

FD
01040
M-5

DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADISTICA
DANE

SISTEMA ESTADISTICO NACIONAL Y TERRITORIAL - SENT

DIVISION DE CALIDAD E INTERVENTORIA ESTADISTICA

ESTADISTICA DESCRIPTIVA

Santafé de Bogotá, febrero de 1998

Créditos:

Calidad Estadística: Myriam Raquel Cifuentes N.
Luis Alejandro Montenegro R.

División de Ediciones: Eduardo Alvarado Núñez

Carátula: Mario Guevara M.

Corrección: Hilda Baquero, Miguel Acosta.

Fotomecánica: Luis Alirio Rodríguez.

Impresión: Juan Carlos Sánchez.

INTRODUCCION

¿Qué es la estadística? ¿Cómo funciona y de qué manera se involucra en nuestro contexto laboral?.

El DANE, como rector de la producción estadística del país, se ve abocado a resolver problemas relacionados con todos los aspectos que involucra el desarrollo de una encuesta (investigación), desde su planificación, diseño y ejecución hasta la producción y análisis de resultados; por esto, se hace necesario que todo funcionario que esté directa o indirectamente relacionado con los procesos de producción de información, tenga conocimiento sobre algunos métodos y procedimientos útiles en el desarrollo de la actividad estadística y es hacia allá que se enfoca el presente documento.

Antes de definirla, es conveniente examinar algunos problemas que pudieran presentarse e ilustrar a través de ellos los aspectos estadísticos esenciales.

Tómese por ejemplo, un comercio pequeño en el que cada día se debe decidir la cantidad de unidades de determinado artículo de venta perecedero que hay que ordenar. La decisión puede basarse en las ventas observadas de este artículo en el pasado, así como en los costos en que se incurre por quedar con un remanente almacenado al final del día, o no satisfacer completamente la demanda por haberse agotado el artículo en mención.

Considérese por otro lado, los planes de inspección para los artículos comprados por una planta manufacturera. Sobre la base de una inspección, cada lote de artículos recibido debe ser aceptado, si se considera que satisface los requerimientos establecidos por el contrato, o rechazado y devuelto al proveedor. El plan posiblemente consista en la inspección de solo 10% de artículos de cada lote, observando cuantos de estos están defectuosos por no satisfacer los mencionados requerimientos. La decisión, entonces se basaría en el número de artículos defectuosos de una muestra del lote.

Supóngase de igual manera, que se desea obtener información acerca del sector agrícola, con el fin de que los encargados de la planeación y formulación de políticas tengan una guía en la toma de decisiones; para ello se realiza una encuesta. La metodología del estudio debe ser científica, las definiciones deben ser claras y no ambiguas, los procedimientos eficientes y libres de error, el personal debe ser experto o capacitado.

Dadas las restricciones interrelacionadas de personal, presupuesto y tiempo, el diseño de dicha encuesta se realiza por muestreo, esto es, el estudio no involucrará una completa enumeración de la población. El planteamiento anterior exige el uso adecuado de métodos estadísticos, que tienen que ver con la planificación de la encuesta, el diseño del cuestionario, el diseño del plan de muestreo, diseño del método de control y aseguramiento de la calidad de la información recopilada, calidad de las estimaciones realizadas, etc.

Este documento, resume brevemente, algunos métodos utilizados en la realización de análisis descriptivo y puede servir de guía para quienes no poseen conocimiento acerca del tema.

CONTENIDO

	Pág. No.
Introducción	3
Generalidades	7
Distribuciones unidimensionales de frecuencias	15
Medidas descriptivas numéricas	18
Medidas de localización	19
Medidas de variabilidad	25
Dispersión relativa	27
Ejercicios de aplicación	30
Ejercicios propuestos	39
Bibliografía	42

GENERALIDADES

Estadística: es la ciencia que crea, desarrolla y aplica métodos que ayuden a tomar decisiones frente a la incertidumbre.

La estadística, es un campo del conocimiento humano que permite al investigador deducir y evaluar conclusiones acerca de una población a partir de información proporcionada por una muestra. Más específicamente, la estadística trata de métodos y técnicas que pueden usarse en:

1. Recolección y clasificación de datos
2. Interpretación y análisis de datos
3. Deducción y evaluación de conclusiones y de su confiabilidad con base en datos muestrales.

Inferencia estadística: es un proceso de inducción lógica que partiendo de los datos de una parte establece un juicio sobre todo el conjunto, obteniéndose una medida de la incertidumbre para la consecuencia que se infiere.

Población: es un conjunto completo de individuos, objetos que poseen alguna característica común observable. Así, todos los establecimientos industriales de Colombia, constituyen una población; cuando interese conocer por ejemplo, la tendencia y estructura de la actividad industrial colombiana.

Los *datos individuales* de una población se llaman *unidades elementales*. Definir una población es, en un sentido, limitar el contenido de las unidades elementales. Estas poseen ciertas características, conocidas a veces como rasgos o propiedades, que pueden ser de naturaleza *cualitativa o cuantitativa*. El resultado de observar una unidad elemental se llama observación. Puesto que, en una investigación, se hace una observación de cada unidad elemental escogida, podemos considerar una población como la totalidad de todas las observaciones pertinentes que podrían hacerse en un problema de decisiones dado.

La definición de una población y las características de sus unidades depende de la naturaleza del problema de decisiones que se plantea.

Por lo tanto la población está definida por:

- Unidades que la constituyen (personas, hogares, establecimientos, etc.)
- Características que la definen (es decir las características que permiten identificarla).
- Localización geográfica (es decir en dónde están ubicadas)
- Período de referencia o tiempo bajo consideración (últimos meses, a partir de...)

Es preciso distinguir entre población objeto y población sujeto; la primera, es aquella para la que se desea obtener la información, la segunda, es la población encuestada - es la población realmente representada en la encuesta.

Ejemplo:

Encuesta sobre gastos de los hogares

Población objeto: residentes de Cundinamarca

Población sujeto: residentes de Cundinamarca excluyendo: población institucionalizada, personas sin lugar de residencia fijo.

Encuesta sobre los graduados:

Población objeto: personas que cumplieron con todos los requisitos para graduarse de un programa universitario, de enfermería u otro programa postsecundario durante 1996, *excluyendo* seminarios religiosos, escuelas militares y colegios empresariales.

Población encuestada: la misma.

En la realidad, la población objeto y la población encuestada pueden ser bastante diferentes. En muchos casos, se redefine la población objeto para que sea la población realmente encuestada; para hacerlo, hay que reconocer las diferencias.

Las unidades estadísticas se clasifican como:

Unidad de muestreo¹: unidad seleccionada o a seleccionar

Unidad de análisis: unidad hacia la cual se hacen las inferencias

Unidad informante: unidad que suministra la información

Unidad de referencia: unidad acerca de la cual se obtiene la información

Ejemplos:

- Se desea conocer información trimestral sobre mercado laboral de personas mayores de 15 años.

