UNIVERSIDAD DE CHILE
VICERRECTORÍA DE ASUNTOS ACADÉMICOS
DEPARTAMENTO DE EVALUACIÓN, MEDICIÓN Y REGISTRO EDUCACIONAL

NOCIONES BÁSICAS DE ESTADÍSTICA UTILIZADAS EN EDUCACIÓN

SANTIAGO, septiembre de 2008

NOCIONES BÁSICAS DE ESTADÍSTICA UTILIZADAS EN EDUCACIÓN

¿QUÉ ES LA ESTADÍSTICA?

La estadística es una disciplina que diseña los procedimientos para la obtención de los datos, como asimismo proporciona las herramientas que permiten extraer la información.

Los métodos estadísticos constituyen uno de los medios por los que el hombre trata de comprender la generalidad de la vida. Los métodos objetivos y controlados que permiten abstraer grupos de tendencias de muchos individuos aislados, son llamados métodos estadísticos. Estos son fundamentalmente los mismos, independientemente de que se apliquen en el análisis de fenómenos físicos, en el estudio de mediciones educacionales, en el estudio de datos provenientes de experimentos biológicos, o del análisis cuantitativo del material en economía

Los ejemplos de estas nociones básicas son tomados de aquellos usados en educación y principalmente en la etapa de término de la Educación Media y su postulación a la Educación Superior.

Estadística descriptiva

La estadística descriptiva es un conjunto de procedimientos que tienen por objeto presentar masas de datos por medio de tablas, gráficos y/o medidas de resumen. De acuerdo a lo anterior, la estadística descriptiva es la primera etapa a desarrollar en un análisis de información.

Tablas de Frecuencias:

Una forma de presentar ordenadamente un grupo de observaciones, es a través de tablas de distribución de frecuencias. La estructura de estas tablas depende de la cantidad y tipo de variables que se están analizando, siendo las más simples las que se refieren a una variable.

EJEMPLO : Se tienen las notas de una prueba de matemática para 1000 alumnos de enseñanza media de un determinado colegio. Se resume la información en la siguiente tabla de frecuencia.

NOTA	FRECUENCIA
1,2	1
1,4	2
1,6	3
1,8	8
2,0	15
2,2	18
2,4	19
2,6	22
2,8	25
3,0	26
3,2	28
3,4	31
3,6	35
3,8	38
4,0	45

NOTA	FRECUENCIA
4,2	46
4,4	48
4,6	52
4,8	58
5,0	60
5,2	56
5,4	54
5,6	51
5,8	50
6,0	46
6,2	44
6,4	40
6,6	32
6,8	31
7,0	18

En una tabla se pueden distinguir los siguientes tipos de frecuencias:

Frecuencia Absoluta : Es el número de repeticiones que presenta una observación. Se denota por n_i

Frecuencia Relativa : Es la frecuencia absoluta dividida por el número total de datos. Se denota por f_i

Frecuencia Absoluta Acumulada : Es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado. La última frecuencia absoluta acumulada es igual al número de casos. Se denota por N_i

Frecuencia Relativa Acumulada : Es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos. Se denota por F_i

Para el ejemplo propuesto se determinaron las distintas frecuencias, las que se muestran en la siguiente tabla:

NOTA	FREC. ABSOLUTA	FREC. ABSOLUTA ACUMULADA	FREC. RELATIVA	FREC RELATIVA ACUMULADA
1,2	1	1	0,001	0,00
1,4	2	3	0,002	0,00
1,6	3	6	0,003	0,01
1,8	8	14	0,008	0,01
2,0	14	28	0,014	0,03
2,2	18	46	0,018	0,05
2,4	19	65	0,019	0,07
2,6	22	87	0,022	0,09
2,8	25	112	0,025	0,11
3,0	26	138	0,026	0,14
3,2	27	165	0,027	0,17
3,4	31	196	0,031	0,20
3,6	35	231	0,035	0,23
3,8	38	269	0,038	0,27
4,0	45	314	0,045	0,31
4,2	46	360	0,046	0,36
4,4	48	408	0,048	0,41
4,6	52	460	0,052	0,46
4,8	58	518	0,058	0,52
5,0	60	578	0,060	0,58
5,2	56	634	0,056	0,63
5,4	54	688	0,054	0,69
5,6	51	739	0,051	0,74
5,8	50	789	0,050	0,79
6,0	46	835	0,046	0,84
6,2	44	879	0,044	0,88
6,4	40	919	0,040	0,92
6,6	32	951	0,032	0,95
6,8	31	982	0,031	0,98
7,0	18	1000	0,018	1
TOTAL	**1000**			

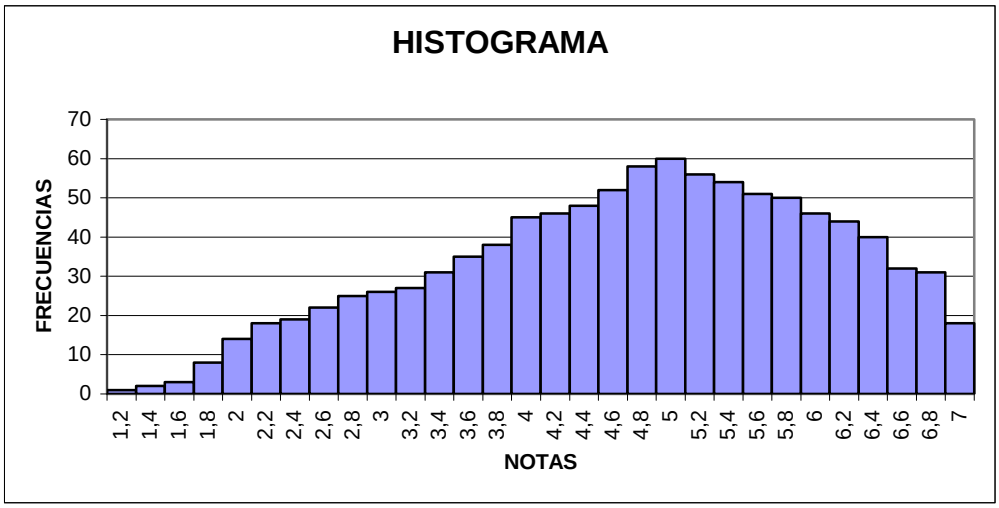
Nota:

Si la frecuencia relativa y relativa acumulada la multiplicamos por 100, los valores obtenidos representan porcentajes, lo que facilita la interpretación de los datos.

