucesivamente. Para obtener los límites superiores, se le resta uno al límite inferior posterior. Se tiene que tomar en cuenta que en la última clase esté contenido el dato mayor.

Sexto.

Corresponde ahora calcular la frontera inferior de la clase. Puesto que los valores están dados en números enteros y como las fronteras deben darse con un decimal más, tomamos como frontera inferior el valor de la primera clase inferior menos 0.05 (si los valores se hubieran dado con un decimal, se le restaría 0.005) y como frontera superior el valor de la primera clase superior más 0.05 (si los valores se hubieran dado con un decimal, se le sumaría 0.005)

Intervalo de clase			Frontera Inferior	Frontera Superior	Amplitud de clase
15	-	21	14.5	21.5	7
22	-	28	21.5	28.5	7
29	-	35	28.5	35.5	7
36	-	42	35.5	42.5	7
43	-	49	42.5	49.5	7
50	-	56	49.5	56.5	7
57	-	63	56.5	63.5	7

Séptimo.

Una vez construidos los diversos intervalos de clase, se cuenta el número de elementos que cae en cada uno, obteniéndose así las respectivas frecuencias.

Tabla No. 3
Distribución de frecuencia
Velocidades de un grupo de conductores en una autopista

Intervalo de clase			frecuencia
15	-	21	7
22	-	28	26
29	-	35	12
36	-	42	5
43	-	49	4
50	-	56	1
T o t al			**55**

2.1.3 Distribución de frecuencia relativa

La distribución de frecuencias es una tabla resumen en la que los datos originales se condensan o agrupan para facilitar el análisis de los datos. Sin embargo, para ampliar el análisis, es deseable formar la distribución de frecuencia relativa o la distribución de porcentaje, dependiendo de si se prefieren fracciones o porcentajes.

La **frecuencia relativa (fr)** es la relación entre la frecuencia de un intervalo y el número total de datos:

$$\mathbf{fr = f_i/n}$$

La **frecuencia porcentual (fr%)** es la expresión en porcentaje de la frecuencia relativa:

$$\mathbf{(fr\%) = fr * 100}$$

Tabla No. 4
Distribución de frecuencias absoluta y relativa
Velocidades de un grupo de conductores en una autopista

Intervalo de clase			frecuencia absoluta f	frecuencia realtiva fr	frecuencia porcentual fr%
15	-	21	7	0.1273	12.73
22	-	28	26	0.4727	47.27
29	-	35	12	0.2182	21.82
36	-	42	5	0.0909	9.09
43	-	49	4	0.0727	7.27
50	-	56	1	0.0182	1.82
T o t al			**55**	**1.0000**	**100.00**

2.1.4 Distribuciones de frecuencia acumulada

Otras representaciones de datos que facilitan el análisis y la interpretación son las distribuciones acumulativas. Éstas pueden desarrollarse a partir de la tabla de distribución de frecuencia, de la tabla de distribución de frecuencia relativa y de la tabla de distribución de frecuencia porcentual

La **frecuencia acumulada (F)** indica el número de observaciones acumuladas en cada intervalo. Para calcularla, en cada intervalo se consideran las frecuencias anteriores, de tal forma que el último intervalo contenga el total de observaciones. Dependiendo de la preferencia o necesidad para presentar los resultados, esta frecuencia puede calcularse utilizando las frecuencias relativas, en este caso se denomina **frecuencia relativa acumulada (Fr)** o bien si se usa la frecuencia porcentual **frecuencia relativa porcentual acumulada (Fr%).**

Tabla No. 5
Distribución de frecuencia acumulada
Velocidades de un grupo de conductores en una autopista

Intervalo de clase	frecuencia absoluta f	frecuencia acumulada F
15 - 21	7	7
22 - 28	26	33
29 - 35	12	45
36 - 42	5	50
43 - 49	4	54
50 - 56	1	55
T o t al	**55**	

Tabla No. 6
Distribución de frecuencia relativa acumulada
Velocidades de un grupo de conductores en una autopista

Intervalo de clase	frecuencia absoluta f	frecuencia realtiva fr	frecuencia relativa acumulada Fr
15 - 21	7	0.1273	0.1273
22 - 28	26	0.4727	0.6000
29 - 35	12	0.2182	0.8182
36 - 42	5	0.0909	0.9091
43 - 49	4	0.0727	0.9818
50 - 56	1	0.0182	1.0000
T o t al	**55**	**1.0000**	