Unidad de muestreo: hogares

Unidad de análisis: personas económicamente activas

Unidad informante: personas adultas (>15 años) pertenecientes al hogar

Unidad de referencia: personas dentro del hogar.

- Se desea conocer la percepción que tiene la población de Santafé de Bogotá con edades entre 15 y 45 años en materia de educación superior

Unidad de muestreo: hogares

Unidad de análisis: personas con edades entre 15 y 45 años

¹ Las unidades de muestreo contienen los elementos y se usan para seleccionarlos en la muestra. En el muestreo de elementos, cada unidad de muestreo contiene solamente un elemento; pero en el muestreo de conglomerados cualquier unidad de muestreo llamada conglomerado puede contener varios elementos. Kish, pág. 28

Unidad informante: personas con edades entre 15 y 45 años

Unidad de referencia: personas con edades entre 15 y 45 años

Cuando las características son cualitativas, las unidades elementales pueden ser clasificadas sólo como poseedoras o no poseedoras de cierta cualidad o propiedad. Las observaciones cualitativas de unidades elementales, llamadas variables categóricas, a menudo pueden ser expresadas numéricamente. Esto se hace contando las unidades elementales que tienen cierta propiedad. En el proceso de enumerar, se asigna 1 a un elemento que tiene el atributo, y 0 al que carece de éste. Al terminar la cuenta se obtiene el total sumando los unos teniéndose entonces un resultado numérico. Esta expresión cuantitativa de datos cualitativos abre al tratamiento estadístico muchas áreas inalcanzables por medición directa.

Al medir las unidades elementales se debe definir cuidadosamente la unidad elemental y su característica observada. Aunque en muchos casos es sencillo y fácil identificar la unidad elemental en la población, en otros resulta muy difícil, a menos que se defina con precisión la unidad elemental. Por ejemplo, en los problemas de toma de decisiones referentes a situación de una vivienda, a menudo se necesita conocer el número de habitaciones en una zona determinada. Pero, ¿qué es una habitación? Un cuarto de baño, una cocina, una terraza o un salón de entrada, ¿se puede considerar como tal? También se produce confusión cuando se carece de una definición precisa de la característica. La ilustración más obvia es la edad. ¿Cómo registrar la edad de una persona? ¿En su último cumpleaños? ¿En su próximo cumpleaños?. La respuesta depende de la forma como se haya definido la unidad elemental.

Se evita usar la palabra universo como sinónimo de población. El universo es un conjunto hipotético infinito de elementos generado por un modelo teórico, ejemplo: lanzar al aire una moneda un número infinito de veces.

Parámetro: es un valor (o constante) que caracteriza a una población; es cualquier característica de una población que sea medible, por ejemplo la media μ , la desviación estándar σ , la proporción P , etc.

Muestra: es un subconjunto de la población o universo. La muestra, cuando es probabilística, permite realizar inferencia estadística de la población de la encuesta.

Estadístico: es el número resultante de la manipulación de ciertos datos iniciales de acuerdo con determinados procedimientos específicos; comúnmente, usamos un estadístico que se calcula a partir de una muestra para estimar el parámetro de una población. Se representan con letras comunes: $\bar{x}$, s .

Encuesta: es la recolección de información acerca de características de interés de algunas o de todas las unidades de la población, usando conceptos bien definidos, métodos, procedimientos, y la compilación de tal información en formatos de resumen útiles.

Las encuestas se realizan con diferentes propósitos. Cuando el propósito principal es estimar ciertas características o atributos de la población, se habla de una encuesta *descriptiva*; cuando el principal propósito es la verificación de hipótesis estadísticas o la exploración de relaciones entre las características con un enfoque más explicativo que descriptivo, se dice que la encuesta es *analítica*.

Variable: característica o fenómeno que puede tomar diferentes valores. ej: peso, estatura, número de hijos de las familias de un barrio, etc. Generalmente se representa con las letras X, Y.

Los valores que toma la variable se notan con minúscula; ejemplo:

X: edad de 15 personas (años)

$x_1 = 20$	$x_2 = 25$	$x_3 = 27$	$x_4 = 19$	$x_5 = 32$	$x_6 = 30$	$x_7 = 32$	$x_8 = 23$
$x_9 = 22$	$x_{10} = 34$	$x_{11} = 21$	$x_{12} = 26$	$x_{13} = 34$	$x_{14} = 30$	$x_{15} = 23$	

$x_1 = 20$ (la edad de la primera persona observada, es 20 años)

El total, se calcula como la suma de las edades de las 15 personas:

$$T = 20 + 25 + 27 + 19 + \dots + 23$$

De manera más general,

$$T = x_1 + x_2 + x_3 + \dots + x_{15}$$

o, en notación abreviada,

$$T = \sum_{i=1}^{15} x_i$$

La letra griega sigma Σ indica la suma de los valores.

Recapitulando:

Población es un conjunto completo de individuos, objetos, que poseen una característica común observable. Todos los ciudadanos en edad de votar de un país, por ejemplo, constituyen una población.

Son ejemplos de poblaciones estadísticas:

- Supóngase que se desea determinar el ingreso familiar anual medio de una comunidad de 10.000 familias. La población consta entonces de 10.000 números (\$5.000.000, \$6.500.000, 9.000.000,.....) cada uno de los cuales representa el ingreso anual de una determinada familia. La unidad elemental es una familia en la comunidad y la característica a medir, es su ingreso anual.

- Se desea determinar la edad media de los estudiantes de una universidad, cuyo total de alumnos matriculados es de 15.000. La población consta de 15000 números (18, 23,..) cada uno de los cuales indica la edad de un estudiante. La unidad elemental es un estudiante de dicha universidad y la característica a medir, es su edad.

- Se pretende determinar el lugar de nacimiento de los empleados de una empresa que consta de 9.000 empleados. La población consiste de 9.000 términos (Santafé de Bogotá, Cali, Ibagué,...) cada uno de los cuales describe un lugar de nacimiento.

Una población estadística es entonces un conjunto de observaciones medidas o descritas, para cada una de sus unidades elementales. Las observaciones medibles se denominan observaciones cuantitativas (ingresos, edad, número de hijos). Las observaciones no medibles se denominan cualitativas (estado civil de una persona, marca de un automóvil).

Los valores asumidos por las observaciones cuantitativas se llaman valores de la variable.

Las variables pueden clasificarse como discretas o continuas. Una variable discreta puede tomar solo un número limitado de valores en una escala de medida. De otra manera, puede tomar solo valores enteros ej: número de personas, tamaño de una familia, etc.