De esta tabla se pueden sacar conclusiones como:

- 45 alumnos obtuvieron nota 4,0
- 578 alumnos obtuvieron nota inferior o igual a 5,0
- El 1,8 % de los alumnos obtuvo nota 7,0
- El 31 % obtuvo nota 4.0 o inferior a ésta, mientras que el 69% obtuvo una nota superior a 4,0

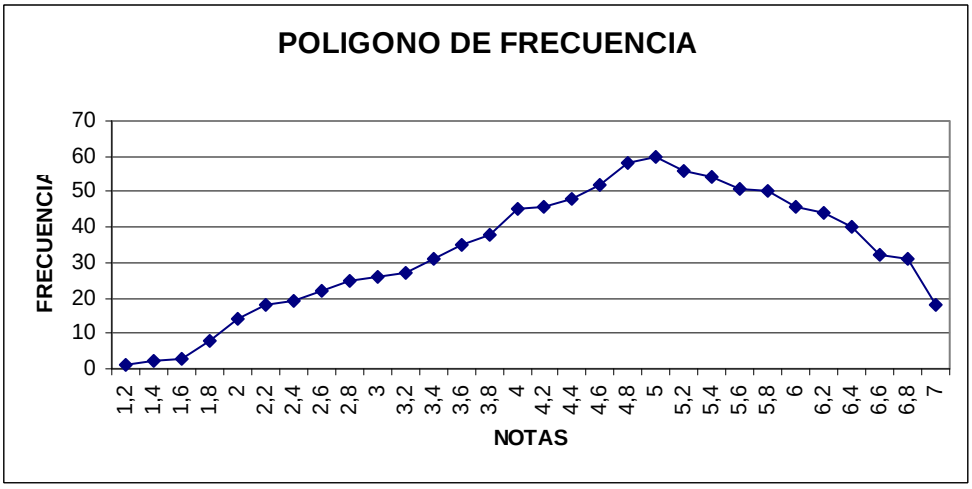
Esta información también puede ser representada en forma gráfica como se muestra a continuación:



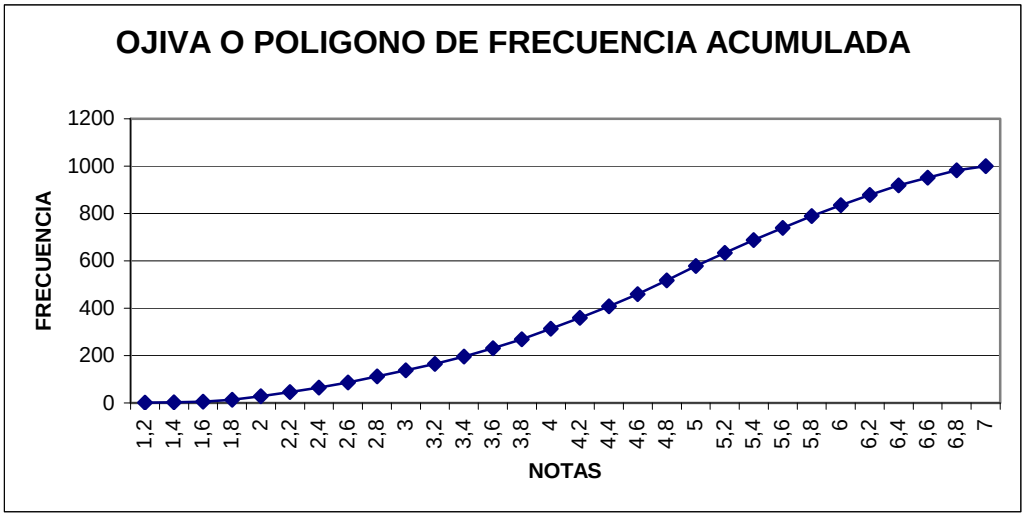
En el histograma se observa gráficamente la distribución de las notas de la prueba, y que los puntos más altos están en las notas 4,8; 5,0 y 5,2 las que coinciden con las frecuencias más altas de la tabla.

Otra forma de representar los datos es a través de un polígono de frecuencias que es un gráfico de puntos en el cual se muestra la distribución dibujada punto por punto representando los valores específicos de la variable bajo estudio.

En el ejemplo se puede observar que se representan los 30 valores que toman las notas. La frecuencia más alta de alumnos la alcanza la nota 5,0



La ojiva o polígono de frecuencia acumulada nos muestra justamente las frecuencias acumuladas. En nuestro ejemplo la Ojiva nos dice que hay alrededor de 800 alumnos que obtuvieron nota 6 o menos en la prueba de matemática.



MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son valores numéricos que tienden a localizar la parte central de un conjunto de datos.

Nos dan un centro de la distribución de frecuencias, es un valor que se puede tomar como representativo de todos los datos. Hay diferentes modos para definir el "centro" de las observaciones en un conjunto de datos. A continuación se presentan los más usados.

La Media aritmética: también denominada promedio, es la que se utiliza principalmente y se define como la suma de los valores de todas las observaciones divididas por el número total de datos. Se representa por $\bar{x}$ o por la letra μ según se calcule en una muestra o en la población, respectivamente.