Tabla No. 7
Distribución de frecuencia relativa porcentual acumulada
Velocidades de un grupo de conductores en una autopista

Intervalo de clase	frecuencia porcentual fr	frecuencia porcentual acumulada Fr
15 - 21	12.73%	12.73%
22 - 28	47.27%	60.00%
29 - 35	21.82%	81.82%
36 - 42	9.09%	90.91%
43 - 49	7.27%	98.18%
50 - 56	1.82%	100.00%
T o t al	**100.00%**	

DÍA 3

2.2 Presentación gráfica de datos

Una vez elaborada la tabla de distribución de frecuencia es importante construir su representación visual. Esta representación revela patrones de comportamiento de la variable en estudio. El tipo de gráfico que se utilice dependerá del tipo de datos y el concepto a representar.

2.2.1 Datos cualitativos

Las gráficas que generalmente se utilizan para resumir datos cualitativos, de atributo o categóricos son las **gráficas de barras** y la **de pastel.**

2.2.1.1 Gráfica de barras

Las gráficas de barras muestran la cantidad de datos que pertenecen a cada categoría como áreas rectangulares de tamaño proporcional. Cada barra sólida, ya sea vertical u horizontal representa un tipo de dato.

2.2.1.2 Gráfica Circular

Denominada también gráfica de pastel o gráfica del 100%, se utilizan para mostrar la cantidad de datos que pertenecen a cada categoría como una parte proporcional de un círculo.

Se forma al dividir un círculo en sectores circulares de manera que:

a) Cada sector circular equivale al porcentaje correspondiente al dato o grupo que representa.
b) La unión de los sectores circulares forma el círculo y la suma de sus porcentajes es 100.

Es aconsejable que el número de elementos comparados dentro de un gráfico circular, no sea mayor de 5, ordenando los segmentos de mayor a menor, iniciando con el más amplio a partir de las 12 como en un reloj. Una manera sencilla de diferenciar los segmentos es sombreándolos de claro a oscuro, siendo el de mayor tamaño el más claro y el de menor tamaño el más oscuro.

DÍA 4

2.2.2 Datos cuantitativos

Una razón fundamental para elaborar una gráfica de datos cuantitativos es mostrar su distribución.

2.2.2.1 Histograma

Una de las formas más comunes de representar una distribución de frecuencias es un histograma.

Un histograma es una gráfica que se construye a partir de la tabla estadística, consiste en rectángulos verticales unidos entre sí, en donde sus lados son los límites reales inferior y superior de clase y cuya altura es igual a la frecuencia de clase. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y el área de los mismos.

2.2.2.2 Polígono de frecuencias

Consiste en una serie de segmentos que unen los puntos cuyas abscisas son las marcas de cada clase y cuyas ordenadas son proporcionales a sus frecuencias respectivas.

El polígono de frecuencias se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase. Para representar el polígono de frecuencias en el primer y último intervalo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia nula, y se unen por una línea recta los puntos del histograma que corresponden a sus marcas de clase. De este modo, el polígono de frecuencias tiene en común con el histograma el que las áreas de la gráfica sobre un intervalo son idénticas.

2.2.2.3 Ojiva

Una gráfica de distribución de frecuencias acumuladas es llamada una ojiva. Se trazan los límites reales superiores contra las frecuencias acumuladas.

Una distribución de frecuencias acumuladas nos permite ver cuántas observaciones están por encima de ciertos valores, en lugar de hacer un mero registro del número de elementos que hay dentro de los intervalos, esto es lo que refleja la ojiva.

Se puede construir una ojiva de una distribución de frecuencias relativas de la misma manera en que trazamos la ojiva de una distribución de frecuencias absolutas. Sólo habrá un cambio: la escala del eje vertical.

2.2.2.4 Gráficas de puntos

La representación gráfica por medio de puntos (o gráfica de puntos) es una de las gráficas más sencillas que se utilizan.

Presenta los datos de una muestra mediante la representación de cada porción de datos con un punto ubicado a lo largo de una escala. Esta escala puede ser vertical y horizontal. La frecuencia del los valores está representada a lo largo de la otra escala.

2.2.2.5 Gráficas lineales

Consisten en una serie de puntos trazados en las intersecciones de las marcas de clase y las frecuencias de cada una, uniéndose consecutivamente con líneas.