Una variable continua es aquella que puede tomar un número infinito de valores en una escala de medidas. Ejemplo: edad, peso, etc.

Las observaciones cualitativas no pueden ser medidas; sólo pueden ser descritas. Estas se llaman variables categóricas (estado civil: soltero, casado, viudo o divorciado).

Es importante enfatizar la diferencia entre una población de variables cuantitativas y una población de variables categóricas. En las primeras por ser medibles, puede calcularse por ejemplo, su media: ingreso familiar medio; en el caso de las segundas, se pueden calcular razones o porcentajes, ejemplo: porcentaje de artículos defectuosos en un determinado lote, etc. .

Debe destacarse que si bien la decisión de un problema dado puede depender, por ejemplo, del valor de la media aritmética de una población, este valor rara vez se calcula utilizando la información acerca de toda la población, sino que habitualmente se estima a partir de una muestra seleccionada adecuadamente.

En muchas investigaciones, la población estadística es infinita y, por lo tanto resulta imposible tener información completa sobre ella; por lo demás, aun cuando esto fuera posible, el gasto y el tiempo que habría que invertir en recolectar y organizar la información pueden resultar muy altos. En tales casos, debe utilizarse la información proveniente de una muestra. Se ve así que el problema de decisión acerca de una población a partir de la información proporcionada por una muestra resulta ser un problema fundamental en estadística.

El diseño estadístico de una investigación: consiste en identificar o determinar el conjunto de técnicas estadísticas que permitan la observación y análisis de la población objeto de estudio. Entre estos tenemos:

- *Censo:* procedimiento mediante el cual se estudia la totalidad de la población objetivo. Algunos ejemplos de encuestas censales: censo de población y vivienda, censo minero nacional, censo de sacrificio de ganado, censo económico.
- *Muestreo:* aplicación de una serie de técnicas que permiten seleccionar una parte de los elementos de una población para someterlos a estudio.

Al realizar una investigación por muestreo es preciso escoger el tipo de muestreo a utilizar teniendo en cuenta aspectos como el nivel de confianza y el error máximo permisible entre otros y de manera análoga determinar el tamaño óptimo de muestra acorde con los objetivos determinados previamente. Algunos ejemplos de encuestas por muestreo: encuesta nacional de hogares, encuesta nacional de calidad de vida, encuesta nacional agropecuaria, muestra mensual manufacturera, encuesta de ingresos y gastos.

- *Registros administrativos:* datos recogidos con un fin administrativo y que no pueden servir para análisis estadísticos generales sino para aplicaciones particulares. Algunas fuentes de datos administrativos: estadísticas vitales (nacimientos y defunciones), salud (admisión en hospitales, enfermedades, costos), educación (estudiantes matriculados, costos), justicia (crímenes reportados, costos o montos), importaciones y exportaciones.

Recuerde que el objetivo de la estadística es hacer inferencias acerca de un conjunto de elementos o unidades (población) con base en la información contenida en una muestra.

Los métodos usados para describir conjuntos de datos numéricos pueden ser clasificados en dos tipos: gráficos y numéricos. A continuación se presentan algunos de ellos.

DISTRIBUCIONES UNIDIMENSIONALES DE FRECUENCIAS

La información estadística puede constar de un gran número de observaciones y, mientras mayor sea su número, mayor puede ser la conveniencia y necesidad de presentarla en forma resumida, la cual puede omitir algunos detalles pero en cambio puede revelar la naturaleza general de la información. Un resumen de tal información se denomina distribución de frecuencias. Ejemplos de ellas son:

**Establecimientos económicos de la ciudad de Páramo,
según actividad**

Actividad	No. de establecimientos
Industria	1.200
Comercio	30.500
Servicios	11.895
TOTAL	43.595

**Años de estudio terminados y aprobados por un grupo de personas
residentes en el barrio Galán, de Santafé de Bogotá**

Años	No. de personas
3	18
4	20
7	12
8	21
9	14
TOTAL	85

**Fuerza laboral masculina de una ciudad,
clasificada por edades**

Edad (años)	No. de varones
14 - 19	218.000
20 - 24	313.000
25 - 55	977.000
> de 55	255.000
TOTAL	1.763.000

Los grupos o categorías que van de 14 - 19, 20 - 24, etc. . se llaman intervalos de clase. Los valores 14, 20,...etc. se llaman límites inferiores y 19, 24,... etc. son los límites superiores de los intervalos de clase.

Principios para la construcción de una distribución de frecuencias

Para variables continuas, deben observarse los siguientes aspectos relacionados con la construcción de los intervalos:

1. El número de intervalos depende del número de datos y de su dispersión. Si el número de intervalos es demasiado pequeño pueden estarse ocultando características importantes de los datos en esta agrupación, si se tienen demasiados intervalos se

pueden presentar muchos intervalos de clase vacíos que resten significado a la distribución. Entonces el número de intervalos de clase será determinado por el número de datos considerando también qué tan uniformes sean los datos. Una muestra pequeña requiere en general pocos intervalos de clase.

2. Como regla general para encontrar la longitud de los intervalos debe dividirse la diferencia de la mayor y la menor de las observaciones entre el número de intervalos de clase, modificando este cociente para tener intervalos de clase de longitud conveniente, cercana al cociente obtenido. Todas las clases deben tener la misma longitud, con posible excepción del primer y último intervalo; esto con el fin de permitir uniformidad en las comparaciones de frecuencias de clase.
3. Se debe comenzar por el primer intervalo, estando seguro de que incluye a la mínima de las observaciones. Se añaden los demás sumando la amplitud escogida.
4. El punto medio (marca de clase) de un intervalo de clase, se obtiene sumando los límites inferior y superior de la clase y dividiendo por dos.
5. La tabla de frecuencias indica la distribución de las frecuencias de los valores de la variable X dentro de varias clases. El término distribución de frecuencias se abrevia normalmente como distribución; por ejemplo, se dice distribución de consumidores por niveles de ingreso, distribución de calificaciones, etc.

REPRESENTACION GRAFICA

La representación gráfica de una tabla de frecuencias da una información concisa y clara acerca de una distribución de frecuencias. Veremos dos tipos de representaciones gráficas: histogramas y polígonos de frecuencias.

HISTOGRAMA

Es un gráfico de barras, que se construye llevando sobre las abscisas los límites reales de clase para el caso continuo, o los valores de la variable en el caso discreto y sobre las ordenadas las frecuencias observadas, para ambos casos. Se emplea especialmente cuando se desea observar el comportamiento de la variable (identificando puntos de mayor o menor frecuencia) o cuando se desea comparar dos variables. Ver ejemplo, pág. 33

POLIGONO DE FRECUENCIAS

Es un gráfico de línea, que se construye llevando sobre el eje de las abscisas las marcas de clase y sobre las ordenadas las frecuencias observadas. Uniendo los puntos resultantes y cerrándolo en la marca de clase del intervalo anterior al primero y posterior al último. Se utiliza preferencialmente cuando se desea comparar dos o más variables.