NOTA	FREC. ABSOLUTA	FREC. ABSOLUTA ACUMULADA	FREC. RELATIVA %	FREC RELATIVA ACUMULADA %	$x_i * n_i$
1,2	1	1	0,1	0,1	1,2
1,4	2	3	0,2	0,3	2,8
1,6	3	6	0,3	0,6	4,8
1,8	8	14	0,8	1,4	14,4
2,0	14	28	1,4	2,8	28,0
2,2	18	46	1,8	4,6	39,6
2,4	19	65	1,9	6,5	45,6
2,6	22	87	2,2	8,7	57,2
2,8	25	112	2,5	11,2	70,0
3,0	26	138	2,6	13,8	78,0
3,2	27	165	2,7	16,5	86,4
3,4	31	196	3,1	19,6	105,4
3,6	35	231	3,5	23,1	126,0
3,8	38	269	3,8	26,9	144,4
4,0	45	314	4,5	31,4	180,0
4,2	46	360	4,6	36,0	193,2
4,4	48	408	4,8	40,8	211,2
4,6	52	460	5,2	46,0	239,2
4,8	58	518	5,8	51,8	278,4
5,0	60	578	6,0	57,8	300,0
5,2	56	634	5,6	63,4	291,2
5,4	54	688	5,4	68,8	291,6
5,6	51	739	5,1	73,9	285,6
5,8	50	789	5,0	78,9	290,0
6,0	46	835	4,6	83,5	276,0
6,2	44	879	4,4	87,9	272,8
6,4	40	919	4,0	91,9	256,0
6,6	32	951	3,2	95,1	211,2
6,8	31	982	3,1	98,2	210,8
7,0	18	1000	1,8	100,0	126,0
TOTAL	**1000**				**4717,0**

La fórmula para calcular el promedio es entonces:

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

En el ejemplo dado que se tiene una distribución de frecuencias el promedio se calcula por:

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

Donde:

n_i : Representa la frecuencia absoluta de cada grupo.
x_i : Corresponde a la clase de cada grupo.
n : Cantidad total de datos.

Aplicando la fórmula se obtiene:

$$\bar{x} = \frac{4717}{1000} = 4{,}717$$

Por lo tanto, la media de notas de los alumnos en la prueba de matemática es de 4,7

Propiedades de la media aritmética:

- Puede ser calculada en distribuciones con escala relativa e intervalar.
- Todos los valores son incluidos en el cálculo de la media.
- Una serie de datos solo tiene una media.
- Es una medida muy útil para comparar dos o más poblaciones.
- Es la única medida de tendencia central donde la suma de las desviaciones de cada valor respecto a la media es igual a cero. Por lo tanto, podemos considerar a la media como el punto de balance de una serie de datos.

Desventajas de la media aritmética

- Si alguno de los valores es extremadamente grande o extremadamente pequeño, la media no es el promedio apropiado para representar la serie de datos.

- No se puede determinar si en una distribución de frecuencias hay intervalos de clase abiertos.

Observaciones:

1. A veces se interpreta erróneamente a la media como aquel valor que es típico, o que se esperaría que la mayoría de las personas tuvieran. Esta interpretación puede ser bastante absurda en algunos casos, por ejemplo, cuando se calcula la media de hijos en un grupo de mujeres, se obtiene que es de 2.3 niños y, obviamente, no se puede esperar encontrar una madre con exactamente 2.3 niños. Todo lo que la cifra dice, es que si dividimos el número total de niños de las mujeres consideradas por el número de mujeres, el resultado es 2.3 niños por mujer. Esto puede ser un conocimiento útil en la comparación de tamaño de familia, de dos o más grupos, pero no sugiere que alguna mujer tenga 2.3 niños.

2. Otras veces se piensa que la media aritmética tiene la característica que la mitad de las observaciones es menor o igual que la media. Este concepto es totalmente errado en algunos casos, por ejemplo, si la distribución es asimétrica a la derecha, como puede ser la distribución de salarios donde hay muchas personas que ganan poco y hay pocas personas que ganan mucho, la media aritmética resultará mucho más grande de lo que uno esperaría encontrar, si se piensa que el valor central debe ser tal que la mitad de las personas tiene un salario inferior a él y la otra mitad un salario superior. Esto se debe a la presencia de unos pocos valores excesivamente grandes que al tener demasiada influencia en el valor de la media aritmética hacen que ella se ubique en una posición más extrema a la esperada. En consecuencia debería pensarse en otras medidas para evaluar un valor central con esta característica.

Mediana:

Se define como el valor que deja igual número de observaciones a su izquierda que a su derecha, es decir, divide al conjunto de datos en dos partes iguales y se denota por Me.

Si los datos no están tabulados la mediana se determina, ordenando las observaciones de menor a mayor y determinando el valor central. Si la cantidad de datos es impar, la mediana se representa justamente por ese valor. En cambio, si la cantidad es par, la mediana es el promedio de los datos centrales.

Si los datos están agrupados la mediana se calcula observando los siguientes pasos: primero se debe determinar cuanto es n/2, luego se verá en cuál intervalo estará contenido este valor. Una vez ubicado el intervalo que lo contiene se procede a reemplazar en la siguiente fórmula:

$$Me = Li + \left[\frac{\frac{n}{2} - (N_i)_{Me-1}}{(n_i)_{Me}} \right] a$$

Donde:

L_i : Es el límite inferior de la clase que contiene la mediana.
$(N_i)_{Me-1}$: Frecuencia absoluta acumulada de la clase que precede (antes) a la clase que contiene a la mediana.
$(n_i)_{me}$: Número de observaciones en la clase que contiene a la mediana.
n_i : Número de observaciones.
a : Amplitud del intervalo seleccionado.

Reemplazando los valores del ejemplo en la fórmula se obtiene:

Para nuestro ejemplo

$$Me = 4{,}8 + \left[\frac{\frac{1000}{2} - 460}{518} \right] 0 = 4{,}8$$

En este caso los datos no están agrupados en intervalo, por lo tanto a = 0

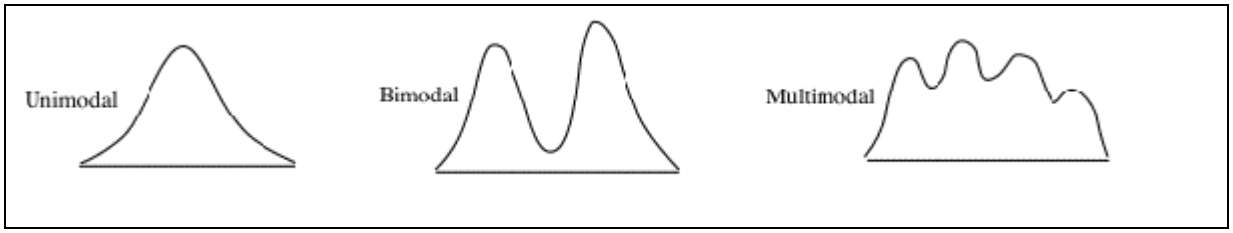
La mediana de los alumnos que rindieron la prueba de matemáticas es de 4,8

Propiedades:

- No le afectan las observaciones extremas.
- Es fácil de calcular.
- Es siempre un valor de la variable.
- La mediana divide el área total del histograma en dos porciones iguales.