DÍA 5

III. Unidad: Análisis descriptivo de datos de una sola variable

---

3. Medidas de posición

Las medidas de posición facilitan información sobre la serie de datos que se está analizando. Estas medidas permiten conocer diversas características de la serie de datos.

3.1 Medidas de tendencia central

Informan sobre los valores medios del conjunto de datos. Son indicadores usados para señalar que porcentaje de datos dentro de una distribución de frecuencias superan estas expresiones, cuyo valor representa el valor del dato que se encuentra en el centro de la distribución de frecuencia, es por esto que se les llama "Medidas de Tendencia Central".

3.1.1 Media

Las media o promedio es una medida de posición que proporciona una descripción compacta de cómo están centrados los datos y una visualización más clara del nivel que alcanza la variable,

puede servir de base para medir o evaluar valores extremos y brinda mayor facilidad para efectuar comparaciones.

Es importante poner en relieve que la notación de promedio lleva implícita la idea de variación y que este número promedio debe cumplir con la condición de ser representativo de conjunto de datos.

El promedio como punto típico de los datos es el valor al rededor del cual se agrupan los demás valores de la variable.

3.1.1.1 Media Aritmética

Es una medida matemática, un número individual que representa razonablemente el comportamiento de todos los datos.

Características de la Media:

a. En su cálculo están todos los valores del conjunto de datos por lo que cada uno afecta la media.

b. La suma algebraica de las desviaciones de los valores individuales respecto a la media es cero.

c. La suma del cuadrado de las desviaciones de una serie de datos a cualquier número A es mínimo si A = X

d. Aunque es confiable porque refleja todos los valores del conjunto de datos puede ser afectada por los valores extremos, y de esa forma llegar a ser una medida menos representativa, por lo que si la distribución es asimétrica, la media aritmética no constituye un valor típico.

CÁLCULO DE LA MEDIA ARITMÉTICA
Para datos no agrupados $\overline{X} = \frac{\sum x_i}{n}$
Para datos agrupados $\overline{X} = \frac{\sum (x_i * f_i)}{\sum fi}$

3.1.1.2 Media ponderada

La media ponderada toma en cuenta la importancia relativa de las observaciones, así, para cada uno de los valores de xi se asigna un factor wi de peso, que depende de la importancia que el investigador desee darle.

CÁLCULO DE LA MEDIA PONDERADA
Para datos no agrupados $\overline{X}_w = \frac{\sum (x_i * w_i)}{\sum w_i}$
Para datos no agrupados $\overline{X}_w = \frac{\sum (x_i * w_i)}{\sum f_i * w_i}$

3.1.1.3 Media geométrica

La media geométrica es útil cuando la variable cambia a lo largo del tiempo, esto es, en el cálculo del promedio de tasas, razones, proporciones geométricas y relaciones de variables. Se utiliza en Matemáticas Financieras y Finanzas para promediar números índices, tasas de cambio, etc.

Esta media se ve afectada por todos los números y valores extremos pero en menor grado que la Media Aritmética, su valor siempre es menor que el de ésta.

Según el tipo de datos que se analice será más apropiado utilizar la media aritmética o la media geométrica.

La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.

La media Geométrica de una serie de números es la raíz n-ésima del producto de esos números:

CÁLCULO DE LA MEDIA GEOMÉTRICA
$\overline{X}_G = \sqrt[n]{x_1 * x_2 * x_3 ... * x_n}$

3.1.2 Mediana

Es el valor de la observación que ocupa la posición central de un conjunto de datos ordenados según su magnitud. Es el valor medio o la media aritmética de los valores medios. La mediana es un valor de la variable que deja por debajo de él un número de casos igual al que deja por arriba.

Geométricamente la mediana es el valor de la variable que corresponde a la vertical que divide al histograma en dos áreas iguales.

Cuando determinados valores de un conjunto de observaciones son muy grandes o pequeños con respecto a los demás, entonces la media aritmética se puede distorsionar y perder su carácter representativo, en esos casos es conveniente utilizar la mediana como medida de tendencia central, es decir que la mediana no presenta el problema de estar influida por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos (no pondera cada valor por el número de veces que se ha repetido).