Un caso particular, es el polígono de frecuencias acumuladas u ojiva, en el cual se lleva sobre la abscisas los límites reales de clase y sobre las ordenadas las frecuencias acumuladas (N_i o H_i). Ver ejemplo, pág. 33

MEDIDAS DESCRIPTIVAS NUMERICAS

Los métodos gráficos son útiles para lograr una descripción de los datos que sea rápidamente captada por el lector. Sin embargo, las técnicas gráficas presentan limitaciones para la descripción y análisis de conjuntos de datos.

Las limitaciones pueden superarse con el uso de medidas descriptivas numéricas. Con base en los datos muestrales es posible calcular un conjunto de números que proporcionen una buena imagen de la distribución de frecuencias de la población y que resulte útil para hacer inferencias acerca de las características de la población.

Las medidas descriptivas numéricas calculadas a partir del total de observaciones de la población se denominan como ya se dijo anteriormente *parámetros* y aquellas calculadas a partir de una muestra se llaman *estadísticos*.

MEDIDAS DE LOCALIZACION (POSICION)

Objetivo: establecer unas medidas que caractericen lo mejor posible una serie de datos. Pueden ser de posición central o no; las medidas de posición central son: la media aritmética, la mediana y la moda. Otras medidas de posición son los percentiles.

Usos: en muchos problemas estadísticos la elección entre media, mediana y moda puede no ser obvia; por lo tanto, pueden ser de utilidad algunos comentarios generales acerca de las ventajas y desventajas de cada una de las tres medidas.

Primero, debe destacarse que la *media aritmética* es la medida más conocida y mejor comprendida por el público en general. Segundo, la media aritmética permite en general, en forma mucho más expedita, el análisis estadístico posterior; además, se puede obtener aún en el caso de no conocer los valores individuales de la serie, por ejemplo: si 10 familias consumen 20 litros de leche, el promedio aritmético será de 2 litros por familia. Por estas razones, puede considerarse la media aritmética o el promedio la medida más comúnmente usada

Sin embargo, la media aritmética tiene la desventaja de ser muy afectada por valores extremos, lo cual no ocurre para la mediana y la moda. Por lo tanto, cada vez que existen valores extremos, la mediana y la moda son preferidas a la media aritmética. Ingreso familiar en una ciudad, salarios por hora en una industria, ventas al detal en los almacenes de una ciudad, son ejemplos de este caso.

La mediana: es una medida de localización adecuada cuando se tiene una distribución de frecuencias asimétrica y que presenta valores extremos. Hay situaciones en que la única medida de posición central que puede calcularse es la mediana, tal como sucede en el caso de una distribución cuyos intervalos extremos no están definidos.

La moda: se caracteriza por la facilidad de su cálculo y su aplicación tanto a datos cuantitativos como a los cualitativos por ejemplo, se desea determinar la talla de camisa más usada por los hombres, la preferencia de las amas de casa por un determinado tipo de jabón, el número de pasajeros que usan el bus ejecutivo, etc. . Puede encontrarse que existen dos modas, una por la mañana alrededor de las ocho y otra por la tarde alrededor de las 7. La distribución es pues, bimodal.

Tanto la media como la mediana tienen un interés especial en la representación de una distribución. El interés radica en que caracterizan a la distribución. La moda, como se ha visto en los ejemplos anteriores, tiene interés por sí misma por ser el valor más típico o predominante, de una distribución.

Las ecuaciones comúnmente usadas para el cálculo de las medidas de localización son las siguientes:

- Para datos brutos

Media aritmética:
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Mediana (Me): n impar es el valor central de la variable ordenada.
 n par es el promedio de los valores centrales de la variable ordenada.

Moda: (Mo) Valor que presenta la mayor frecuencia.

- Para datos agrupados

Media aritmética $\bar{X} = \frac{\sum_{i=1}^n y_i \cdot n_i}{n}$, tanto para variable discreta o continua.

Mediana (Me): variable discreta

Caso a cuando $N_{j-1} < n/2$ $Me = Y_j$

Caso b cuando $N_{j-1} = n/2$ $Me = \frac{Y_{j-1} + Y_j}{2}$

Variable continua

Caso a cuando $N_{j-1} = n/2$ la mediana se obtiene mediante la fórmula: $Me = Y'_{j-1}$

Caso b cuando $N_{j-1} < n/2$ la mediana se obtiene mediante la fórmula

$$Me = Y'_{j-1} + C \frac{n/2 - N_{j-1}}{n_j}$$

Los pasos a seguir para el cálculo de la mediana se pueden resumir, así:

1. Obtener las frecuencias absolutas acumuladas.
2. Buscar la mitad de las observaciones, por medio de $n/2$
3. Localizar el resultado anterior $n/2$ en la columna de las frecuencias absolutas acumuladas; si no aparece, al valor inmediatamente anterior se le denomina N_{j-1} y al inmediatamente superior se le denomina N_j
4. Si N_{j-1} es menor que $n/2$, se dice entonces que la mediana es igual a Y_j

Referirse a las ecuaciones anteriores a o b según el caso

Moda Mo En el caso de variable continua corresponde a la marca de clase del intervalo que presenta la mayor frecuencia; en el caso discreto simplemente es el valor de la variable que presenta la frecuencia más alta.

Otra medida de localización es la *media aritmética ponderada*.

Se utiliza cuando se asocia a los números $x_1, x_2, \dots, x_k$ ciertos factores o pesos $w_1, w_2, \dots, w_k$ que dependen de la significación o importancia de cada uno de los números. En este caso:

$$\overline{X} = \frac{w_1 X_1 + w_2 X_2 + \dots + w_k X_k}{w_1 + w_2 + \dots + w_k}$$

Ejemplo:

Si un examen final de curso se valora como tres veces los exámenes parciales y un estudiante tiene una nota de examen final de 8 y notas de exámenes parciales de 7 y 9, su nota final será:

$$\overline{X} = \frac{1(7) + 1(9) + 3(8)}{1 + 1 + 3} = 8$$

Propiedades de la media aritmética

- La suma algebraica de las desviaciones de un conjunto de números de su media aritmética es cero.

Ejemplo:

X_i	$(X_i - \overline{X}_i)$
8	0.4
3	-4.6
5	-2.6
12	4.4
10	2.4
Σ	0

$$\bar{X} = 38/5 = 7.6$$

- Si f_1 números tienen media m_1 , f_2 tienen media m_2 , f_3 números tienen media $m_3, \dots, f_k$ tienen media m_k entonces la media de todos los números es

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \dots + f_k m_k}{f_1 + f_2 + \dots + f_k}$$

Es decir, una media aritmética ponderada de todas las medias.