Moda:

Es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor. Puede haber más de una moda en una distribución. Se denota por Mo.



En la tabla de frecuencias del ejemplo, se observa claramente que la moda de los alumnos que rindieron la prueba de matemática es 5.

MEDIDAS DE POSICIÓN

Las medidas de posición dividen un conjunto ordenado de datos en grupo con la misma cantidad de individuos.

Percentiles:

Son 99 valores que dividen en cien porciones iguales el conjunto de datos ordenados. Ejemplo, el percentil de orden 15 deja por debajo al 15% de las observaciones, y por encima queda el 85%

Cuando los datos están agrupados en una tabla de frecuencias, se calculan mediante la fórmula:

$$P_k = L_i + \frac{k\left(\frac{n}{100}\right) - N_{i-1}}{n_i} * a$$

con k= 1,2,3,... 99

Donde

L_i : Límite real inferior de la clase del percentil *k.*

n : Cantidad total de datos.

N_{i-1} : Frecuencia acumulada de la clase que antecede a la clase del percentil *k.*

n_i : Frecuencia de la clase del percentil *k.*

a : Longitud del intervalo de la clase del percentil *k.*

Para el ejemplo calcularemos el percentil 87

$$P_{87} = 6{,}2 + \frac{87\left(\frac{1000}{100}\right) - 835}{44} * 0 = 6{,}2$$

El 87% de los alumnos obtuvieron una nota igual o inferior a 6,2

En la publicación de los resultados de pruebas del examen de selección los puntajes se expresan en puntaje estándar asociándose al percentil correspondiente.

Deciles:

Son los nueve valores que dividen al conjunto de datos ordenados en diez porciones iguales, son también un caso particular de los percentiles, pues corresponden a los percentiles 10, 20, 30, 40, 50, 60, 70, 80 y 90.

Para datos agrupados los deciles se calculan mediante la fórmula.

$$D_k = L_i + \frac{k\left(\frac{n}{10}\right) - N_{i-1}}{n_i} * a$$

con k= 1,2,3,... 9

Donde:

L_i : Límite real inferior de la clase del decil *k.*

n : Cantidad total de datos.

N_{i-1} : Frecuencia acumulada de la clase que antecede a la clase del decil *k.*

n_i : Frecuencia de la clase del decil *k.*

a : Longitud del intervalo de la clase del decil *k.*

Para el ejemplo calcularemos el decil 4

$$D_4 = 4{,}4 + \frac{4\left(\frac{1000}{10}\right) - 360}{48} * 0 = 4{,}4$$

El 40% de los alumnos obtuvieron una nota igual o inferior a 4,4

Cuartiles:

Son los tres valores que dividen al conjunto de datos ordenados en cuatro porciones iguales, son un caso particular de los percentiles, correspondiendo a los percentiles 25, 50 y 75.

- El primer cuartil Q1 es el valor de la variable que deja a la izquierda el 25% de la distribución.

- El segundo cuartil Q2 (la mediana), es el valor de la variable que deja a la izquierda el 50% de la distribución.

- El tercer cuartil Q3 es el valor de la variable que deja a la izquierda el 75% de la distribución.

Para el ejemplo, se tienen los siguientes cuartiles

Q_1: $\frac{n}{4}$ = 250 Primero $N_i \rangle n/4$ = 269 ; luego Q_1 =3,8

El 25% de los alumnos obtuvieron una nota igual o inferior a 3,8

Q_2: $\frac{2n}{4}$ = 250 Primero $N_i \rangle 2n/4$ = 518 ; luego $Q_{2=}$ 4,8

El 50% de los alumnos obtuvieron una nota igual o inferior a 4,8

Q_3: $\frac{3n}{4}$ = 250 Primero $N_i \rangle 3n/4$ = 789 ; luego $Q_{3=}$ 5,8

El 75% de los alumnos obtuvieron una nota igual o inferior a 5,8, o bien, el 25% de los alumnos tuvieron nota superior a 5,8.

Quintiles

Son los cuatro valores que dividen al conjunto de datos ordenados en cinco porciones iguales, son un caso particular de los percentiles, correspondiendo a los percentiles 20, 40, 60, 80.

- El primer quintil es el valor de la variable que deja a la izquierda el 20% de la distribución.
- El segundo quintil es el valor de la variable que deja a la izquierda el 40% de la distribución.
- El tercer quintil es el valor de la variable que deja a la izquierda el 60% de la distribución.
- El cuarto quintil es el valor de la variable que deja a la izquierda el 80% de la distribución.

$$K_k = L_i + \frac{k\left(\frac{n}{5}\right) - N_{i-1}}{n_i} * a$$

con k = 1, 2, 3, 4

Donde:

L_i : Límite real inferior de la clase del quintil *k.*

n : Número de datos.

