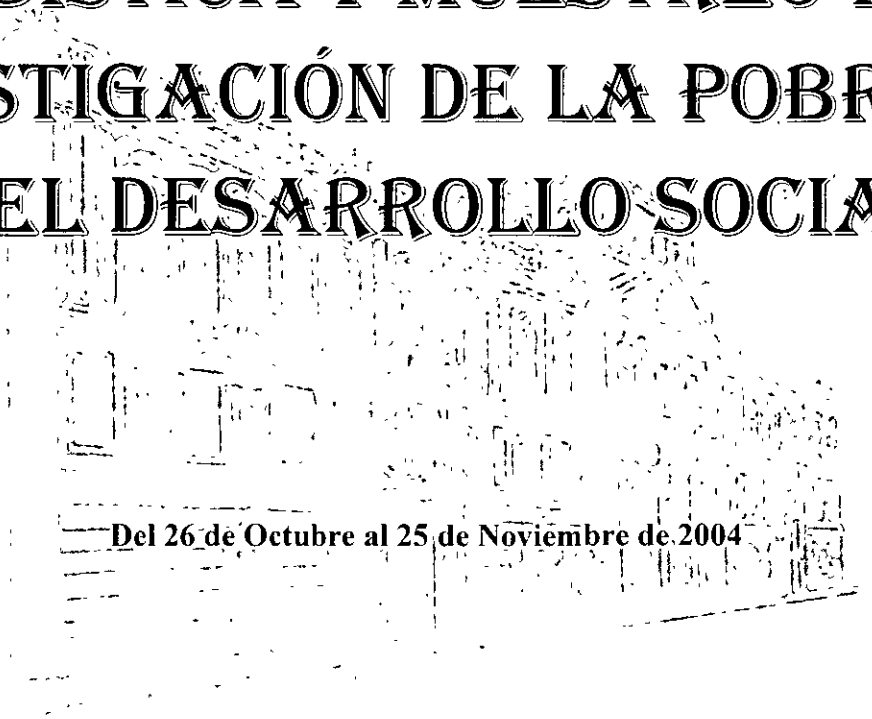
FACULTAD DE INGENIERÍA UNAM
DIVISIÓN DE EDUCACIÓN CONTINUA

CURSOS INSTITUCIONALES

ESTADÍSTICA Y MUESTREO PARA INVESTIGACIÓN DE LA POBREZA Y EL DESARROLLO SOCIAL



Del 26 de Octubre al 25 de Noviembre de 2004

APUNTES GENERALES

CI - 198

Instructor: Lic. Tonatiu Suárez

GOBIERNO DEL DISTRITO FEDERAL
OCTUBRE/NOVIEMBRE DE 2004

Palacio de Minería, Calle de Tacuba No. 5, Primer piso, Delegación Cuauhtémoc, CP 06000, Centro Histórico, México D.F.,
APDO Postal M-2285 ■ Tels: 5521.4021 al 24, 5623 2910 y 5623.2971 ■ Fax: 5510.0573

1. ¿Qué es la estadística?

- La estadística es un método científico de análisis que se utiliza en las ciencias sociales.

- El objetivo principal de la estadística es utilizar una muestra de datos para inferir características de toda la población.

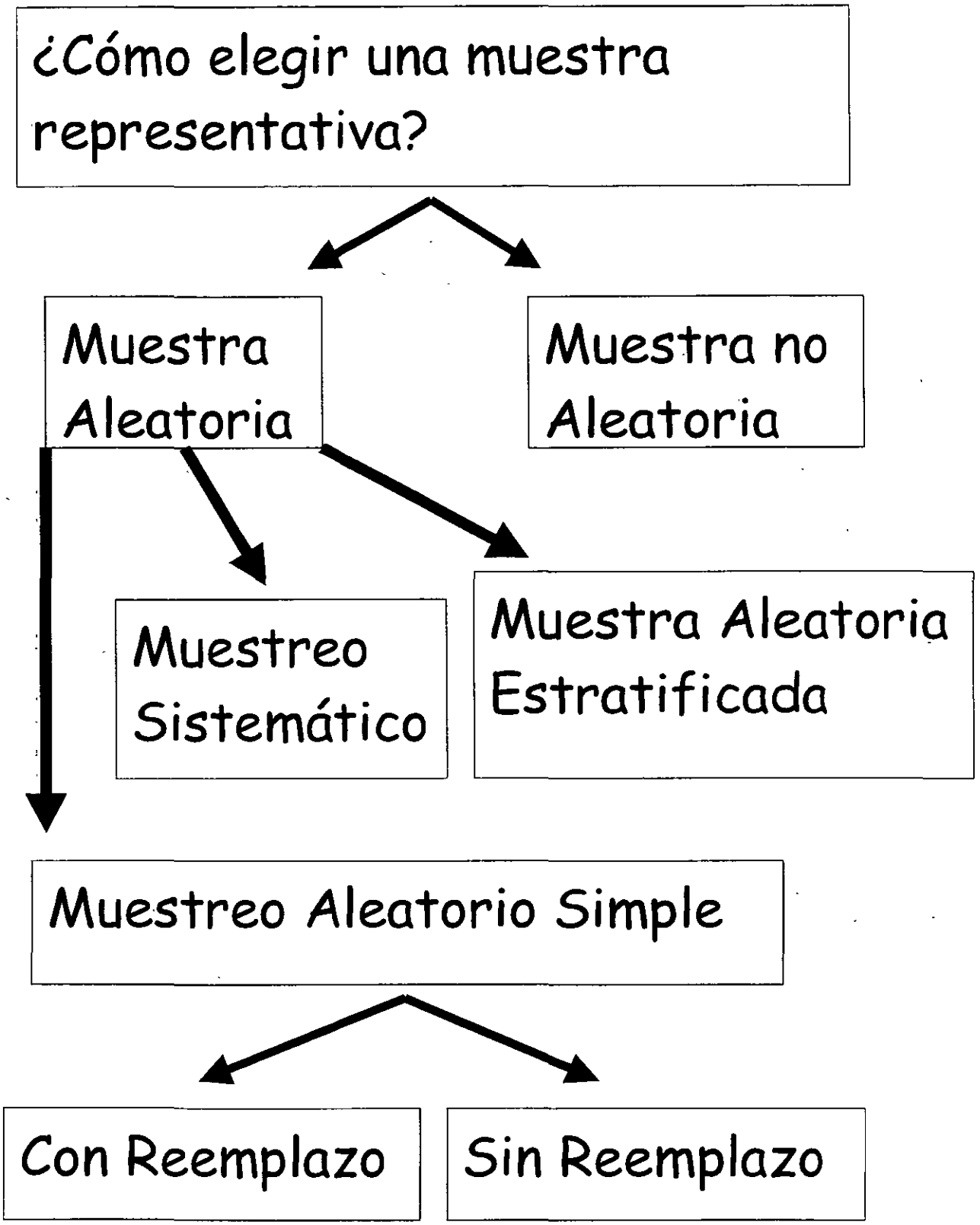
- Para realizar tal tipo de inferencia es necesario contar con una muestra representativa de la población bajo análisis.

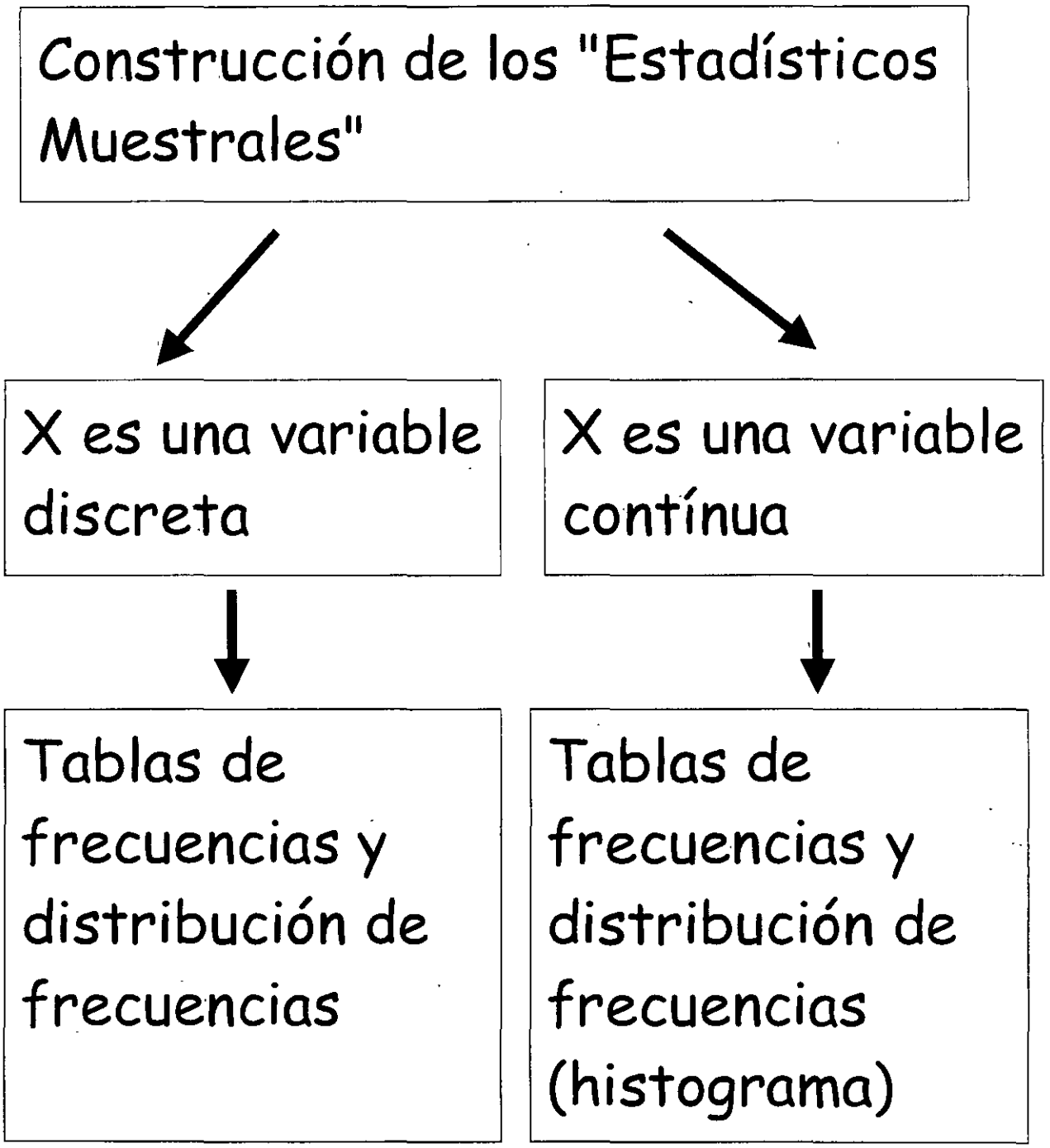
2. Objetivos del curso.