Ejercicio

En una compañía con 80 empleados, 60 ganan 9.000 pesos por hora y 20 ganan 11.000 por hora.

- Determinar la media del salario por hora
- ¿ Sería la respuesta de a) la misma si los 60 empleados ganaran un salario *medio* por hora de 9.000 pesos y los 20 empleados ganaran un salario *medio* por hora de 11.000? Probar la respuesta
- ¿ Es el salario medio por hora representativo?

NOTA: en general, la media de una muestra es igual a la media ponderada de las submuestras, tomándose como ponderación los tamaños de esas submuestras

$$n=20 \quad n_1=12 \quad \bar{Y}_1=1.333$$

$$\bar{Y}=2.15 \quad n_2=8 \quad \bar{Y}_2=3.375$$

$$\bar{X} = \frac{\bar{X}_1 n_1 + \bar{X}_2 n_2}{n} \quad \text{o} \quad \bar{Y} = \frac{\bar{Y}_1 n_1 + \bar{Y}_2 n_2}{n}$$

Muestra

Y_i	n_i	$Y_i n_i$
0	2	0
1	4	4
2	6	12
3	5	15
4	3	12
Σ	20	43

Submuestra 1

Y_i	n_i	$Y_i n_i$
0	2	0
1	4	4
2	6	12
Σ	12	16

Submuestra 2

Y_i	n_i	$Y_i n_i$
3	5	15
4	3	12
Σ	8	27

$$\bar{Y} = 43/20 = 2.15$$

$$\bar{Y}_1 = 16/12 = 1.333$$

$$\bar{Y}_2 = 27/8 = 3.375$$

La media aritmética de las dos submuestras en conjunto será igual:

$$\bar{Y} = \frac{\bar{Y}_1 n_1 + \bar{Y}_2 n_2}{n_1 + n_2}$$

$$\bar{Y} = \frac{(1.333) + 3.375(8)}{20} = 2.15$$

- La media aritmética de una constante por una variable es igual al producto de la constante por la media aritmética de la variable

$$M(KY) = k \bar{Y}$$

- La media aritmética de una constante más una variable, es igual a la constante más la media aritmética de la variable

$$M(Y_i + K) = \bar{Y} + K$$

- La suma de los cuadrados de las desviaciones, es mínima cuando dichas desviaciones se tomen respecto a la media aritmética.

- La media de la suma de dos variables, es igual a la suma de las medias de dichas variables ; esta propiedad puede extenderse a las diferencias entre variables

$$M(X_i + Y_i) = M(X_i) + M(Y_i) = \bar{X} + \bar{Y}$$

Observaciones

En una variable discreta el resultado obtenido al aplicar la media aritmética debe ser el mismo, cualquiera que sea el método, tanto para datos agrupados como para no agrupados

En la variable continua, el resultado obtenido al aplicar la fórmula para datos no agrupados difiere del resultado obtenido aplicando la fórmula para datos agrupados, debido a la pérdida de información primero por agruparse los datos en intervalos y segundo por calcularse la media utilizando las marcas de clase.

MEDIDAS DE VARIABILIDAD (DISPERSION)

Objetivo: medir la distancia promedio entre los valores de la variable y su media aritmética. Son utilizados para indicar el grado de uniformidad de los datos.

La dispersión puede ser medida mediante el recorrido o rango, la desviación media, el recorrido intercuartílico, la varianza y la desviación estándar. Trataremos únicamente la desviación estándar.

La desviación estándar, en un sentido amplio, mide la desviación promedio de cada valor de la variable respecto de la media aritmética. Esto es, la desviación estándar es la raíz cuadrada del promedio de los cuadrados de las diferencias entre los valores de la variable y la media aritmética. Tiene como ventaja que permite un mejor desarrollo para un análisis estadístico posterior.

La razón por la que se prefiere la desviación típica a otras medidas radica en sus propiedades matemáticas. A continuación se define el término desviación, después el concepto de varianza y, por último, la desviación típica o estándar.

Por desviación se entiende la diferencia entre un valor individual x_i y la media ($x_i - \bar{x}$).

La varianza de una población de N observaciones se define como el promedio del cuadrado de las desviaciones con respecto a su media μ . La varianza de la población se denota por σ^2 y está dada por la fórmula :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

La varianza de una muestra de n observaciones se define como la suma de los cuadrados de las desviaciones de las observaciones respecto de su media, $\bar{X}$ dividida esta suma entre $(n-1)$. La varianza de la muestra se denota por S^2 y está dada por la fórmula.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

La desviación estándar de un conjunto de n observaciones $x_1, x_2, x_3, \dots, x_n$, es igual a la raíz cuadrada positiva de la varianza.

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

La varianza se mide en términos del cuadrado de las unidades originales. Si las observaciones están medidas en centímetros la varianza está dada en centímetros cuadrados. Al tomar la raíz cuadrada de la varianza, se obtiene la desviación estándar, con lo que se regresa a las unidades originales de las observaciones.

INTERPRETACION DE LA DESVIACION ESTANDAR

Una buena comprensión del significado de la desviación estándar depende del conocimiento de la relación existente entre la desviación estándar y la distribución normal²; ya que cuando los datos obedecen a esta distribución, tomando unidades de desviación típica a lo largo de eje horizontal con origen en la media, resulta la superficie encerrada por la curva, dividida en zonas de área bien determinada. Este tema a pesar de su importancia no se profundizará por no estar dentro de los objetivos del presente escrito.

DISPERSION RELATIVA

Se utiliza cuando se quiere comparar la variabilidad entre dos o más conjuntos de datos. Esto puede hacerse fácilmente con sus respectivas varianzas o desviaciones estándar cuando las variables se dan en las mismas unidades y cuando sus medias son aproximadamente iguales. Cuando faltan estas condiciones, puede usarse alguna medida relativa de dispersión. Una medida relativa de variabilidad frecuentemente usada es el coeficiente de variación, CV, que es simplemente la razón de la desviación estándar a la media y está generalmente expresado como un porcentaje.

El coeficiente de variación es independiente de las unidades utilizadas, por esta razón es útil para comparar distribuciones donde las unidades pueden ser diferentes.

$$CV = \frac{S}{\bar{x}}$$

Ejemplo: considérese que un grupo de estudiantes realiza dos pruebas. La primera resulta con una media de 60 puntos, una desviación típica de 6 puntos y un máximo de 100 puntos. La segunda tiene una media de 700 puntos y una desviación típica de 7 puntos con un máximo de 1.000 puntos. ¿Cuál de las dos pruebas tiene una variación mayor?

² Distribución de probabilidad continua: ver Mendenhall, pág 161.