Características de la mediana

a. Es una medida de tendencia central no afectada por los valores extremos.

b. No está definida algebraicamente.

c. Cuando la localización del elemento central puede ser determinada y los límites de clase mediana son conocidos, la mediana para la distribución de frecuencias puede ser calculada por interpolación, no importando que ésta contenga intervalos abiertos, cerrados, iguales o diferentes.

d. La suma de los valores absolutos, sin considerar el signo, de las desviaciones individuales respecto a la mediana es mínimo.

e. La mediana en caso de una distribución asimétrica, no resulta desplazado del punto de tendencia central.

f. Si el universo tiene curtosis excesiva la mediana como estadístico, varía menos que cualquier otra medida.

g. Si la mediana se calcula por interpolación y hay lagunas en los valores de la clase mediana o los datos son irregulares, esta medida no es buena ya que su ubicación puede resultar falsa.

h. Si se desea ubicar las condiciones de un elemento en una clase, la mediana resulta se indicada, ya que por comparación pone en evidencia si un elemento está en la mitad superior a ella o en la inferior.

CÁLCULO DE LA MEDIANA

Para datos no agrupados:
1ero. Se ordenan los datos ascendentemente.
2do. La mediana corresponde al dato que está en la posición central.

Para datos agrupados
1ero. Se calcula la **clase de la mediana**, la cual corresponde a la clase cuya frecuencia acumulada es mayor o igual a n/2.

2do. En la **clase de la mediana** se aplica la siguiente fórmula:

$$M_e = L_{me} + \left[\frac{n/2 - F}{f_{me}} \right] * A$$

Donde : L_{me} = Límite real inferior de la clase de la mediana
F = frecuencia acumulada de la clase anterior a la mediana
f_{me} = frecuencia absoluta de la clase de la mediana
A = amplitud del intervalo de la clase de la mediana

3.1.3 Moda

Es el valor de un conjunto de datos que ocurre más frecuentemente, se considera como el valor más típico de una serie de datos.

Para datos agrupados se define como Clase Modal el intervalo que tiene más frecuencia.

La moda puede no existir o no ser única, las distribuciones que presentan dos o más máximos relativos se designan de modo general como bimodales o multimodales respectivamente.

Características de la Moda:

a. Representa más elementos que cualquier otro valor

b. No está afectada por los valores extremos pero para datos continuos es dudoso su cálculo.

c. La moda para una distribución de frecuencias de datos agrupados no puede ser calculada exactamente, el valor de la moda puede ser afectado por el método de agrupación de los intervalos de clase.

d. La moda no permite conocer la mayor parte de los datos.

e. Algunas veces el azar interviene de manera importante y hace que un valor no representativo se repita frecuentemente.

f. Puede usarse para datos cuantitativos como cualitativos.

g. La moda como estadístico, varía mucho de una muestra a otra.

h. Cuando se tienen dos o más modas es difícil su interpretación

i. Tiene la ventaja de que los datos desproporcionados con respecto al resto no la distorsionan, pero no se presta para un tratamiento matemático.

CÁLCULO DE LA MODA

Para datos no agrupados:
La moda corresponde al dato o datos que se repiten con más frecuencia.

Para datos agrupados
1ero. Se localiza la **clase modal**, la cual corresponde a la clase que tenga la mayor frecuencia.

2do. En la **clase modal** se aplica la siguiente fórmula:

$$Mo = L_{mo} + \left[\frac{D_1}{D_1 + D_2} \right] * A$$

Donde : L_{mo} = Límite real inferior de la clase modal
D_1 = diferencia entre la frecuencia de la clase modal y la clase anterior
D_2 = diferencia entre la frecuencia de la clase modal y la clase posterior
A = amplitud del intervalo de la clase modal

DÍA 6

3.2 Medidas de tendencia no central

Las medidas de posición no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suelen utilizar una serie de valores que dividen la muestra en tramos iguales.

3.2.1 Deciles

Son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.

3.2.2 Cuartiles

Son 3 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.

3.2.3 Percentiles

Son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

CÁLCULO DE PERCENTILES

Para datos no agrupados:

1ero. Se ordenan los datos ascendentemente.

2do. Se calcula el índice (posición) del percentil con la siguiente fórmula

$$i = \left(\frac{p}{100}\right) * n$$

En donde *p* es el número de percentil de interés y *n* es la cantidad de observaciones.

3ero. Si *i no es entero, se redondea*. El valor entero inmediato *mayor* que *i* indica la posición del *p*-***èsimo*** percentil.

Si *i sí es entero*, el *p*-èsimo percentil es el promedio de los valores de los datos ubicados en los lugares *i* e *i* + 1.