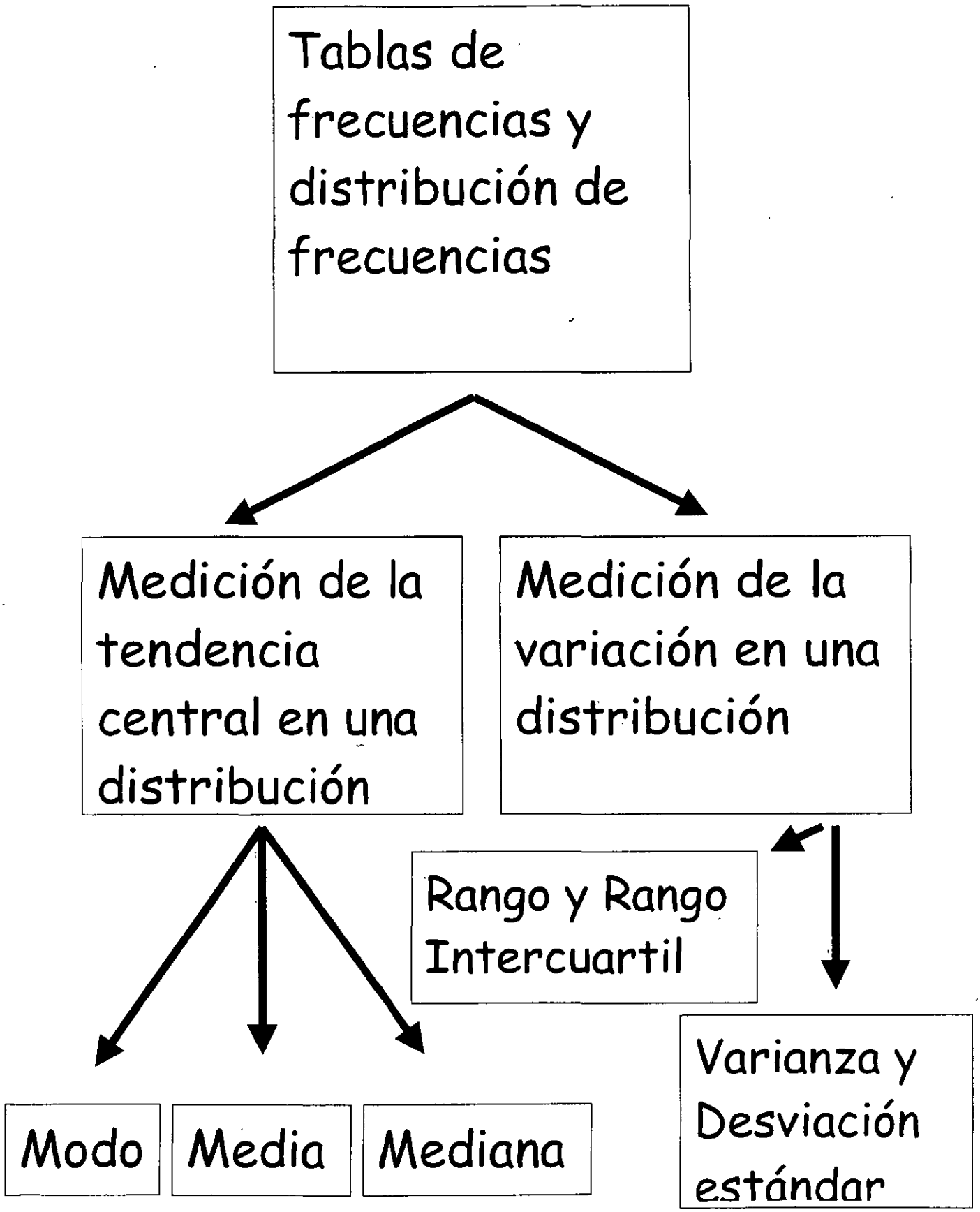
¿Cómo seleccionar una muestra representativa?

Caracterización de los datos muestrales en unos pocos números llamados "estadísticos muestrales".

Uso de los "estadísticos muestrales" para hacer inferencias acerca de la población.


¿Cómo elegir una muestra representativa?
Muestra Aleatoria
Muestra no Aleatoria
Muestreo Sistemático
Muestra Aleatoria Estratificada
Muestreo Aleatorio Simple
Con Reemplazo
Sin Reemplazo


Construcción de los "Estadísticos Muestrales"
X es una variable discreta
X es una variable contínua
Tablas de frecuencias y distribución de frecuencias
Tablas de frecuencias y distribución de frecuencias (histograma)


Tablas de frecuencias y distribución de frecuencias
Medición de la tendencia central en una distribución
Medición de la variación en una distribución
Modo
Media
Mediana
Rango y Rango Intercuartil
Varianza y Desviación estándar

Hay dos métodos básicos para seleccionar datos desde una población.

1. **Muestra Aleatoria:** Cada elemento de la población tiene la misma chance de ser seleccionado.

2. **Muestra No Aleatoria:** Algunos elementos de la población tienen mayor chance de ser seleccionados.

Muestra Aleatoria Simple: Selección de N elementos de una población de forma tal que cualquier posible combinación de N elementos de esa población tiene la misma probabilidad de ser seleccionada.

Muestra Aleatoria Simple sin reemplazo: Cuando cada elemento seleccionado para la muestra no se devuelve a la población para su posible re-selección.

Muestra Aleatoria Simple con reemplazo: Cuando cada elemento seleccionado para la muestra se devuelve a la población para su posible re-selección.

Muestreo Sistemático: Los elementos de la población se ordenan en una secuencia aleatoria y se selecciona cada n-ésimo elemento de la secuencia.

Muestreo Aleatorio Estratificado: Se divide a la población en subgrupos y en cada uno de estos subgrupos se realiza un muestreo aleatorio simple.

Tablas y Gráficos de Frecuencia

Suponga que un consultor quiere caracterizar la asistencia al trabajo de los empleados de una compañía financiera. Para ello obtiene una muestra aleatoria de 20 empleados en el año 1998. Hagamos que X represente el número de días que un empleado estuvo ausente durante 1998. Para los 20 empleados (ordenados alfabéticamente) los días de ausencia son:

X = 6, 4, 4, 6, 0, 4, 3, 6, 1, 3, 8, 3, 6, 0, 1, 6, 11, 5, 10, 8.

X se denomina una variable aleatoria discreta.

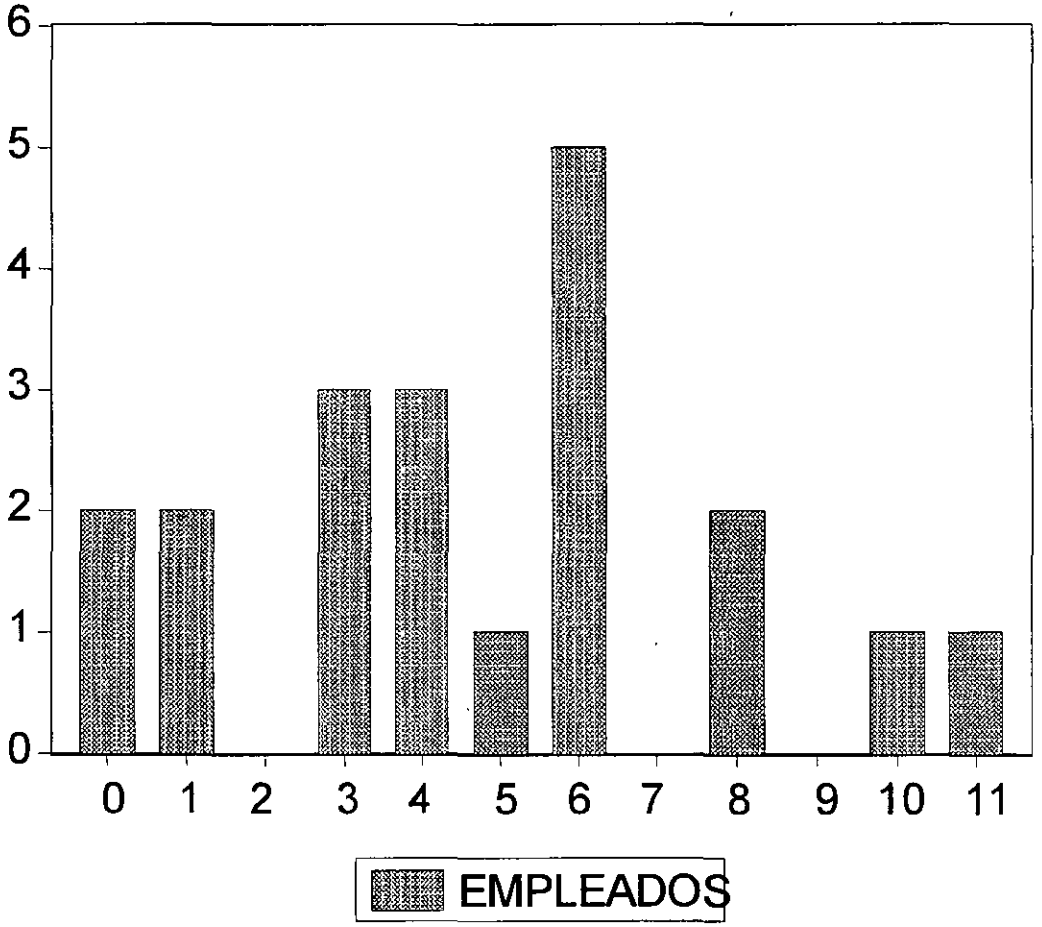
- Una variable aleatoria es aquella variable que toma valores dependiendo del resultado de un experimento.

- Una variable aleatoria es discreta cuando solo toma un número contable de valores.