Prueba	Media	Desviación típica	N
A	60	6	100
B	700	7	1.000

Desde un punto de vista absoluto, se observa que los 7 puntos corresponden a una variación mayor que la de 6 puntos, pero desde un punto de vista relativo, mediante el

$$CV = \frac{S}{\bar{X}} \text{ tenemos:}$$

Primera prueba, $CV = 6 / 60 = 1/10$

Segunda prueba, $CV = 7/700 = 1/100$

Como se ve, la dispersión relativa de la segunda prueba es 1/10 de la primera.

Otra aplicación del CV es su empleo para comprobar resultados experimentales y para la estimación de la desviación típica.

Supongamos que un científico ha obtenido los siguientes datos referentes a los pesos de elefantes y ratas:

Animales	Media ($\bar{X}$) libras	Desviación típica (s) libras
Elefantes	24.000	1.285
Ratas	1.05	0.16

Claramente se observa que los pesos de los elefantes varían mucho más que los pesos de las ratas. Sin embargo, los pesos de los elefantes pueden variar menos como un porcentaje de su propia media que los pesos de las ratas. Si calculamos sus respectivos coeficientes de variación tenemos:

$$CV (ELE) = 1285 / 24000 = 0.054 \quad (5.4\%)$$

$$CV (RAT) = 0.16 / 1.05 = 0.152 \quad (15.2\%)$$

Así, vemos que la variabilidad de los pesos de las ratas es casi tres veces mayor, sobre una base relativa, que la variabilidad de los pesos de los elefantes. La desviación estándar de los primeros es 15.2% de su media, mientras que la de los últimos sólo es 5.4% de su media

Otros ejemplos del empleo del CV pueden encontrarse en el muestreo donde se utiliza para comprobar el grado de confianza de una varianza estimada.

EJERCICIOS DE APLICACION

1. Leer el siguiente texto: “ Una vez recolectados los datos en forma ordenada, es necesario presentarlos en forma tal que se facilite su comprensión y su posterior análisis y utilización. Para ello se ordenan en cuadros numéricos y luego se representan en gráficos para variable discreta mediante diagramas de frecuencias tanto para absolutas o relativas.”

a) Construir una tabla de frecuencias absolutas, según el número de letras que componen cada palabra (rr y ll, debe considerarse una sola letra).

b) Considerando el número de letras por palabra, qué valor tienen: x_4 , x_{27} , x_{12} , x_{34} , x_2

c) Calcular las frecuencias relativas, relativas acumuladas y absolutas acumuladas.

d) Puede ocurrir que $h_4 > h_7$ (en general) Que $H_4 < H_7$?

e) Graficar: n_i , H_i

Desarrollo

Sea X: número de letras que componen cada palabra.

$x_1 = 3$	$x_{11} = 12$	$x_{21} = 2$	$x_{31} = 7$	$x_{41} = 8$
$x_2 = 3$	$x_{12} = 2$	$x_{22} = 9$	$x_{32} = 9$	$x_{42} = 8$
$x_3 = 12$	$x_{13} = 5$	$x_{23} = 8$	$x_{33} = 1$	$x_{43} = 9$
$x_4 = 3$	$x_{14} = 3$	$x_{24} = 1$	$x_{34} = 5$	$x_{44} = 2$
$x_5 = 5$	$x_{15} = 3$	$x_{25} = 11$	$x_{35} = 2$	$x_{45} = 11$
$x_6 = 2$	$x_{16} = 2$	$x_{26} = 4$	$x_{36} = 11$	$x_{46} = 5$
$x_7 = 5$	$x_{17} = 8$	$x_{27} = 3$	$x_{37} = 2$	$x_{47} = 4$
$x_8 = 8$	$x_{18} = 2$	$x_{28} = 2$	$x_{38} = 8$	$x_{48} = 9$
$x_9 = 2$	$x_{19} = 11$	$x_{29} = 7$	$x_{39} = 4$	$x_{49} = 1$
$x_{10} = 9$	$x_{20} = 1$	$x_{30} = 2$	$x_{40} = 8$	$x_{50} = 9$

Tabla de frecuencias

Y_i	n_i	h_i	N_i	H_i
1	4	0.08	4	0.08
2	11	0.22	15	0.30
3	6	0.12	21	0.42
4	3	0.06	24	0.48
5	5	0.1	29	0.58
7	2	0.04	31	0.62
8	7	0.14	38	0.76
9	6	0.12	44	0.88
11	4	0.08	48	0.96
12	2	0.04	50	1.00
Σ	50	1.00		

$x_{4,}=3$, $x_{27}=3$, $x_{12}=2$, $x_{34}=5$, $x_2=3$

¿ Puede ocurrir que $h_4 > h_7$? En general puede suceder que $h_4 > h_7$ dependiendo de la frecuencia de ocurrencia de cada uno de los valores de la variable.

¿ $H_4 < H_7$? siempre ocurrirá porque son frecuencias acumuladas.

Diagramas de frecuencias absolutas

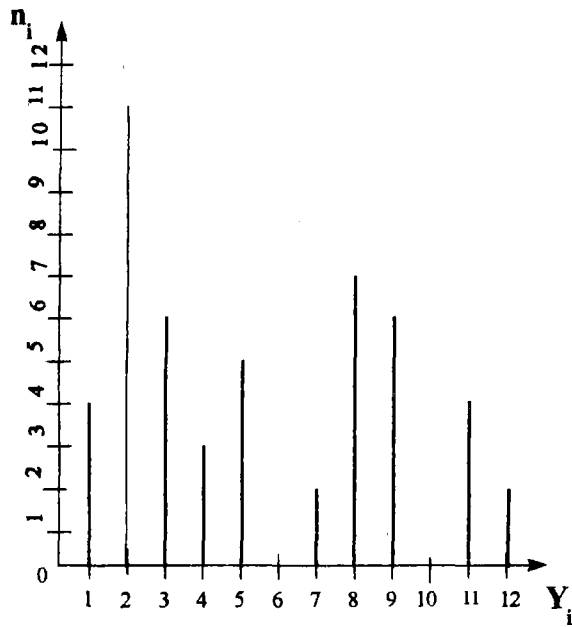
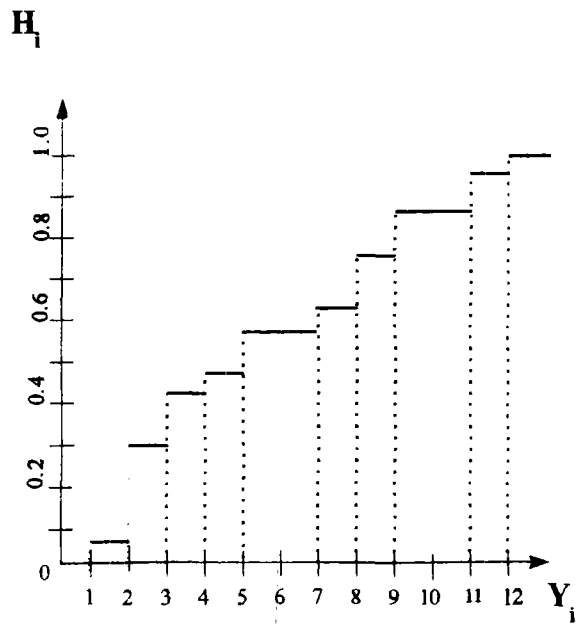