N_{i-1} : Frecuencia acumulada de la clase que antecede a la clase del quintil *k.*

n_i : Frecuencia de la clase del quintil *k.*

a : Longitud del intervalo de la clase del quintil *k.*

Para el ejemplo calcularemos el quintil 3

$$K_3 = 5{,}2 + \frac{3\left(\frac{1000}{5}\right) - 578}{56} * 0 = 5{,}2$$

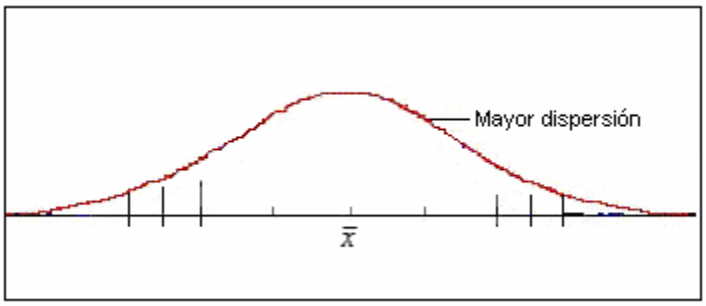
El 60% de los alumnos obtuvieron una nota igual o inferior a 5,2 o bien, el 40% de los alumnos tuvieron nota superior a 5,2

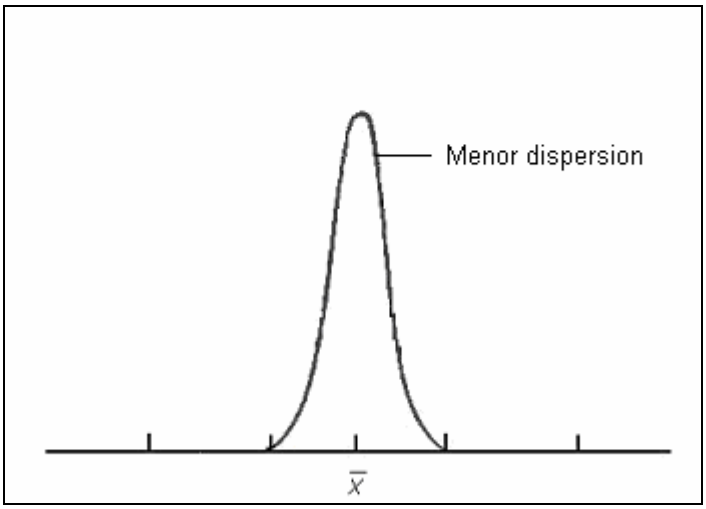
MEDIDAS DE DISPERSIÓN

Las medidas de dispersión indican la mayor o menor concentración de los datos con respecto a las medidas de centralización

Desviación estándar: también llamada **desviación típica**, es una medida de dispersión usada en estadística que nos dice cuánto tienden a alejarse los valores puntuales del promedio en una distribución. Específicamente, la desviación estándar es "el promedio de la distancia de cada punto respecto del promedio". Se suele representar por una S o con la letra sigma, σ, según se calcule en una muestra o en la población.

Una desviación estándar grande indica que los puntos están lejos de la media, y una desviación pequeña indica que los datos están agrupados cerca de la media.





La fórmula para calcular la desviación estándar es:

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

En el ejemplo dado que se tiene una distribución de frecuencias, la desviación se calcula por:

$$S = \sqrt{\frac{\sum_{i=1}^{K} n_i x_i^2 - \frac{\left(\sum_{i=1}^{K} n_i x_i\right)^2}{n}}{n-1}}$$

NOTA	FREC. ABSOLUTA	FREC. ABSOLUTA ACUMULADA	FREC. RELATIVA %	FREC RELATIVA ACUMULADA %	$x_i * n_i$	$x_i^2 * n_i$
1,2	1	1	0,1	0,1	1,2	1,44
1,4	2	3	0,2	0,3	2,8	3,92
1,6	3	6	0,3	0,6	4,8	7,68
1,8	8	14	0,8	1,4	14,4	25,92
2,0	14	28	1,4	2,8	28,0	56,00
2,2	18	46	1,8	4,6	39,6	87,12
2,4	19	65	1,9	6,5	45,6	109,44
2,6	22	87	2,2	8,7	57,2	148,72
2,8	25	112	2,5	11,2	70,0	196,00
3,0	26	138	2,6	13,8	78,0	234,00
3,2	27	165	2,7	16,5	86,4	276,48
3,4	31	196	3,1	19,6	105,4	358,36
3,6	35	231	3,5	23,1	126,0	453,60
3,8	38	269	3,8	26,9	144,4	548,72
4,0	45	314	4,5	31,4	180,0	720,00
4,2	46	360	4,6	36,0	193,2	811,44
4,4	48	408	4,8	40,8	211,2	929,28
4,6	52	460	5,2	46,0	239,2	1100,32
4,8	58	518	5,8	51,8	278,4	1336,32
5,0	60	578	6,0	57,8	300,0	1500,00
5,2	56	634	5,6	63,4	291,2	1514,24
5,4	54	688	5,4	68,8	291,6	1574,64
5,6	51	739	5,1	73,9	285,6	1599,36
5,8	50	789	5,0	78,9	290,0	1682,00
6,0	46	835	4,6	83,5	276,0	1656,00
6,2	44	879	4,4	87,9	272,8	1691,36
6,4	40	919	4,0	91,9	256,0	1638,40
6,6	32	951	3,2	95,1	211,2	1393,92
6,8	31	982	3,1	98,2	210,8	1433,44
7,0	18	1000	1,8	100,0	126,0	882,00
TOTAL	**1000**				**4717,0**	**23970,12**

Reemplazando en la fórmula los valores del ejemplo se obtiene:

$$S^2 = \frac{23970{,}12 - \frac{4717^2}{1000}}{999} = 1{,}72$$

$$S = \sqrt{S^2} = 1{,}3114$$

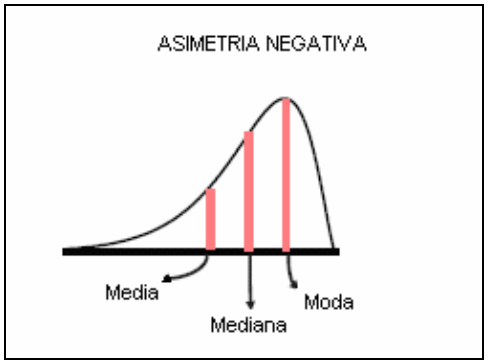
La desviación estándar en las notas de la prueba de matemática es de 1,3.