Para resumir los datos de la muestra podemos construir una tabla con la frecuencia (número de empleados por cada día de ausencia) y la frecuencia relativa (número de empleados por cada día de ausencia sobre el número total de empleados)

Número de días ausente	Frecuencia	Frecuencia Relativa
0	2	2/20=0.10
1	2	2/20=0.10
2	0	0/20=0.00
3	3	3/20=0.15
4	3	3/20=0.15
5	1	1/20=0.05
6	5	5/20=0.25
7	0	0/20=0.00
8	2	2/20=0.10
9	0	0/20=0.00
10	1	1/20=0.05
11	1	1/20=0.05
Total	20	1

Desde esta tabla podemos graficar la distribución de frecuencias:

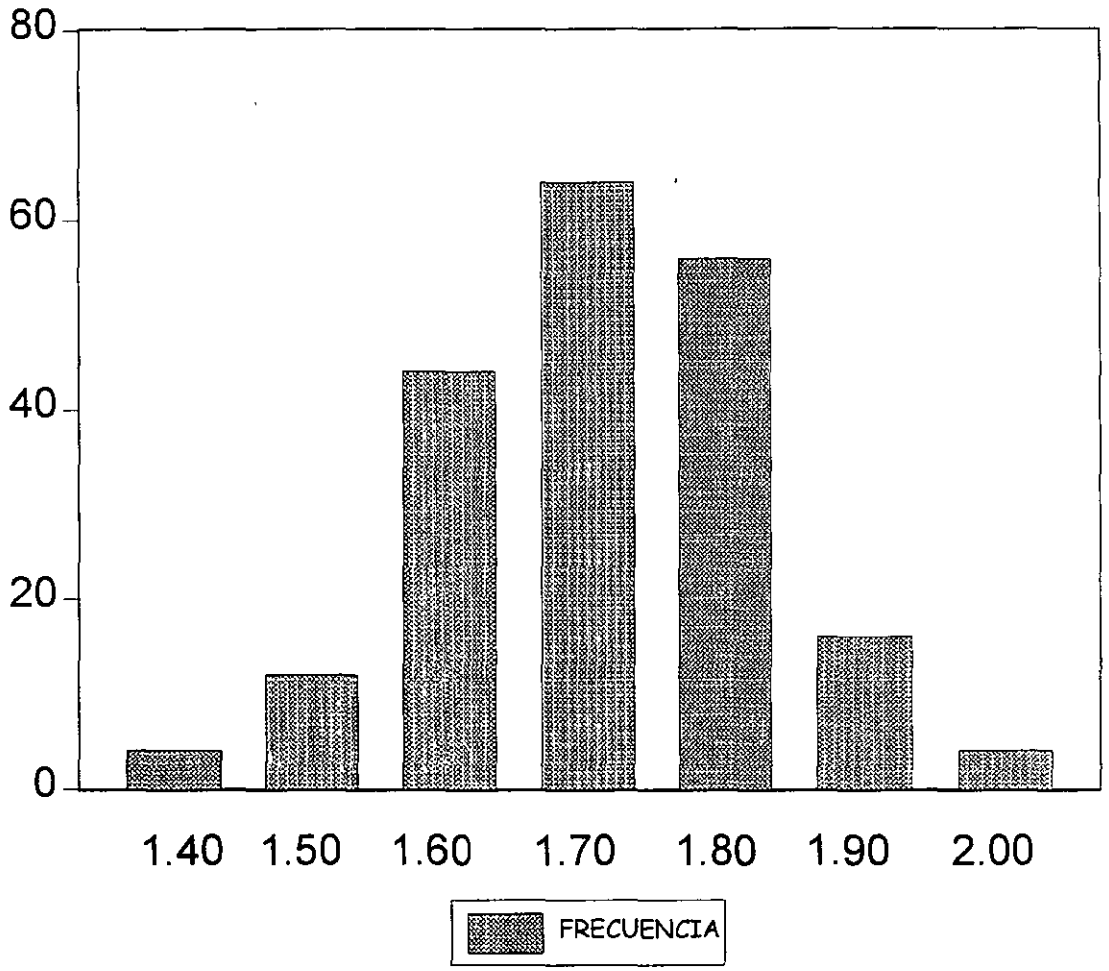


Si la variable aleatoria fuera contínua, es decir que adopta cualquier valor en un intervalo real, ya no podríamos hablar de la "frecuencia" para un valor determinado de X porque nunca observaríamos el mismo valor exactamente. En este caso lo que podemos hacer es calcular las frecuencias dentro de un determinado intervalo.

Ejemplo: Supongamos que tomamos una muestra aleatoria de 200 personas de la Capital Federal y registramos su altura, que denotamos con la letra X.
Como X puede tomar cualquier valor, por ejemplo 1,6543... metros de altura, lo que debemos hacer es construir intervalos de altura y contar el número de personas con alturas incluídas en cada intervalo:

Intervalo	Centro	Frecuencia	Frecuencia Relativa
1.35-1.45	1.40	4	0.02
1.45-1.55	1.50	12	0.06
1.55-1.65	1.60	44	0.22
1.65-1.75	1.70	64	0.32
1.75-1.85	1.80	56	0.28
1.85-1.95	1.90	16	0.08
1.95-2.05	2.00	4	0.02
Total		200	1.00

Para graficar la distribución de frecuencias utilizamos un histograma o gráfico de barras.


80
60
40
20
0
1.40
1.50
1.60
1.70
1.80
1.90
2.00
FRECUENCIA

La división de los datos ordenados por su frecuencia en centécimos se denomina **PERCENTIL.**

Por ejemplo, una persona de 1.55 metros de altura tiene solo 8% de gente más baja que él y por lo tanto su altura se dice que es el **octavo percentil** de la distribución.

Los percentiles que dividen a los datos en cuatro cuartos tienen nombres especiales.

El percentil 25 y el 75 se llaman **PRIMER y TERCER CUARTIL (Q_1 y Q_3)**. El percentil 50 se denomina **MEDIANA (Q_2)**.

Las fórmulas utilizadas para calcular estas medidas son:

$$Q_1 = L_{s1} + [(\% \text{ hasta } 25)/(\% \text{ en } 25)] \times A_{Q1}$$

Donde:

L_{s1} es el límite superior del intervalo anterior al intervalo que contiene a Q_1.

A_{Q1} es el ancho del intervalo que contiene a Q_1.

$Q_2 = L_{s2} + [(\%\ \text{hasta}\ 50)/(\%\ \text{en}\ 50)] \times A_{Q2}$

Donde:

L_{s2} es el límite superior del intervalo anterior al intervalo que contiene a Q_2.

A_{Q2} es el ancho del intervalo que contiene a Q_2.

Ejemplos:

$Q_1 = 1.55 + (17/22) \times 0.10 = 1.6272$

$Q_2 = 1.65 + (20/32) \times 0.10 = 1.7125$

$Q_3 = L_{s3} + [(\%\ \text{hasta}\ 75)/(\%\ \text{en}\ 75)] \times A_{Q3}$

Donde:

L_{s3} es el límite superior del intervalo anterior al intervalo que contiene a Q_3.

A_{Q3} es el ancho del intervalo que contiene a Q_3.

CENTRO DE LA DISTRIBUCION

Hay tres formas de definir el centro de una distribución:

Modo: El modo de una distribución se define como el valor de mayor frecuencia.

Mediana: La mediana es la observación que divide en dos partes iguales a la distribución.

Media o Promedio: Supongamos una muestra de N observaciones denotadas por X_1, X_2,...,X_N. La media o promedio se obtiene sumando todos los valores y dividiendo por el tamaño muestral:

$$Media: \quad X = \frac{1}{N}\sum_{i=1}^{N} X_i$$

En caso de tener datos agrupados en intervalos la media se calcula como:

$$Media: \quad X = \frac{1}{N}\sum_{i=1}^{C} X_i f_i$$

Donde:
C es el número de intervalos y f_i es la frecuencia del intervalo i.

- ¿Cuál de las tres medidas del centro de una distribución es la apropiada?

VARIACION DE LA DISTRIBUCION

Las medidas de la variación de una distribución son:

Rango: El rango es simplemente la distancia entre el mayor y el menor valor de la muestra.

Rango Intercuartil (IQR): El rango intercuartil es la distancia entre el primer y el tercer cuartil:

$$IQR = Q_3 - Q_1$$

Desvío medio absoluto (MAD):

$$MAD = \frac{1}{N}\sum_{i=1}^{N}|X_i - X|$$

Desvío medio cuadrático (MSD):

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

Varianza y Desvío Estándar:

$$Varianza: \; S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - X)^2$$

Para datos agrupados en intervalos la definición se modifica a:

$$Varianza: \; S^2 = \frac{1}{N-1} \sum_{i=1}^{C} (X_i - X)^2 f_i$$

Para compensar el hecho de haber elevado al cuadrado las desviaciones con respecto a la media se puede tomar la raíz cuadrada de la varianza:

Desvío Estándar : $S = \sqrt{S^2}$

Ejemplo:

Datos X	Desvíos	Desvíos Absolutos	Desvíos Cuadrados
10	-30	30	900
20	-20	20	400
30	-10	10	100
50	10	10	100
90	50	50	2500
Suma	0	120	4000

$$Media: \quad X = \frac{200}{5} = 40$$

$$MAD = \frac{120}{5} = 24$$

$$MSD = \frac{4000}{5} = 800$$

$$S^2 = \frac{4000}{4} = 1000$$

$$S = \sqrt{\frac{4000}{4}} = 32$$

¿Porque usamos N-1 como divisor en la fórmula de la varianza?

PROBABILIDAD

Una probabilidad indica la posibilidad de que un evento futuro ocurra. La probabilidad varía entre cero y uno, reflejando el rango de posibilidades desde imposible (cero) hasta totalmente cierto (uno).

Los elementos básicos de la teoría de las probabilidades son los resultados del experimento bajo estudio. Estos resultados se denominan **EVENTOS**.

La colección de todos los resultados posibles de un experimento se denomina **ESPACIO MUESTRAL.**

Ejemplo: Si arrojamos una moneda al aire la probabilidad de obtener "cara" es 0.50. Esto significa que existe una posibilidad del 50% de obtener una "cara". El espacio muestral está dado por el conjunto {"cara", "ceca"} y los eventos (resultados del experimento de arrojar una moneda) son "cara" y "ceca".

Cuando asignamos probabilidades a los resultados de un experimento, se deben cumplir dos requerimientos:

(a) La probabilidad asignada a cada evento debe estar entre cero y uno.
(b) Las probabilidades de todos los eventos deben sumar uno.

Existen tres formas diferentes de asignar probabilidades a los eventos de un experimento:

1. Frecuencia Relativa: asigna probabilidades en base a datos históricos ó en base a experimentación. Este tipo de probabilidad se define como el número de veces que un evento ocurre dividido el número total de veces que se hace el experimento.
2. Probabilidad Subjetiva: la probabilidad subjetiva refleja sentimientos u opiniones acerca de la posibilidad de que un resultado determinado ocurra.