Diagrama de frecuencias relativas acumuladas



2. Las estaturas (en centímetros) de un grupo de personas asistentes a una competencia deportiva en la ciudad de Cali, son las siguientes:

153	123	129	132	147	138	137	134	131	147
138	128	134	148	125	139	146	145	148	135
152	128	146	143	138	138	122	146	137	151
145	124	132	138	144	141	137	146	138	146
152	136	160	159	157	150	160	142	148	130

- a) Agrupar los datos anteriores con una amplitud de intervalo $i= 5$
- b) Interpretar n_2 , N_3 , H_4 , h_2
- c) ¿ Qué porcentaje de personas tienen estatura promedio de 142 centímetros?
- d) ¿ Cuántas personas tienen estatura promedio de 152 centímetros o más?
- e) Construir el polígono de frecuencias absolutas.
- f) Construir el histograma de frecuencias relativas
- g) Construir la ojiva de frecuencias absolutas

Desarrollo

Distribución de asistentes a una competencia deportiva en la ciudad de Cali por estatura

Y_i	Clases	n_i	h_i	N_i	H_i
122	120- 124	3	0.06	3	0.06
127	125 -129	4	0.08	7	0.14
132	130 -134	6	0.12	13	0.26
137	135 -139	12	0.24	25	0.50
142	140 -144	4	0.08	29	0.58
147	145 -149	12	0.24	41	0.82
152	150 -154	5	0.10	46	0.92
157	155 -159	2	0.04	48	0.96
162	160 -164	2	0.04	50	1.00
Σ		50	1.00		

$n_2=4$ cuatro personas tienen estatura promedio de 127 kilogramos.

$N_3=13$ trece personas tienen estatura promedio de 132 kilogramos o menos

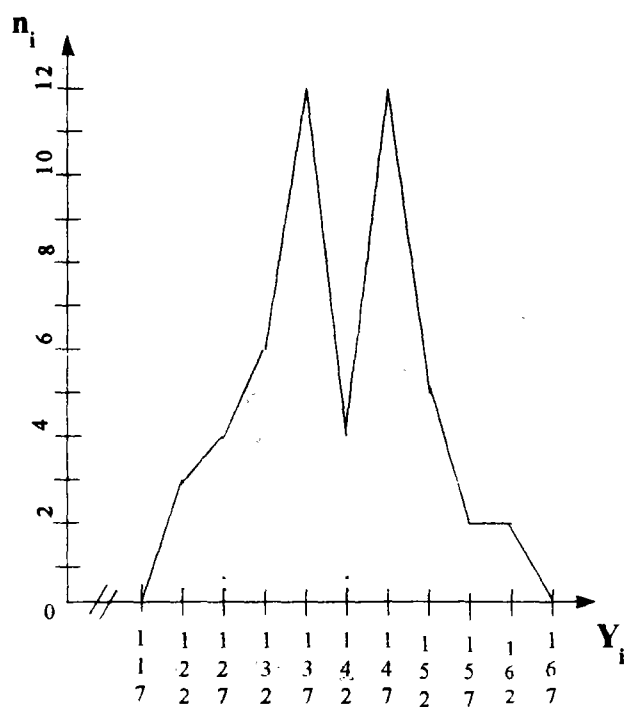
$H_4=0.50$ el 50% de las personas tienen estatura promedio de 137 kilogramos o menos

$h_2=0.08$ el 8% de las personas tienen estatura promedio de 127 kilogramos

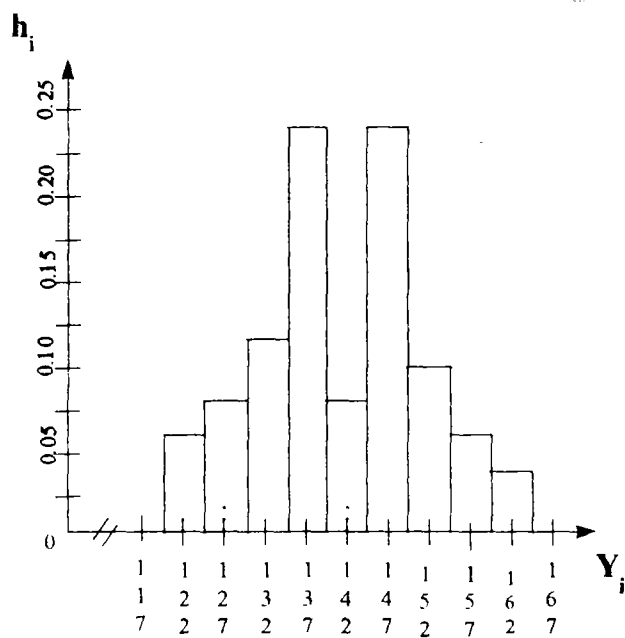
El porcentaje de persona con estatura promedio de 142 centímetros es 8% (h_5)

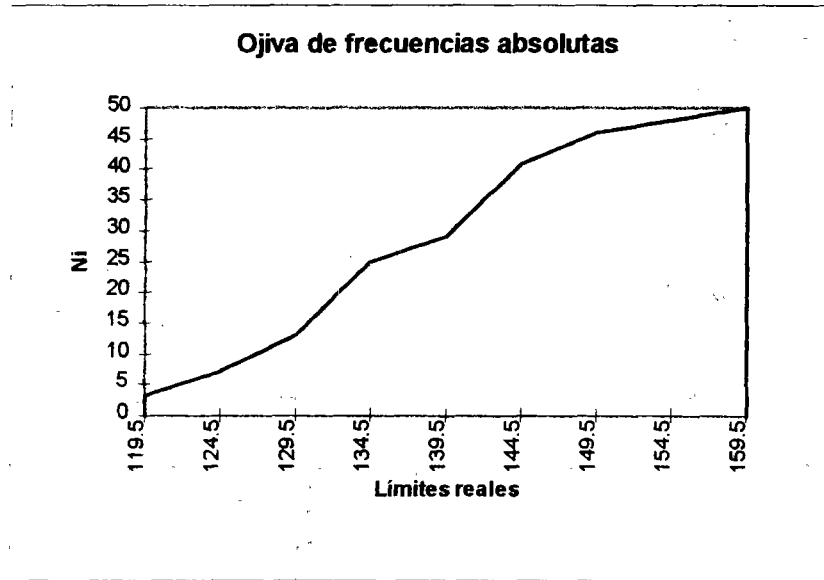
El total de personas con estatura promedio de 152 centímetros o más es 9 ($n_7 + n_8 + n_9$)

Polígono de frecuencias absolutas



Histograma de frecuencias relativas





3. La siguiente distribución corresponde al número de hijos de 20 familias observadas.

Calcular la media aritmética, la mediana, la moda y la desviación estándar.