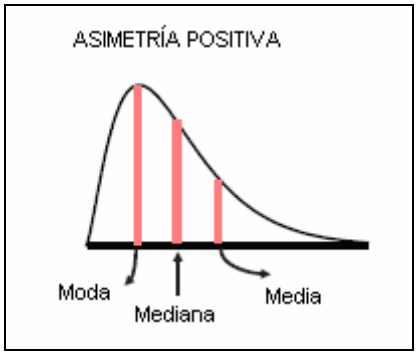
MEDIDAS DE FORMA

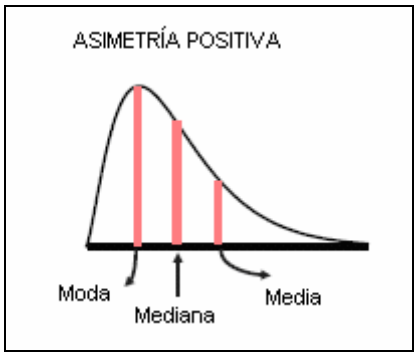
Las distribuciones pueden tener diferentes formas, y una manera de caracterizar la forma es observar su simetría. Una distribución de frecuencias puede ser simétrica o asimétrica. Para saber si es simétrica tenemos que tomar una referencia, es decir, ver respecto a qué es simétrica. El coeficiente de asimetría de Pearson, mide la desviación de la simetría, expresando la diferencia entre la media y la mediana con respecto a la desviación estándar del grupo de mediciones.
Su fórmula es:

$$\boxed{A_s = \frac{3(\bar{x} - Me)}{S_x}}$$

- Si $A_s = 0$ diremos que la distribución es simétrica, en ese caso las desviaciones a la derecha y a la izquierda de la media se compensan.
- Si $A_s < 0$ diremos que es asimétrica negativa ya que la mayoría de las observaciones están a la derecha de la proyección de la media.
- Si $A_s > 0$ diremos que es asimétrica positiva ya que la mayoría de las observaciones están a la izquierda de la proyección de la media.





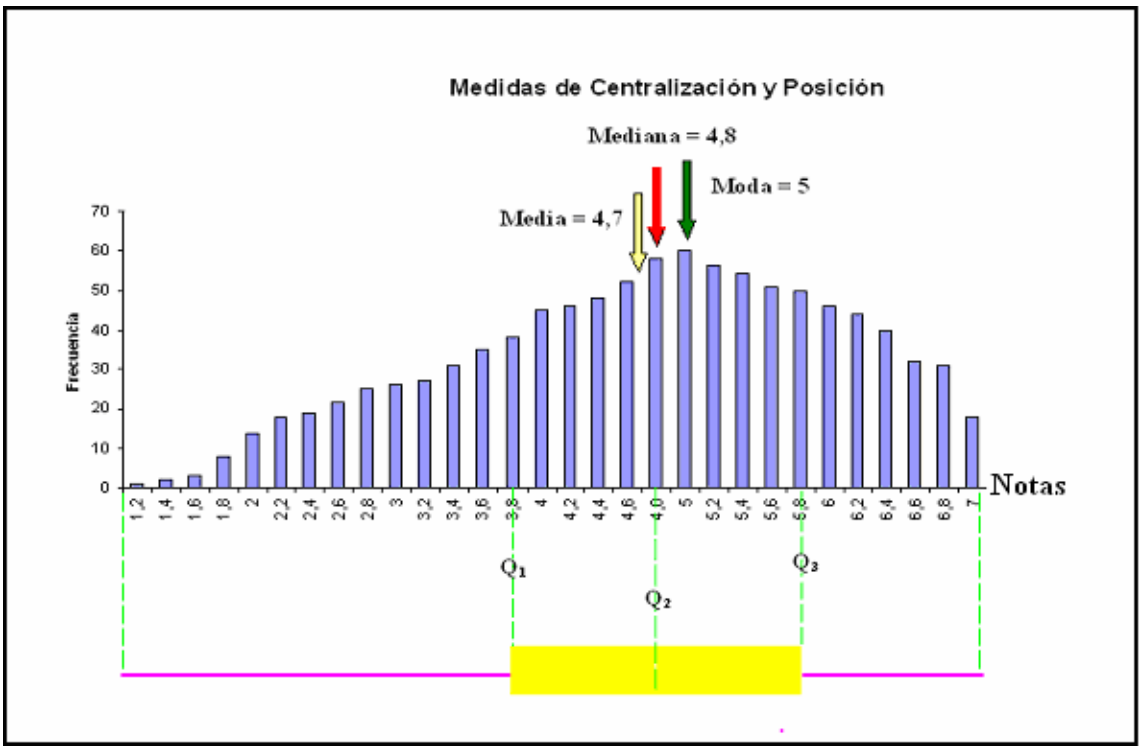


Reemplazando en la fórmula los valores del ejemplo se obtiene:

$$A_s = \frac{3(4{,}717 - 4{,}8)}{1{,}3114} = \text{-0,1898}$$

Por lo tanto, las notas de los alumnos tienen una distribución ligeramente asimétrica negativa.

En el siguiente histograma se pueden observar las medidas de tendencia central y posición de nuestro ejemplo, además, se puede ver fácilmente que la distribución es asimétrica negativa.

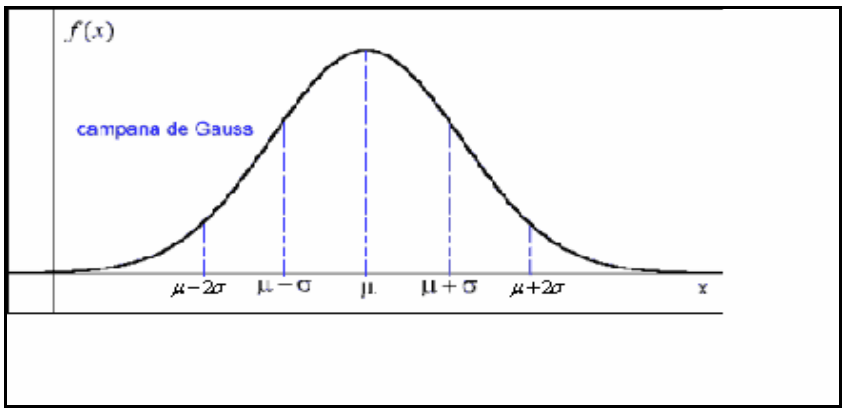


DISTRIBUCIÓN NORMAL

La distribución normal es una de las distribuciones más usadas e importantes. Se ha desenvuelto como una herramienta indispensable en cualquier rama de la ciencia, la industria y el comercio.