3. Probabilidad Clásica: el método de probabilidad clásica se basa en el supuesto de que todos los eventos de un experimento son igualmente probables. En el caso más simple:

$$\text{Probabilidad de Ocurrencia} = \frac{\text{Casos Favorables}}{\text{Casos Posibles}}$$

Definiciones

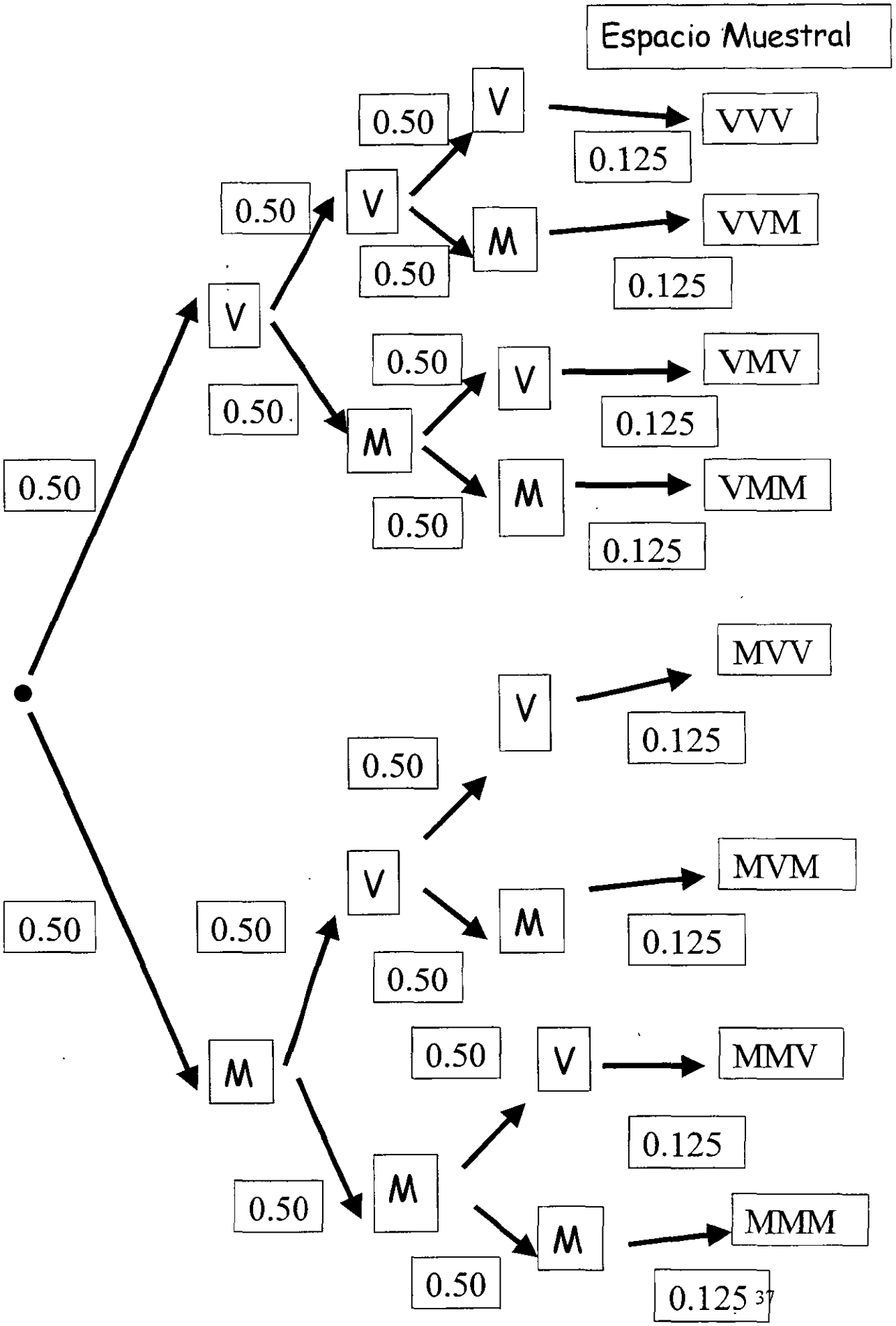
El **COMPLEMENTO** de cualquier evento es la colección de resultados que no está contenida en ese evento.

Un **EVENTO CONJUNTO** es un evento que tiene dos o más características.

Una lista de eventos es **COLECTIVAMENTE EXHAUSTIVA** si incluye todos los eventos que pueden ocurrir.

Dos eventos son **MUTUAMENTE EXCLUSIVOS** si la ocurrencia de uno de ellos impide la ocurrencia del otro.

Ejemplo:
Supongamos que el "experimento" consiste en tener una familia de tres hijos. Un resultado típico de este experimento sería VMV, esto es tener un varón, luego una mujer y por último otro varón. ¿Cómo podríamos encontrar la probabilidad de este resultado?

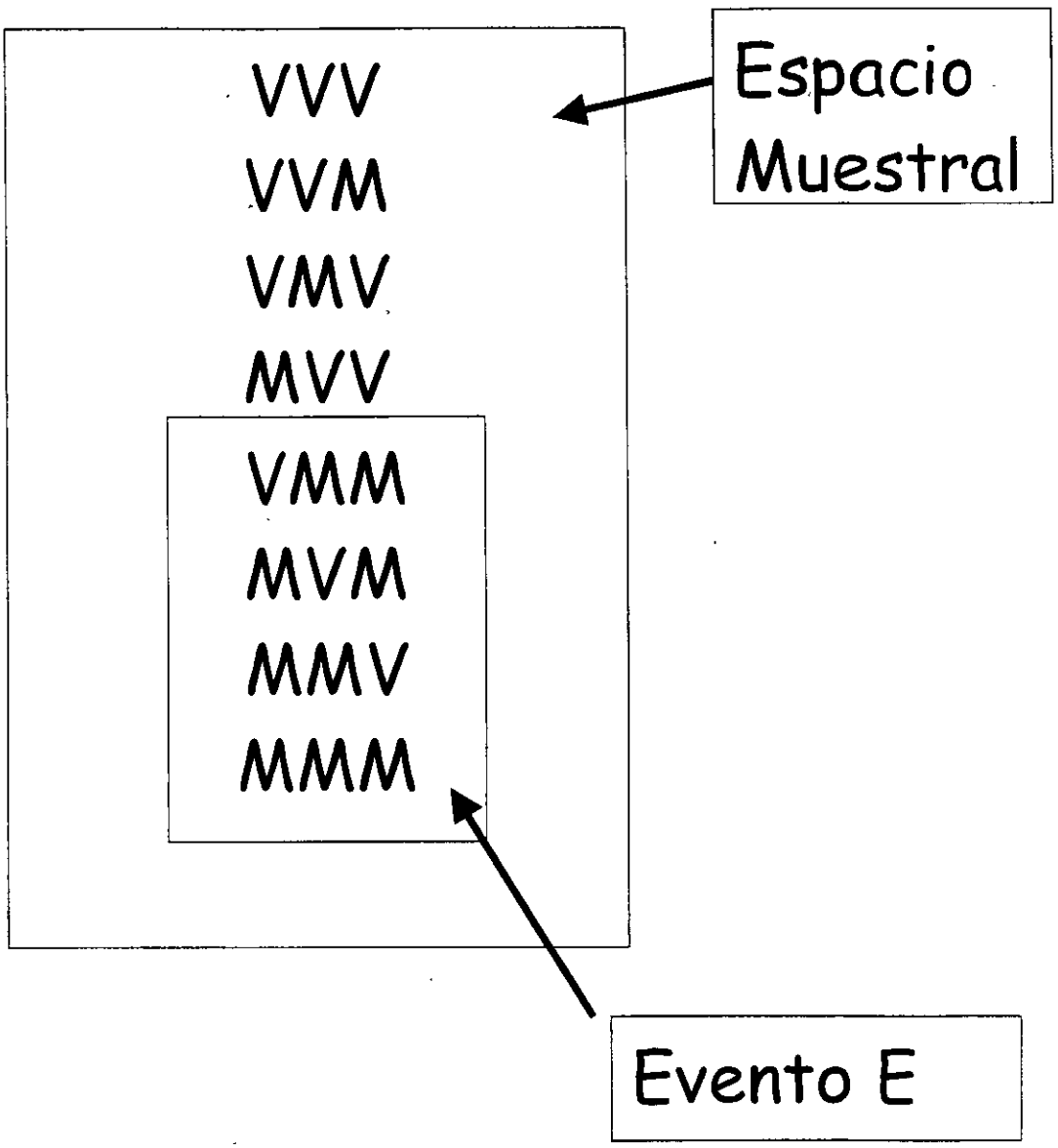

Espacio Muestral
0.50
V
0.50
V
0.50
V
VVV
0.125
0.50
M
VVM
0.125
0.50
V
VMV
0.125
0.50
M
0.50
M
VMM
0.125
0.50
V
MVV
0.125
0.50
V
0.50
M
MVM
0.125
0.50
0.50
M
0.50
V
MMV
0.125
0.50
M
0.50
M
MMM
0.125

Espacio Muestral	Probabilidades
VVV	1/8
VVM	1/8
VMV	1/8
VMM	1/8
MVV	1/8
MVM	1/8
MMV	1/8
MMM	1/8

Definamos el evento "E : al menos dos mujeres". Este evento está representado por los siguientes resultados:

E = {VMM, MVM, MMV, MMM}

En términos de un diagrama de Venn:



¿Cuál es la probabilidad de E?

$$\Pr(E) = 0.125+0.125+0.125+0.125$$
$$= 0.500$$

Es decir que la probabilidad de un evento es la suma de las probabilidades de todos los resultados incluidos en ese evento.

Ejemplo:
encuentre las probabilidades de los eventos "G: menos de dos mujeres" y "H: todos los hijos del mismo género".