Para datos agrupados

1ero. Se calcula el índice del percentil

$$i = \left(\frac{p}{100}\right) * \sum f$$

2do. Se localiza la **clase del percentil**, la cual corresponde a la clase cuya frecuencia acumulada es mayor o igual a *i*.

3ro. En la **clase del percentil** se aplica la siguiente fórmula:

$$P_i = L_{pi} + \left[\frac{\left(\frac{p}{100}\right) * \sum f - F}{f_{pi}}\right] * A$$

Donde :

L_{pi} = Límite real inferior de la clase del percentil

F = frecuencia acumulada de la clase anterior al percentil

F_{pi} = frecuencia absoluta de la clase del percentil

A = amplitud del intervalo de la clase del percentil

DÍA 7

3.3 Medidas de dispersión

Estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Es importante medir la dispersión por las siguientes razones:

a) La dispersión proporciona información adicional que permite juzgar la confiabilidad de la medida de tendencia central, si los datos se encuentran muy dispersos, la posición central es menos representativa de los datos, como un todo, que cuando éstos se agrupan más cerca del valor de la media.

b) Existen problemas característicos para datos muy dispersos, por lo que es necesario reconocer esa dispersión alta para poder abordar ese tipo de problemas.

c) Cuando se desea comparar diferentes muestras, si no se desea tener una alta dispersión de valores con respecto del centro de distribución, o esto presenta riesgos inaceptables, se necesita reconocerla y evitar elegir distribuciones que tengan las dispersiones más grandes.

Asimismo, la dispersión puede medirse desde tres enfoques, la distancia, la dispersión promedio y la dispersión relativa.

3.3.1 Medidas de distancia

La dispersión puede medirse en términos de la diferencia entre dos valores seleccionados del conjunto de datos, a continuación se presentan tres de las llamadas medidas de distancia.

3.3.1.1 Rango

Es la diferencia entre el más alto y el más pequeño de los valores observados. El rango es fácil de entender y de calcular, pero su utilidad como medida de dispersión es limitada, pues solo toma en cuenta el valor más grande y el más pequeño y ninguna otra observación del conjunto de datos, restándole importancia a las variaciones entre todas las demás observaciones.

RANGO (R)
R = Dato mayor – Dato menor

3.3.1.2 Rango Intercuartilico

El rango intercuartilico mide aproximadamente qué tan lejos de la mediana se debe ir en cualquiera de las dos direcciones antes de recorrer una mitad de los valores del conjunto de datos.

RANGO INTERCUARTÍLICO (RIQ)
$RIQ = Q_3 - Q_1$

3.3.1.3 Rango Interpercentìlico

Es una medida de dispersión de la diferencia entre los valores del percentil 90 y el percentil 10.

RANGO INTERPERCENTÌLICO (RIP)
$RIP = P_{90} - P_{10}$

3.3.2 Medidas de desviación promedio

Las descripciones más completas de la dispersión son aquellas que manejan la desviación promedio respecto a alguna medida de tendencia central. En esta clasificación las más utilizadas son la varianza y la desviación estándar. Ambas medidas dan una distancia promedio de cualquier observación del conjunto de datos respecto a la media de la distribución.

3.3.2.1 Varianza

Medida del cuadrado de la distancia promedio entre la media y cada observación de la población.

VARIANZA DE POBLACIÓN (σ^2)
$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
VARIANZA DE DATOS AGRUPADOS DE POBLACIÓN (σ^2)
$\sigma^2 = \frac{\sum f * (xi-\mu)^2}{N}$

VARIANZA DE UNA MUESTRA (S^2)
$$S^2 = \frac{\sum (x - \overline{X})^2}{(n-1)}$$

VARIANZA DE DATOS AGRUPADOS DE MUESTRA (S^2)
$$S^2 = \frac{\sum f * (xi - \overline{X})^2}{(n-1)}$$

3.3.2.2. Desviación Estándar

Se calcula obteniendo la raíz cuadrada positiva de la varianza. Esta medida de dispersión tiene las mismas unidades que los datos originales, a diferencia de la varianza en la que las unidades están expresadas por los cuadrados de las unidades.

3.3.3 Dispersión relativa

La desviación estándar es una medida absoluta de la dispersión que expresa la variación en las mismas unidades que los datos originales, el coeficiente de variación es una medida relativa de dispersión que relaciona la desviación estándar y la media, expresando la desviación estándar como porcentaje de la media, la unidad de media es entonces "porcentaje", en lugar de las unidades de los datos originales. El interés del coeficiente de variación es que al ser un porcentaje permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desviación típica, ya que viene expresada en las mismas unidas que los datos de la serie.