Muchos eventos reales y naturales tienen una distribución de frecuencias cuya forma es muy parecida a la distribución normal, llamada también campana de Gauss por su forma acampanada.

La forma de la campana de Gauss depende de los parámetros μ y σ. La media indica la posición de la campana, de modo que para diferentes valores de la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de S, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.



La distribución de probabilidad normal y su curva tiene las siguientes características:

- La curva normal tiene forma de campana. La media, la moda y la mediana de la distribución son iguales y se localizan en el centro de la distribución.

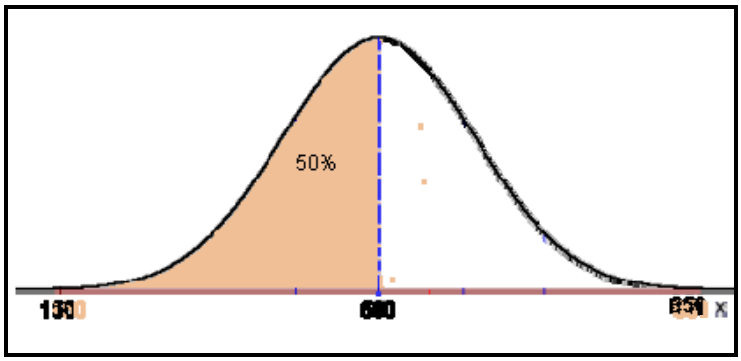
- La distribución de probabilidad normal es simétrica alrededor de su media. Por lo tanto, la mitad del área bajo la curva está antes del punto central y la otra mitad después, es decir, la mitad de curva tiene un área de 0,5. El área total bajo la curva es igual a 1.

- La escala horizontal de la curva se mide en desviaciones estándar.

- La forma y la posición de una distribución normal dependen de los parámetros μ y σ, por lo que hay un número infinito de distribuciones normales.

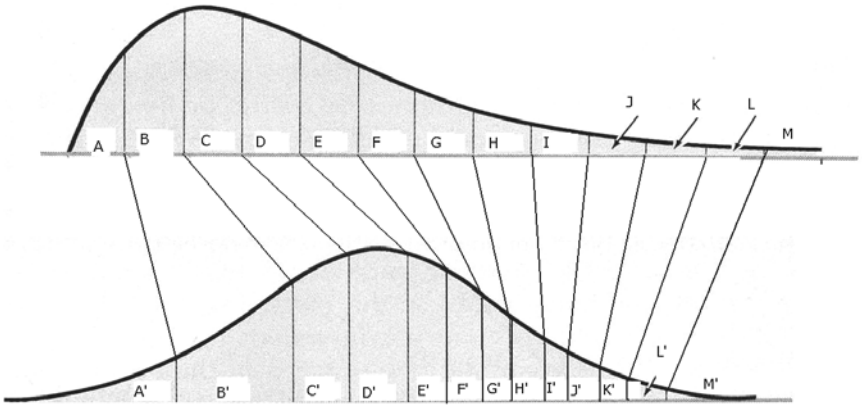
ÁREA BAJO LA CURVA EN UNA DISTRIBUCIÓN NORMAL

El área bajo la curva, entre dos puntos, indica la probabilidad de que la frecuencia se encuentre entre dichos valores. Así, por ejemplo, un puntaje en la PSU, que tiene una distribución con media de 500, significa que bajo 500 puntos se encuentra el 50% de la población. Esto se obtiene de ver la probabilidad entre 150 (menor valor en la prueba) y 500 puntos, que es justamente la mitad. Lo mismo ocurre hacia la derecha, dado que la curva normal es simétrica, por lo tanto el promedio es igual a la mediana y al modo.



NORMALIZACIÓN

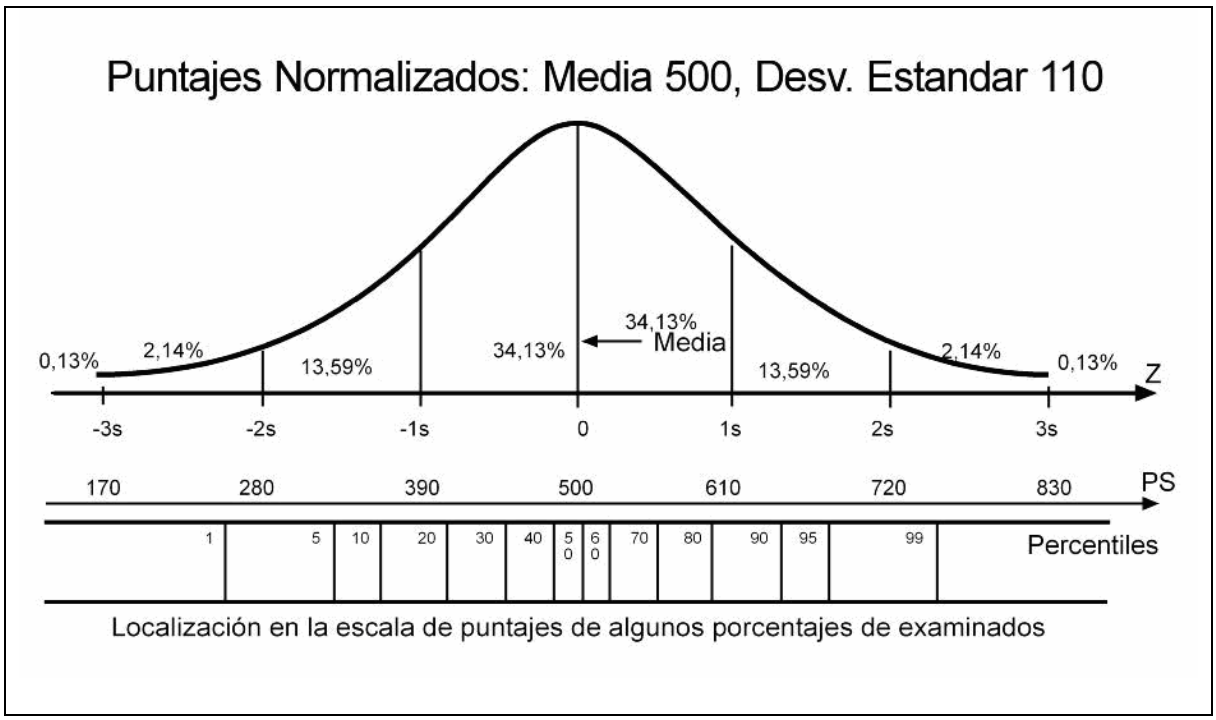
Se asocia con la acción de transformar una distribución cualquiera a una distribución normal. Corresponde ajustar los datos de la distribución “inicial” a una distribución normal. En este caso se cambia la forma de la distribución original manteniendo la proporción de casos entre valores contiguos.