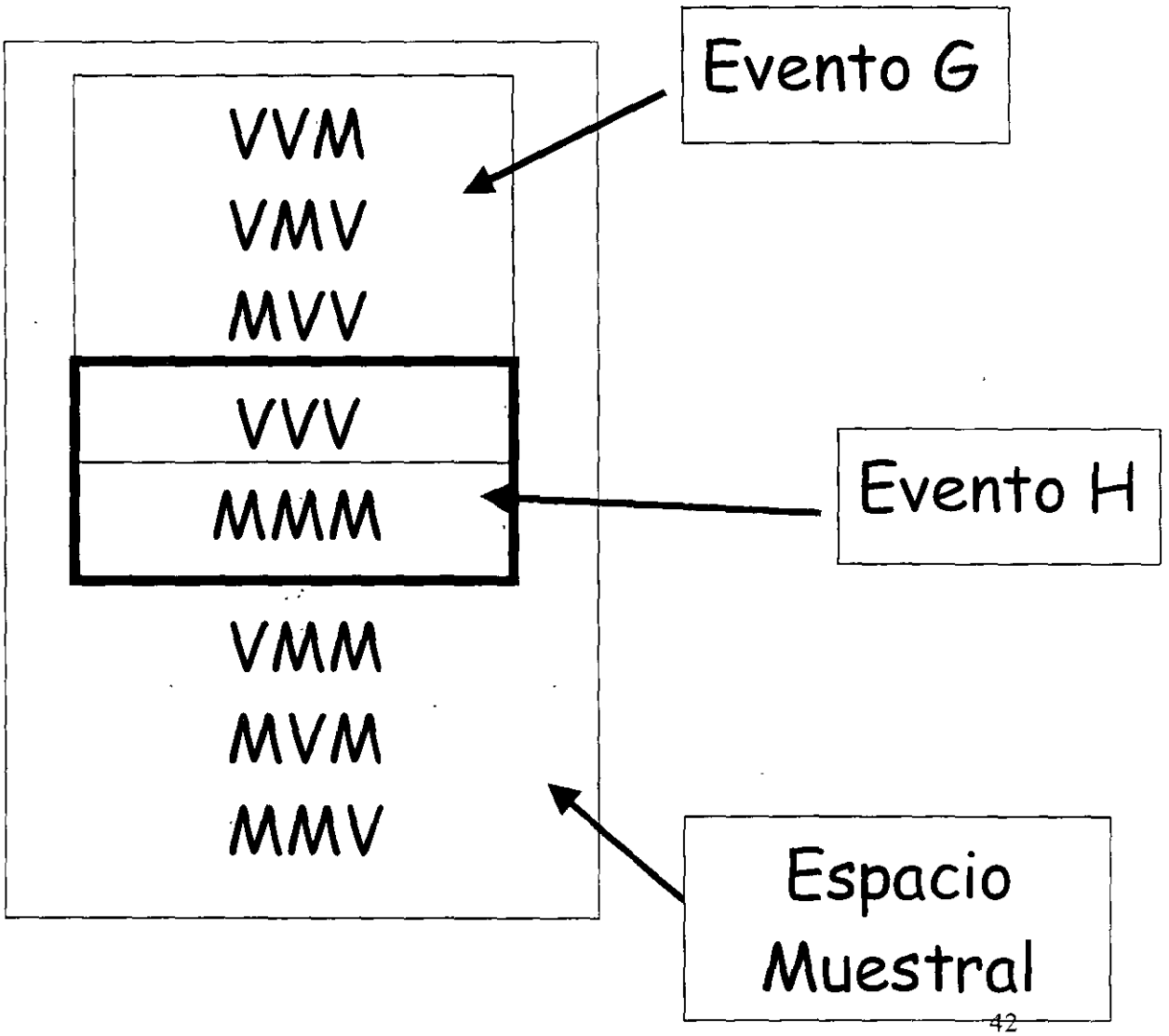
Ahora supongamos que los padres estarían desilucionados si tuvieran menos de dos mujeres o si fueran todos sus hijos del mismo género. ¿Cuál es la probabilidad de que esto ocurra?

Aqui tenemos un evento conjunto, es decir, el evento "G ó H".
Para esto definimos a este evento conjunto como el conjunto de elementos que pertenecen al evento G ó al H ó a ambos.

Es decir:

"G ó H" = {VVV, VVM, VMV, MVV, MMM}

En términos de una figura,



Como tenemos cinco resultados en el evento conjunto, la probabilidad del mismo es 5/8.

Los padres estarían doblemente desilucionados si tuvieran menos de dos mujeres y si fueran todos sus hijos del mismo género. Es decir que esto es el evento "G y H". ¿Cuál es la probabilidad de que esto ocurra?

Del espacio muestral podemos ver que solo existe un punto que pertenece tanto a G como a H:

Evento "G y H" = { VVV }

En general definimos,

"G y H" = Todos los puntos que estan en ambos eventos "G" y "H"

Como solo hay un evento en "G y H", su probabilidad es 1/8.

También podemos encontrar la probabilidad de "G y H" desarrollando una fórmula.

Primero consideremos dos eventos que no tengan ningún punto en común. Por ejemplo, evento "I = { VVV }" y el evento G.

Como los eventos no se superponen, se denominan mutuamente excluyentes, es decir que si uno ocurre el otro no puede ocurrir.

En este caso es obvio que la probabilidad de "I ó G" es:

$$\begin{aligned} Pr(I \text{ ó } G) &= Pr(I) + Pr(G) \\ &= 1/8 + 4/8 \\ &= 5/8 \end{aligned}$$

Esta fórmula de simple adición no siempre funciona, por ejemplo:

$$Pr(G \text{ ó } H) \neq Pr(G) + Pr(H)$$

$$5/8 \neq 1/8 + 2/8$$

En este caso la fórmula no funciona porque al sumar Pr(G) y Pr(H) estamos contando dos veces la intersección. Restando de la fórmula Pr(G y H) elimina este doble conteo.

Por lo tanto, la fórmula general es:

Pr(G ó H) =
Pr(G) + Pr(H) - Pr(G y H). (*)

En nuestro ejemplo,

5/8 = 4/8 + 2/8 - 1/8

La fórmula (*) también se aplica en los casos en que los eventos son mutuamente excluyentes porque en ese caso Pr(I y G) = 0.

Es decir que en el caso especial de eventos mutuamente excluyentes:

Pr(I ó G) = Pr(I) + Pr(G)

Por otra parte, en general para el complemento de cualquier evento E, E^c tenemos la siguiente fórmula:

$Pr(E) + Pr(E^C) = 1,$

Ya que los puntos en E y en E^C completan el espacio muestral.

Probabilidad Condicional

Es el cálculo de la probabilidad después de que se conoce una condición (evento).

Por ejemplo, en el caso de la familia de tres hijos supongamos que se conoce que el evento G ha ocurrido (menos de dos mujeres).

Cuál es la probabilidad de que se haya producido el evento H?

Esto es, imaginemos que podemos repetir el "experimento" un gran número de veces y consideremos solo los casos en los que G ha ocurrido, cuán frecuentemente ocurrirá H?

Esto es lo que se llama probabilidad de H condicional a G y se denota por:

$$Pr(H \mid G)$$

En general,

$Pr(H \mid G) = Pr(H \text{ y } G)/Pr(G)$

Ó,

$Pr(H \text{ y } G) = Pr(H \mid G) * Pr(G)$

Ejemplo.

La siguiente tabla muestra los porcentajes de participación en la fuerza de trabajo de hombres y mujeres en USA en 1985.

	Hombre	Mujer
Empleado	51.9%	40.9%
Desempleado	3.9%	3.3%
Total	55.8%	44.2%

(a) Cuál es la tasa de desempleo?
(b) Cuál es Pr(Des. | hombre)?
(c) Cuál es Pr(Des. | mujer)?

Independencia

Dos eventos F y E se denominan estadísticamente independientes si $Pr(F \mid E) = Pr(F)$.

Esto es, la ocurrencia de E no afecta la ocurrencia del evento F.

Implicancias:

$Pr(E \text{ y } F) = Pr(E) * Pr(F)$

Resúmen de fórmulas

	Pr(A ó B)	Pr(A y B)
Regla General	Pr(A) + Pr(B) - Pr(A y B)	Pr(B|A)*Pr(A)
Caso Especial	Pr(A) + Pr(B) si A y B son mutuamente excl.	Pr(A)*Pr(B) si $A \perp B$

Teorema de Bayes

Consideremos el teorema de Bayes usando un ejemplo.

El gerente de marketing de una Jugetería está planeando introducir un nuevo juguete en el mercado. En el pasado el 40% de los juguetes lanzados al mercado por la compañía han sido exitosos. Antes de lanzar cada juguete, se realiza un análisis de mercado y se escribe un reporte que puede ser favorable o desfavorable al lanzamiento. En el pasado, 80%

de los lanzamientos exitosos recibieron reportes favorables y un 30% de los lanzamientos no exitosos también recibieron reportes favorables. El departamento de marketing de la empresa desearía saber la probabilidad de que el nuevo juguete fuera exitoso:

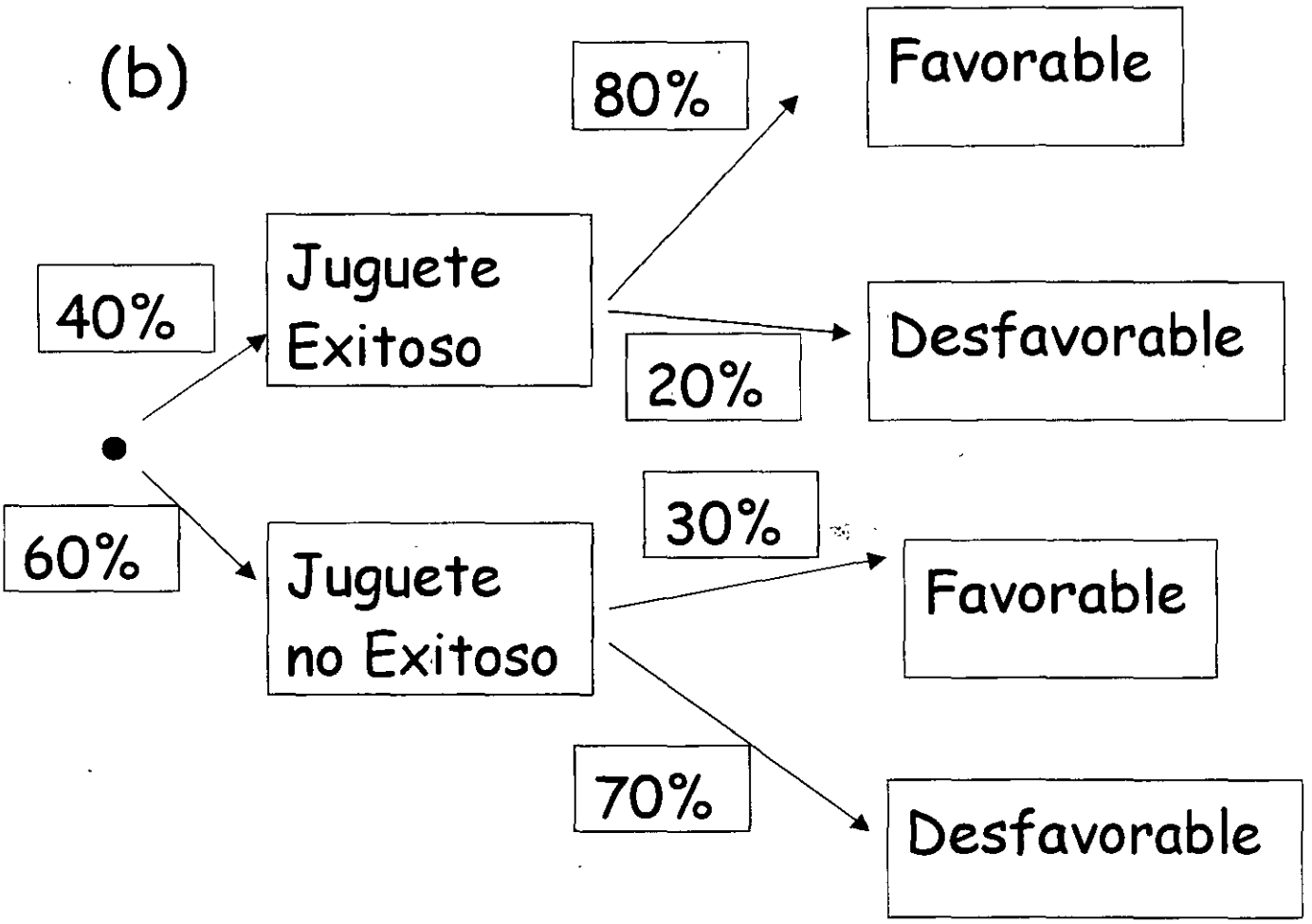
(a) Antes del reporte del análisis de mercado.