COEFICIENTE DE VARIACIÓN (CV)
$$CV = \frac{\sigma}{\mu} * 100$$

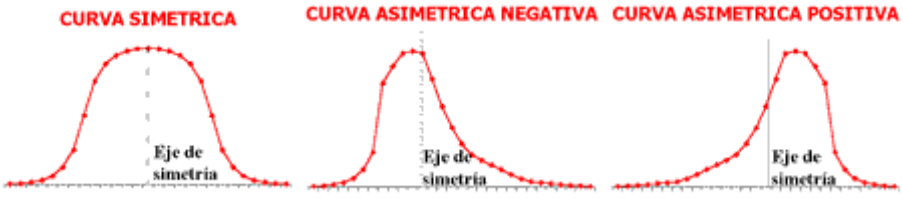
DÍA 8

3.4 Medidas de forma

La forma es la manera en que los datos se distribuyen, es decir la forma que tiene la curva que representa la serie de datos muestrales. La forma se mide en dos aspectos: Sesgo o Asimetría y Curtosis o Apuntamiento.

3.4.1 Sesgo (Asimetría)

Mide si la curva de la gráfica que representa a los datos es simétrica respecto al eje vertical, si lo es se dice que la hay simetría (distribución Simétrica o Insesgada) y si no lo es se dice que es Asimétrica o Sesgada.



Existen varias formas de calcular el sesgo de una distribución, a continuación se presentan dos de ellas:

COEFICIENTES DE SESGO DE PEARSON (SK_1 Y SK_2)
$Sk_1 = \frac{\overline{X} - Mo}{S}$ $\qquad$ $Sk_2 = \frac{3(\overline{X} - Me)}{S}$

SESGO EN FUNCIÓN DE CUANTILES O FRACTILOS	
Sesgo cuartìlico	**Sesgo percentìlico**
$Sk_3 = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$	$Sk_4 = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$

El signo en los coeficientes de sesgo determina la asimetría:

El signo positivo corresponde a una distribución asimétrica positiva
El signo negativo corresponde a una distribución asimétrica negativa

Asimismo, cuando el coeficiente de sesgo es igual a 0, indica que la distribución es simétrica.

3.4.2 Curtosis (apuntamiento)

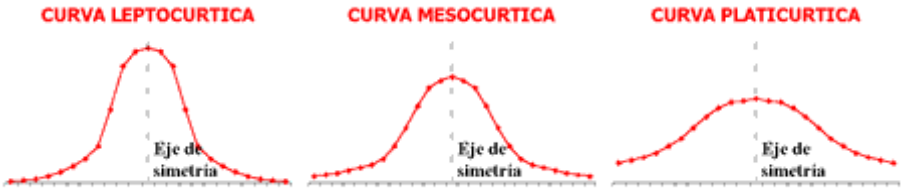
Mide la altura o grado de apuntamiento de la gráfica que representa a los datos (eje horizontal).

Se definen 3 tipos de distribuciones según su grado de curtosis:

Distribución mesocúrtica: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).

Distribución leptocúrtica: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

Distribución platicúrtica: presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



El **Coeficiente de Curtosis** analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución, se calcula a través de la siguiente fórmula:

COEFICIENTE DE CURTOSIS CENTILICO
$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}}$

Si

K > 0.263 la distribución es Leptocúrtica
K = 0.263 la distribución es Mesocúrtica
K < 0.263 la distribución es Platicúrtica

DÍA 9

4. Presentación y análisis de datos de dos variables

4.1 Tablas de contingencia.

En determinadas ocasiones el encargado del análisis estadístico clasifica una unidad experimental de acuerdo con dos variables cualitativas, con lo que genera datos bivariados.

Cuando se registran dos variables categóricas se puede resumir la información contando el número de observaciones que caen en cada una de las distintas intersecciones de los niveles de categoría. Los valores resultantes se presentan en un arreglo llamado **tabla de contingencia r x c.**

	Columnas				
Renglones		**1**	**2**	**. . .**	**c**
	1	O_{11}	O_{12}	. . .	O_{1c}
	2	O_{21}	O_{22}	. . .	O_{2c}
	. **.** **.**	. . .	. . .	. . .	. . .
	r	O_{r1}	O_{r2}	. . .	O_{rc}