EJEMPLO: NORMALIZACIÓN DE LAS PRUEBAS (PSU)

Los puntajes de las PSU se normalizan desde el Proceso de Admisión 2005, con una media de 500 puntos y desviación estándar de 110 puntos, truncando los extremos en 150 y 850 puntos. El 99% central de los puntajes se normalizan con un promedio de 500 y desviación estándar 110; el 0,5% de cada extremo se ajusta interpolando linealmente.

En el caso de la prueba de Ciencias, se normalizará luego de estimar la equivalencia de puntajes entre sus tres versiones, empleando el módulo común como base para establecer dicha equivalencia.



Ejemplo:

Si en la PSU de Lenguaje y Comunicación, rendida en la Admisión 2007 por 212.723 postulantes, el 15,87 % de éstos tienen 610 o más puntos, esto significa que 33.759 de ellos sacaron 610 o más puntos y el resto obtuvo puntajes menores.

Nota:

Por normalización se entiende una transformación de la distribución de los puntajes corregidos, manteniendo el orden. Para ello se calculan los percentiles asociados a cada puntaje corregido, y luego se identifica su equivalente en puntaje estandarizado en la distribución normal (puntaje Z). Este puntaje Z es finalmente convertido a la escala que se desee, en este caso, con promedio 500 y desviación estándar 110, obteniéndose el puntaje final PS, haciendo PS=110*Z+500.

A N E X O

MEDIA ARITMÉTICA PONDERADA: En ocasiones no todos los valores de la variable tienen el mismo peso. Esta importancia que asignamos a cada variable, es independiente de la frecuencia absoluta que tenga. Será como un aumento del valor de esa variable, en tantas veces como consideremos su peso.

Por lo tanto la media aritmética ponderada se utiliza cuando a cada valor de la variable (x_i) se le otorga una ponderación o peso distinto de la frecuencia o repetición. Para poder calcularla se tendrá que tener en cuenta las ponderaciones de cada uno de los valores que tenga la variable

Se la suele representar como:

$$\overline{X}w = \frac{\sum x_i w_i n_i}{\sum w_i n_i}$$

Siendo w_i la ponderación de la variable x_i y $\sum w_i$ la suma de todas las ponderaciones.

Un ejemplo es la obtención de la media ponderada de los puntajes según las distintas ponderaciones dadas por las universidades para alguna carrera específica:

Ponderación o “peso”	=	NEM	=	20%
		LyC	=	25%
		MAT	=	25%
		CS	=	30%
Puntajes	=	NEM	=	600
		LyC	=	680
		MAT	=	620
		CS	=	650

$$\bar{x} = \frac{20\% \cdot 600 + 25\% \cdot 680 + 25\% \cdot 620 + 30\% \cdot 650}{20\% + 25\% + 25\% + 30\%}$$

$$\bar{x} = \frac{12.000 + 17.000 + 15.500 + 19.500}{100\%}$$

$$\bar{x} = \frac{64.000}{100\%} = 640 ptos.$$

Esta misma fórmula se emplea para calcular el promedio de un grupo a partir del conocimiento del promedio y de la cantidad de casos que hay en cada subgrupo de él.

Un ejemplo de este caso es el cálculo del promedio de notas en Educación Media a partir de los promedios de 1º, 2º, 3º y 4º año medio.

Normalmente los postulantes suman los promedios de sus notas de enseñanza media y lo dividen por 4, ignorando la ponderación de cada promedio por cuanto el número de asignaturas de cada curso es distinto.

A continuación, se describen otros conceptos de “media” de escasa utilización en educación.

LA MEDIA GEOMÉTRICA: en una cantidad finita de números (digamos 'n' números) es la raíz n-ésima del producto de todos los números.

Se calcula con la siguiente fórmula

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Por ejemplo, la media geométrica de 2 y 18 es

$$\sqrt[2]{2 \cdot 18} = \sqrt[2]{36} = 6$$

Otro ejemplo, la media de 1, 3 y 9 sería

$$\sqrt[3]{1 \cdot 3 \cdot 9} = \sqrt[3]{27} = 3$$

Sólo es relevante la media geométrica si todos los números son positivos. Si uno de ellos es 0, entonces el resultado es 0. Si hay un número negativo (o una cantidad impar de ellos) entonces la media geométrica es, o bien negativa o bien inexistente en los números reales.

En muchas ocasiones se utiliza su trasformación en el manejo estadístico de variables con distribución no normal.

La media geométrica es relevante cuando varias cantidades son multiplicadas para producir un total.

MEDIA ARMÓNICA: Es la inversa de la media aritmética de los inversos de los valores de la variable, se representa por H, y responde a la siguiente expresión:

$$H = \frac{n}{\sum \frac{n_i}{xi}}$$

Esta media no es aconsejable en distribuciones de variables con valores pequeños. Se suele utilizar para promediar variables tales como productividades, velocidades, tiempos, rendimientos, cambios, etc.

Ventajas e inconvenientes:

- En su cálculo intervienen todos los valores de la distribución.
- Su cálculo no tiene sentido cuando algún valor de la variable toma valor cero.
- Es única.

Como ejemplo se muestra el caso de las edades de las tres personas 80, 55 y 30 años.

$$H = \frac{3}{1/80 + 1/55 + 1/30} = \frac{3}{\frac{1650 + 2400 + 4400}{132000}} = \frac{3}{8450/132000} = \frac{132000 \cdot 3}{8450}$$

$$= \frac{396000}{8450} = 46{,}86 \text{ años}$$