(b) Si el reporte es favorable.

(c) Si el reporte es desfavorable.

(a) Pr(exitoso)=0.40

(b)



Del árbol de probabilidades podemos extraer el espacio muestral:

Espacio Muestral =

{ EF, ED, NEF, NED}

Las respectivas probabilidades son:

Pr(EF) = 0.32
Pr(ED) = 0.08
Pr(NEF) = 0.18
Pr(NED) = 0.42
Total = 1

La probabilidad que necesitamos es:

$$\Pr(E \mid F) = \Pr(E \text{ y } F)/\Pr(F)$$
$$= 0.32/(0.32+0.18)$$
$$= 0.64$$

$$\Pr(F) = \Pr(E \text{ y } F) + \Pr(NE \text{ y } F)$$
$$= 0.32 + 0.18 = 0.50$$

(d) $$\Pr(E \mid D) = \Pr(E \text{ y } D)/\Pr(D)$$
$$= 0.08/(0.08+0.42)$$
$$= 0.16$$

$$\Pr(D) = \Pr(E \text{ y } D) + \Pr(NE \text{ y } D)$$
$$= 0.08 + 0.42 = 0.50$$

Formalmente, la fórmula de Bayes es:

$$Pr(E|F) = Pr(E \text{ y } F)/(Pr(EyF)+Pr(NEyF))$$

Donde,

$Pr(EyF) = Pr(F|E)*Pr(E)$

$Pr(NEyF) = Pr(F|NE)*Pr(NE)$

Las probabilidades iniciales, antes de realizar cualquier test se denominan probabilidades anteriores. Las probabilidades obtenidas después del testeo se denominan probabilidades posteriores.

Distribuciones de probabilidad

1) Variables Aleatorias Discretas

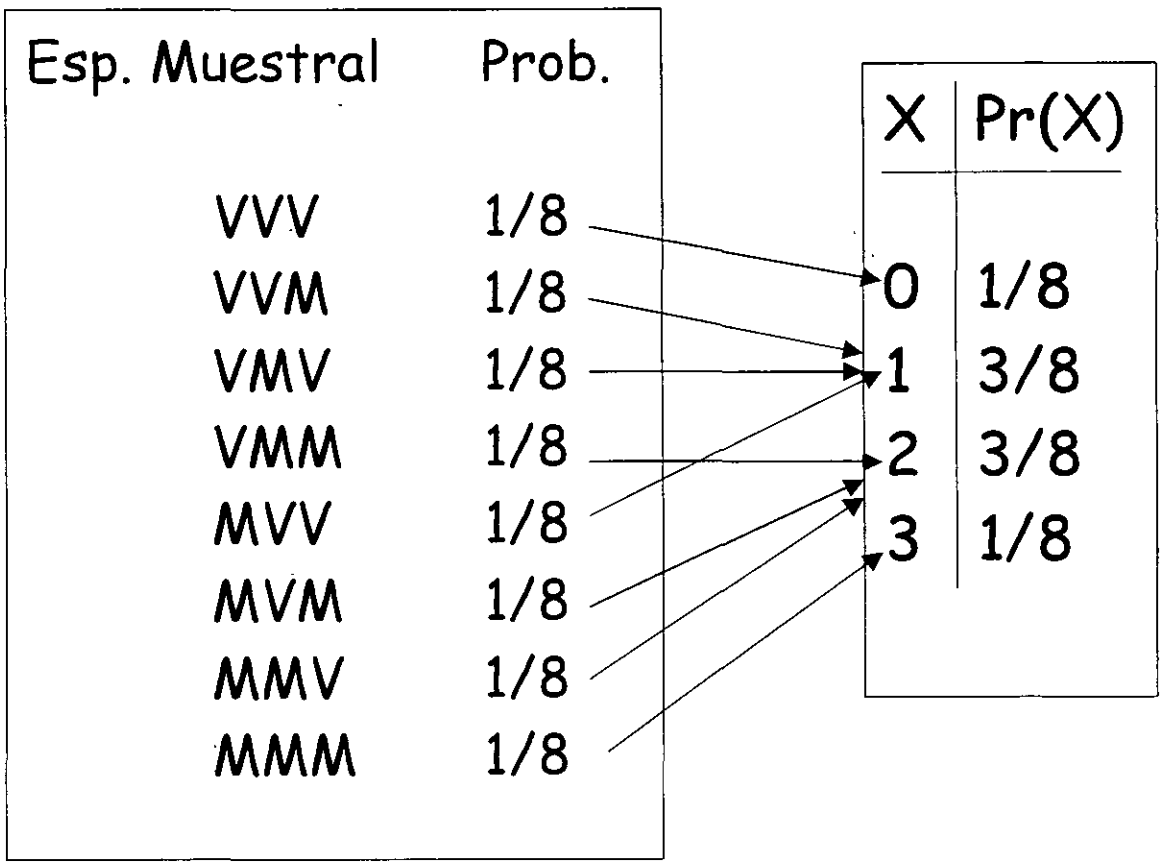
Consideremos nuestro ejemplo de la familia con tres hijos. Supongamos que esta familia está interesada en el número de mujeres que tendrá. Este es un ejemplo de una variable aleatoria que denotaremos por la letra X. Es decir,

$X \equiv$ número de hijas mujeres.

Los valores posibles de X son:

$X = \{0, 1, 2, 3\}$

Estos resultados no son igualmente probables.



Calculando las probabilidades de todos los eventos se obtiene la distribución de probabilidad de X.

Entonces, una variable aleatoria discreta X, toma valores con probabilidades especificadas por su distribución de probabilidad Pr(X).

El espacio muestral original se reduce a un nuevo y más conveniente espacio muestral.

Media y Varianza

De la misma forma que calculamos la media y la varianza de una muestra de observaciones desde las tablas de frecuencias, podemos calcular la media y la varianza de una variable aleatoria X, desde su distribución de probabilidad Pr(X).

$$\textit{Media:} \quad \mu = \sum_i X_i \Pr(X_i)$$

$$\textit{Varianza:} \quad \sigma^2 = \sum_i (X_i - \mu)^2 \Pr(X_i)$$

Distribución Binomial

Una de las variables aleatorias discretas mas comunes es la denominada BINOMIAL.

El ejemplo clásico de variable binomial es:

E = número de "caras" en varias tiradas de una moneda.

<u>Supuestos Básicos</u>

1. Suponemos que hay "n" pruebàs

2. En cada prueba, un cierto evento de interés puede ocurrir o no. Si ocurre, decimos que es un "exito" y si no ocurre es un "fracaso". Sus respectivas probabilidades son π y $(1-\pi)$, y no cambian en cada prueba.
3. Asumimos que las pruebas son estadísticamente independientes.

Entonces E, el número total de "exitos" en "n" pruebas se denomina variable binomial.

La fórmula general de la binomial cuando en cada prueba hay una probabilidad π de "exito" es la probabilidad de exactamente E "exitos" en "n" pruebas:

$$p(E) = \binom{n}{E} \pi^{E} \, (1-\pi)^{(n-E)}$$

El coeficiente binomial se define como,

$$\binom{n}{E} = \frac{n!}{E! \, (n-E)!}$$

Donde n! = n*(n-1)*(n-2)*...*3*2*1

Un punto importante es que la distribución binomial solo puede utilizarse si las pruebas son **independientes**.

Ejemplo

Supongamos que una producción de 40,000 microondas incluye 32,000 (80%) sin defectos. El departamento de control de calidad, no conociendo este número, toma una muestra aleatoria de 10 microondas para estimar la calidad total.

Cuál es la probabilidad de que en la muestra haya 5 microondas sin defectos y 5 defectuosos?

Solución

Cada uno de los 10 sucesivos microondas en la muestra puede ser considerado una "prueba", entonces "n=10". Para el primer microondas la probabilidad de "exito" (sin defectos) es 32,000/40,000=0.80.
Dependiendo de si el primer microondas fue un "exito" o no, el segundo microondas tiene una

probabilidad de "exito" de 31,999/39,999 ó 32,000/31,999 que para los fines prácticos sigue dando 80%. Es decir que la segunda prueba es prácticamente independiente de la primera. Repitiendo este argumento para los restantes microondas tenemos una distribución binomial con n=10, π=0.80 y E=5, es decir

$$p(5) = \binom{10}{5} 0.80^5 \ (1-0.80)^{(10-5)}$$

Lo que da aproximadamente 0.026.

Es decir que en una muestra aleatoria de 10 microondas, hay una probabilidad del 2.6% de tener 5 microondas sin defectos y 5 defectuosos.

En la práctica tenemos tablas con las probabilidades de la distribución binomial calculadas.

Ejemplo

Una muestra de 5 votantes de la elección presidencial de 1984 en

USA (cuando el 60% votó por los republicanos) es seleccionada en forma aleatoria.

(a) el número de votantes republicanos en la muestra puede variar de 0 a 5. Tabular su distribución de probabilidad.

(b) Calcular la media y la desviación estándar.

(c) Cuál es la probabilidad de que haya exactamente 3 votantes republicanos en la muestra?

(d) Calcule la probabilidad de que en la muestra haya una mayoría de votos republicanos y asi refleje correctamente la mayoría en la población.

(e) Grafique las respuestas (a)-(d).

Solución

(a) Cada votante seleccionado constituye una prueba. En cada prueba la probabilidad de voto a los republicanos es π=0.60. En un total de n=5

pruebas, queremos la probabilidad de E "exitos", donde E=0,1,2,3,4 y 5. Si observamos la tabla de la binomial para n=5 y π=0.60 tenemos:

E	P(E)	E*P(E)	$(E-\mu)$	$(E-\mu)^2$	$(E-\mu)^2 p(E)$
0	0.010	0	-3	9	0.090
1	0.077	0.077	-2	4	0.308
2	0.230	0.460	-1	1	0.230
3	0.346	1.038	0	0	0
4	0.259	1.036	1	1	0.259
5	0.078	0.390	2	4	0.312
	1	(b) $\mu=3$			$\sigma^2=1.20$

De la tabla $p(3) = 0.346 \cong 35\%$ (c)

y

$$p(3)+p(4)+p(5)=0.346+0.259+0.078$$
$$=0.683$$

Distribuciones Contínuas

Distribución Normal

Para muchas variables contínuas, la distribución de probabilidad es una curva con forma de campana denominada curva NORMAL o curva GAUSSIANA en honor al científico alemán Karl Gauss (1777-1855).

Esta distribución es la más común y útil de las distribuciones en ciencias sociales.

Distribución Normal Estándar

La más simple de las distribuciones normales es la distribución normal estándar. Se denomina con la letra Z y se distribuye alrededor de una media $\mu=0$ y una desviación estándar $\sigma=1$.