Yi	ni	Yi*ni	Ni	Yi ² ni
0	2	0	2	0
2	3	6	5	12
4	7	28	12	112
6	4	24	16	144
7	4	28	20	196
Σ	20			464

Media aritmética: $\bar{X} = \frac{86}{20} = 4.3$

Mediana (Me) = 4

Moda (Mo) = 4

$$S^2 = \frac{464}{20} - 43^2 = 23.2 - 18.49 = 4.71$$

$$S = \sqrt{4.71} = 2.17$$

4. La siguiente tabla muestra las edades de 50 personas asistentes a un seminario de Gestión empresarial en la ciudad de Neiva

Clases	n_i	$Y_i \cdot n_i$	N_i	$Y_i^2 \cdot n_i$	Y_i
25 -29	8	216	8	5832	27
30 -34	23	736	31	23552	32
35 -39	10	370	41	13690	37
40 -44	5	210	46	8820	42
45 -49	3	141	49	6627	47
50 -54	1	52	50	2704	52
Σ	50	1725		61225	

Para dicha distribución calcular la media, la mediana, la moda y la desviación estándar.

$$\text{Media aritmética: } \bar{X} = \frac{1725}{50} = 34.5$$

$$\text{Mediana (Me)} = 29.5 + \frac{25 - 8}{23} = 29.5 + 0.74 = 30.24$$

$$\text{Moda (Mo)} = \frac{30 + 34}{2} = 32$$

$$S^2 = \frac{61225}{50} - 34.5^2 = 1224.5 - 1190.25 = 34.25$$

$$\sqrt{34.25} = 5.85$$

Los sueldos mensuales de 70 empleados de oficina son los siguientes:

Ocupación	No. de empleados	Sueldo \$
Recepcionista	2	170.000
Mecanógrafa	12	200.000
Secretaria	8	280.000
Aux. contabilidad	10	465.000
Técnicos electricistas.	24	520.000
Técnicos. mecánicos	14	590.000

Calcular la media aritmética la mediana y la moda

Y_i	n_i	$Y_i n_i$	N_i
170	2	340	2
200	12	2400	14
280	8	2240	22
465	10	4650	32
520	24	12480	56
590	14	8260	70
Σ	70	30064	

Media aritmética: $\bar{X} = \frac{30370000}{70} = 433857.14$

Mediana (Me) = $n/2 = 35$

$N_{j-1} < n/2 \quad 32 < 35 \quad \text{Me} = Y_j = 520000$

Moda (Mo) = $Y_j = 520000$

EJERCICIOS PROPUESTOS

1 La siguiente información corresponde a la ENH etapa 96, para la variable edad del personal ocupado :

Y_i	Clases	n_i	h_i	N_i	H_i
17	10 - 24	7.164	0.196	7.164	0.196
32	25 - 39	16.666	0.455	23.830	0.651
47	40 - 54	9.610			0.914
62	55 - 69	2.732	0.075	36.172	
77	70 - 84	406	0.011	36.578	1.000
92	85 - 99	18	0.000	36.596	1.000
	TOTAL	36.596			

- a) Completar la distribución anterior.
- b) Interpretar n_i , H_i , N_i , h_i
- c) Construir el polígono de frecuencias absolutas
- d) Construir el histograma de frecuencias relativas
- e) Calcular e interpretar la media aritmética
- f) Calcular e interpretar la mediana
- g) Calcular e interpretar la desviación estándar

**Distribución del personal ocupado ENH- etapa 96,
por estado civil**

Estado civil	n_i	h_i	N_i	H_i
Unión libre	6.822	0.186		0.186
casado	13.500	0.369	20.322	
viudo	982		21.304	0.582
separado	3.446	0.094	24.750	0.676
soltero	11.846	0.324	36.596	1.000
TOTAL	36.596	1.000		

a) Completar la distribución anterior

b) Interpretar n_3 , h_4 , N_3 y H_2

c) Construir la gráfica para n_i

3. A partir de los datos de producción de la EAM para la agrupación 311,

Producción (miles de pesos)

34526	36878	34094	37718	42612
60190	62000	21526	57718	5150
55622	43494	58988	38580	34556
63500	62400	45138	6980	32872
13390	41628	38786	34712	53044
6500	48076	37414	62640	10542
54168	47630	26554	44070	9048
63204	60000	62300	38900	46498
56230	43976	47308	58546	63670
32112	17702	61466	62834	56560

Calcular e interpretar

- a) Producción promedio
- b) Producción mediana
- c) Desviación estándar de la producción
- d) Coeficiente de variación

4 Contestar los siguientes puntos, señalando con V o F según sea cierto o falso

- a) Al hacer una gráfica, la variable a estudiar va en el eje de las abscisas y las frecuencias en la ordenada ()
- b) Para calcular las marcas de clase, se suma el límite inferior al superior del intervalo y se divide entre dos ()
- c) El polígono de frecuencias es un gráfico de barras ()
- d) El número de accidentes según sus causas es una variable discreta ()
- e) La suma de las frecuencias absolutas es igual a 1 ()
- f) $H_5 = 1.1$ $H_4 = 0.80$ $h_5 = 0.30$ ()
- g) Un grupo de valores pueden tener más de una media aritmética ()
- h) En una serie sencilla cuando n no es par, la mediana es igual al valor del dato central ()
- i) Si cierta distribución tiene una varianza igual a 144 y otra una desviación estándar igual a 11, puede afirmarse que la primera tiene mayor dispersión ()

BIBLIOGRAFIA

MARTINEZ, Bencardino Ciro. (1982). Estadística apuntes y 600 problemas resueltos. Ecoe, segunda edición.

MENDENHALL, William y REINMUTH, James E. Estadística para administración y economía. Iberoamérica, 3° edición.

ZUWAYLIF, F. (1977). estadística general aplicada. Fondo Educativo Interamericano.

HABER, Audrey y RUNYON, Richard. (1973). Estadística general. Fondo Educativo Interamericano.S.A.

YAMANE, Taro. Estadística. Programas editoriales.

KISH, Leslie. (1979). Muestreo de encuestas. Trillas, primera edición en español.

Impreso en la División de Ediciones del
Departamento Administrativo Nacional de
Estadística, DANE
Santafé de Bogotá, D.C. - Colombia - Octubre 1999