En general,

Cada valor de Z es el número de desviaciones estándar desde la media.

Las probabilidades se denotan por el área bajo la curva. Por ejemplo, si deseamos calcular la probabilidad de que un valor dado de Z sea mayor a 1.5, esta probabilidad corresponde al área sombreada en la figura.

Las probabilidades para Z están calculadas y tabuladas.

Ejemplos:

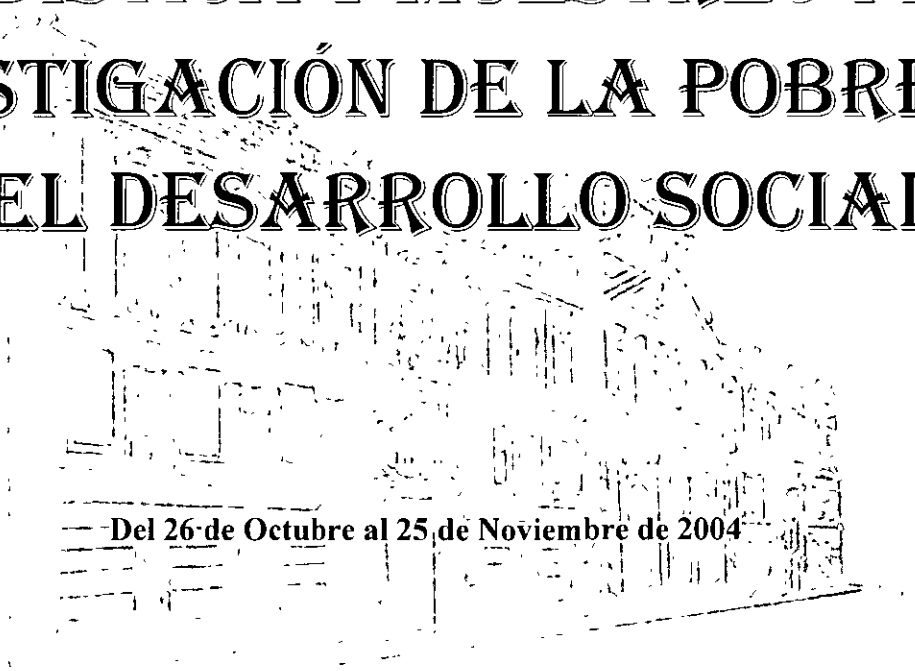
Si Z tiene una distribución normal estándar, encuentre:

a. $Pr(Z < 1.64)$
b. $Pr(Z < -1.64)$
c. $Pr(1 < Z < 1.5)$
d. $Pr(-1 < Z < 2)$
e. $Pr(-2 < Z < 2)$

FACULTAD DE INGENIERÍA UNAM
DIVISIÓN DE EDUCACIÓN CONTINUA

CURSOS INSTITUCIONALES

ESTADÍSTICA Y MUESTREO PARA INVESTIGACIÓN DE LA POBREZA Y EL DESARROLLO SOCIAL



Del 26 de Octubre al 25 de Noviembre de 2004

ANEXOS

CI - 198

Instructor: Lic. Tonatiu Suárez

GOBIERNO DEL DISTRITO FEDERAL
OCTUBRE/NOVIEMBRE DE 2004

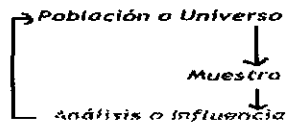
Palacio de Minería, Calle de Tacuba No. 5, Primer piso, Delegación Cuauhtémoc, CP 06000, Centro Histórico, México D.F.,
APDO Postal M-2285 ■ Tels: 5521.4021 al 24, 5623 2910 y 5623.2971 ■ Fax: 5510.0573

Existen 3 tipos de Falsedades:

1. Mentiras
2. Mentiras Detestables
3.

Estadística: Trata del diseño de experimentos o encuestas mediante muestras para obtener una cantidad determinada de información a un costo mínimo; y del uso óptimo de esta (información) se infiere con respecto a una población.

PROCESO EN ESTADÍSTICA



Estadística Moderna:

- Matemáticos de la teoría de probabilidad
- Necesidad de recopilar datos sobre bases nacionales

Estadística Descriptiva: Métodos que implican la recolección , presentación y caracterización de un conjunto de datos para describir en la forma apropiada las diversas características de ese conjunto de datos.

Estadística Inferencial: Métodos que posibilitan la estimación de una característica de una población o la forma de una decisión concerniente a una población, tan sólo con base en los resultados de una muestra.

TIPOS DE DATOS:

- Cualitativos: Categóricos – ejemplo ¿?
- Cuantitativos: #'s "Conteos (numeros enteros) y mediciones".

Números Enteros Descrito: (donde obtengo datos?)

- Publicaciones
- Encuestas casa, telefono, correo, m@il.
- Experimentos
- En donde uno trabaja

Nota: _ El tipo de pregunta depende de que tipo de entrevista es.
_ ¿ En la escala del 1 al 5 que tal le parecio la fiesta?

Grupos Focales: A partir de una pequeña muestra organizan o sacan conclusiones.

Observación Directa: Donde el que hace la encuesta observa directamente.

Los estudios estadísticos dependen de los datos:

- Cuestionario
- Preguntas Ordenadas
- Redacciones
- Opciones (no asumir nada obvio)

TIPOS DE MUESTRAS
(pedazo de población)

No Probalísticos:

- Muestra no representativa, de la cual no se puede proyectar resultados "Muestra Sesgada".

a)cuota: first 25 personas que vea de una en una.
b) Juicio: 25 mujeres que trabajen de pelo negro.
c)Trozo: Grupo de conveniencia. Encontrar al grupo de un solo. "al mismo tiempo se hace a todos".

- Estudio informal para detectar problemas.

- No son proyectables a atribuirse a otros estudios; no hay que generalizar.

Probabilísticas:

- Problema Práctico: saber cuantos son o hay saber tamaño de la población.

Basadas en la forma de elegir:

a) **Aleatorio Simple**: la muestra al azar para el estudio.

b) **Sistemática:** Para sacar "random" hay muestras sistemáticas; "1 en K" (sistema de elección).

- El uso del sistema A o B, depende de la muestra que se trabaja o población:

Poblaciones:

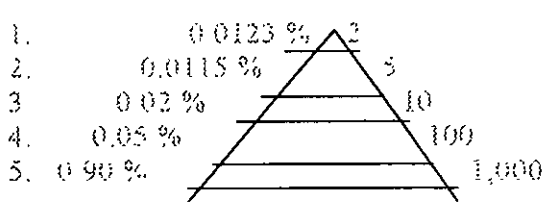
_ Aleatoria o desordenada (A y B). sentados en donde sea.

_Ordenada (B): Orden alfabético y numérico. Años, sentados por # de carnét.

_ Cíclicas (A): población de días, BTM, ROBIN, GATUBELA.

C) Estratificada: Califica un grupo sin traslape, en un solo nivel.

- Origen Étnico
- Sexo
- Nivel Social
- Salario
- Ej:



- Grupos Homogéneos
- Uno toma en cuenta el # y cantidad de "estratos"
- Al hacer categorías hay que tener cuidado.
- Al clasificar en "profesión" hay que tener en mente sus especialidades.
- Cuando uno estratifica , se obtiene el menor "error".

D) Conglomeración (Racimos):

- Grupos heterogéneos (distantes), pero que hay otros similares
- Ej: 2 secciones de estadística (am y pm) grupos diferentes, con similitud en lo que estudian.
- Ej2

Venta de Zapatos

Z	x		x		
O		x		x	
N			x	x	
A	x	x			x
S		x		x	

Resumen:

¿Se sabe el # de población? Si No

Probabilística	**No Probabilística**
• Aleatoria Simple	* Cuota
• Sistemática	* Juicio
• Estratificada	* Trozo
• Conglomerada	

Presentación de Datos:

Puede ser en diferentes métodos:

1) Métodos Explicativos: diagramas, tablas, gráficas; pero el problema es decidir cual usar, para poder explicar estadísticamente las cosas.

DIAGRAMAS:

1) Tallo – Hoja (sting – leave): **Para presentar información.**

- Dejar la última libre.
- Se ordena fácilmente.
- Si se repite el número, se puede ver fácilmente.

Ej:

Numeros: 6'0 , 7'0, 7'9, 6'4, 4'6, 5'7, 7'8, 6'1, 5'6, 6'7, 6'4, 10'3, 10'8, 12'5, 11'9.

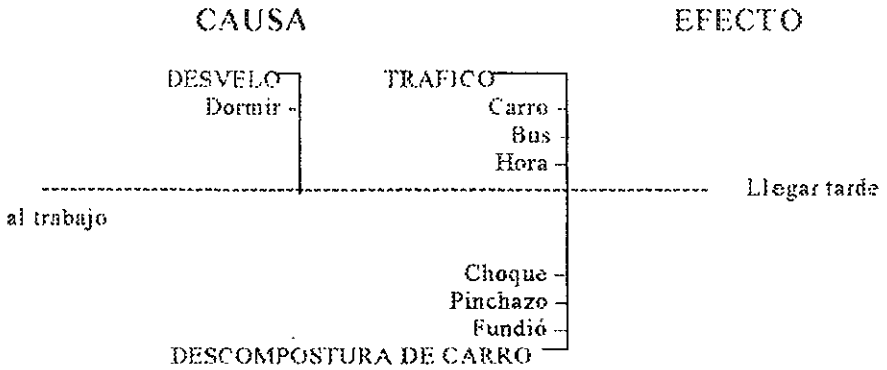
Tallo	Hoja
4	6
5	7, 6
6	0, 4, 1, 7, 4
7	0, 9, 8
10	3, 8
11	9

* Si se gira la tabla se ve una grafica, sabemos donde hay mas números

12 5

2) Esqueleto de Pescado (ISHIKAWA): >))))°> **Control de Calidad**

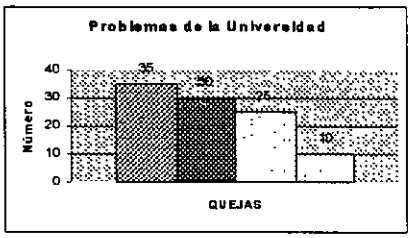
- Hay que conocer la causa – efecto, que nos puede ayudar a conocer y comprender una situación.



3) PARAPETO:

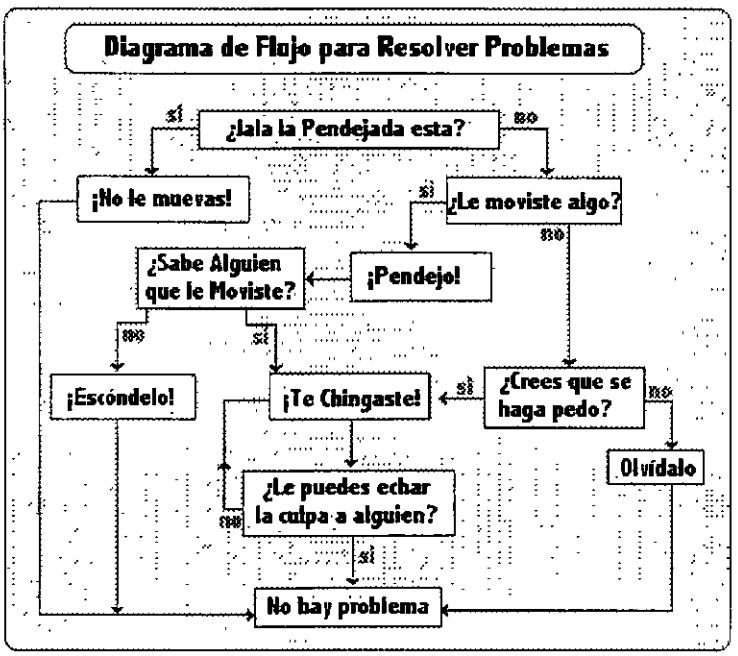
- Parapeto decía que el 80% de los problemas vienen del 20% de las causas, la mayoría de quejas vienen de pocas causas.
- Si se eliminan los 2 problemas + grandes, se elimina o se tranquilizan los secundarios.

Quejas	Número
Parqueo	35
Baños	25
Cafetería	10
Audiovisuales	30



4) Flujo "Proceso":

- Saber como interactua algo; una empresa, etc.



2) Métodos Numéricos: Medidas de Resumen descriptivas.

Se puede dar una medida de:

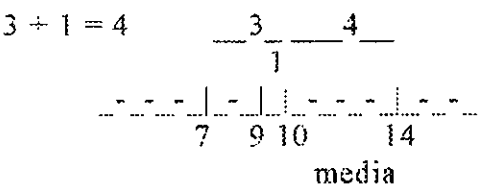
Posición	Dispersión	Forma
• ¿Cómo están los datos con referencia a algo fijo? • Buscar un punto de diferencia • Cómo se distribuyen los datos respecto a algo fijo. • Pueden ser: - *Tendencia Central*: Media, mediana, moda. - Tendencia no Central: Cuantiles, Percentiles, por centiles.	• Interesada en que tanto están separados o que dispersos están. a)Varianza b) Desviación Estándar c) Rango d) Coeficiente de Variación	• Que forma tienen los datos. • Medidos como - Sesgo - Curtosis - Asimetría

TENDENCIA CENTRAL:

A) MEDIA: (promedio)

Ej: 7, 9, 14 Media = (7 + 9 + 14) / 3 = 10

3 + 1 = 4



- La suma de los números de la Izquierda es igual al de la suma de la derecha.-
- Punto de equilibrio las diferencias de Izquierda y derecha.-
- "Diferente camino, mismo kilómetros"
- VENTAJAS. Toma en cuenta todos los datos
- DESVENTAJA:
  - 9, 10, 11 estan más pegados, es decir, la medida es la misma pero diferente distancia.
  - El Promedio puede MENTIR, ya que hay gente que sube o baja el promedio. **"Jordan / Salario"**
  - **OJO**: **Ingreso percápita, producto nacional bruto.**

B) MEDIANA:

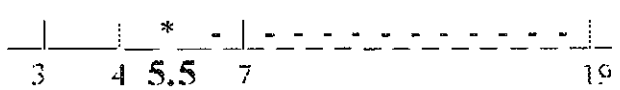
- Lo que esta en medio (el 50% a la Iz. y el 50% a la der.)
- Los datos deben estar ordenados.
- VENTAJA: Para saber que esta en medio.
- DESVENTAJA: No toma en cuenta todos los datos, solo los de en medio.

EJ: 7, **9**, 14 El 9 es la mediana.

7, 3, 19, 4, 22 luego ordenados: 3, 4, 7, 19

* Si no tiene mitad impar, (4 + 7) / 2 = 5.5

Mediana = 5.5



EJ:

Salario Medio

Están abajo _____|_____ Están arriba
4, 000

C) MODA:

- Lo que más se repite.-
- VENTAJA: Única con la que se puede calcular datos Cualitativos. Busca lo común.

- EJ:

7 verde La **moda** es verde por ser común.
3 rojo
1 azul

TENDENCIA NO CENTRAL:

a) PARTES POR MIL: Califica en mil.

b) DECILES: Califica del 1 – 10 (2% = 20%)

c) CUANTILES: Califica por 100.

d) PERCENTIL:

- VENTAJA: Disfraza la Información.
- Cambia calificaciones o notas de números a porcentajes.
- El porcentaje dice la posición que ocupa sobre los demás, y no da la nota.
- Forma más fácil para dar de manera positiva un resultado.

EJ: Si en un examen hay 100 niños.

Nota		Percentil	
62	D	100 %	
61		99 %	
61		99 %	
55		33%	significa que esta sobre el 33 % de los demás.-

PARA COMPLETAR necesitamos las medidas de:

DISPERSIÓN

1) **RANGO:** restar el valor + grande, menos el valor + pequeño.

7, 9, 11, 15 $R = 15 - 7 = 8$

11, 10, 9 $R = 11 - 9 = 2$

* Pero hay que ver que tan separados están los datos.

2) **VARIANZA**: es el promedio menos la media, todo al cuadrado, dividido # de datos.

$\sum (D.S.^2)/ n$

Datos		**Varianza**	
7	$7 - 10 = -3$	$-3^2 = 9$	
8	$8 - 10 = -2$	$-2^2 = 4$	**Varianza**
15	$15 - 10 = \underline{5}$	$\underline{5^2 = 25}$	**$38/3 = 12.66^2$**
	0	38	

Para saber la dimensional de algo como el $$, si se da 12.66^2 no se entendería, por eso la D.S. le quita el ^{2} poniéndole una raíz, quedando así $ 12.66

Dispersión o Separación: Dependiendo para que se use es buena o mala. (lenguaje , idioma "español alemán portugués).

3) **DESVIACIÓN STÁNDAR**: Datos menos la media.
Toma en cuenta todos los datos.
Es la raíz de la varianza.

$\sqrt{38/3} = 5$

4) **COEFICIENTE DE VARIACIÓN**: CV = (D.S. / media) * 100

- VENTAJAS: Comparar cosas muy distintas; que no tengan sentido de relación y da solo dimensiones.

EJ: El precio promedio de

ORO	es $125 onz.
PAPEL	es $1,400 ton.
CAFÉ	es $250 qq.

Comparar precios d lo mismo por 5 día o más.-

Se calcula la Desviación Estándar:

D.S.		**CV**
$10	(10/125)*100 =	8 %
$150	(150/1400)*100 =	10.7 %
$ 5	(5/250)*100 =	3 %

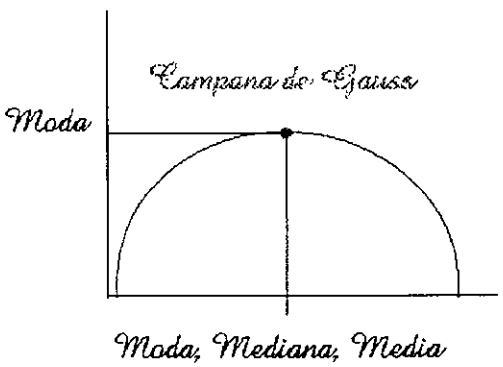
Significa que el Café por tener el 3 % , es el elemento que menos cambios tiene. Sería mas seguro invertir en este.

FORMA:

Sabemos hacer la Forma, teniendo ya la tendencia y descripción; es decir cuando ya tenemos la POSICIÓN (tendencia central, mediana, moda, media, tendencia no central); y tenemos el RANGO (recorrido indescriptible, varianza, desviación estándar, coeficiente de variación); entonces proseguimos a saber la forma.

Puede ser Asimétrica, Curtosis y Sesgo.

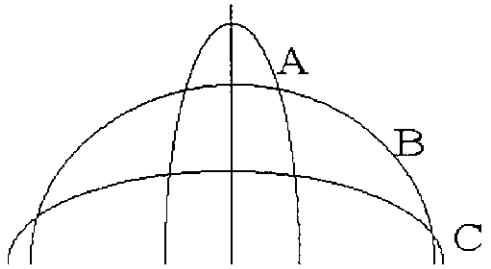
A) Simétrica: es una curva simétrica.



Media: 20
Moda: 20.01
Mediana: 21.1

Si no son iguales , no es simétrica.

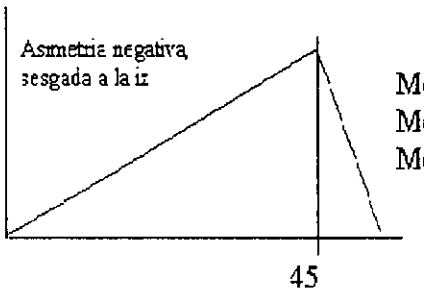
B) Desviación Standard: Que tan ancha es la curva, pueden tener curvas simétricas pero con dif. Desviación Standard.



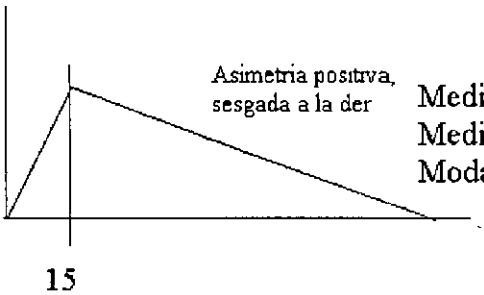
Todas tienen misma media, moda y mediana.-

A: menor D.S.
C: mayor D.S.

C) Asimétrico:



Media: 15
Mediana: 30
Moda: 45



Media: 45
Mediana: 30
Moda: 15

D) Curtosis: Medida de que tan puntiagudo es algo. (este dato lo da la compu).

PLATICURTICA	MASOKURTICA	LEPTOKURTICA
no hay moda	Simetria normal	Puntiaguda
menor a 3	es = a 3	mayor a 3

Características: se aplica a curvas simétricas debe ser menor o mayor o = a 3.

"TEOREMA DE TCHEVICHEFF"

La Desviación Estándar, la media y forma, sirven para averiguar la mayoría de los datos..

K	$1 - 1/K^2$
1	0
2	¾
3	8/9

| 69 -+ **2** (20) | 75% de datos
| 69 + - 40 |

[29, 109]

| 69 -+ **3** (20) |
| 69 + - 60 |

[9, 129] 89% de datos

Media: 69
D. S.: 20

Fórmula:
Media +- # (D S.)

Regla Empírica: Solo se utiliza en la simétrica, por que la media, moda y mediana son iguales; cubre un buen % de datos.

Fórmula: **Mediana +- (D.S.)**