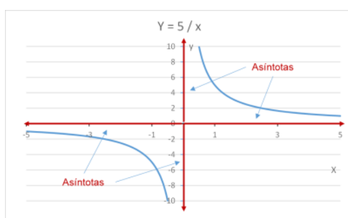
Unidad 4 Análisis y aplicación de las medidas de tendencia central, variabilidad, ubicación y forma.

Distribución normal ($\bar{N}$)

En las ciencias del comportamiento las distribuciones tienden a adoptar una forma característica conocida como *curva normal* o *distribución gaussiana*. La *curva normal* es una curva simétrica y asintótica, definida por la *media*, *desviación estándar o típica* y por el valor de la ordenada máxima. Los fenómenos conductuales tienden a distribuirse normalmente. Dentro de la curva normal, existen una serie de elementos importante para identificar y describir los fenómenos psicológicos, así como la forma en que estos se observan en el mundo real. Para esto existe la clasificación de las diversas medidas que arroja la estadística descriptiva: Medidas de tendencia central (modo, media y mediana), medidas de ubicación o posición (percentiles, decil y cuartiles), medidas de dispersión (rango, desviación media, rango o rango semicuartilar, desviación estándar y varianza.), medidas de forma (asimetría o sesgo y agrupamiento o curtosis).

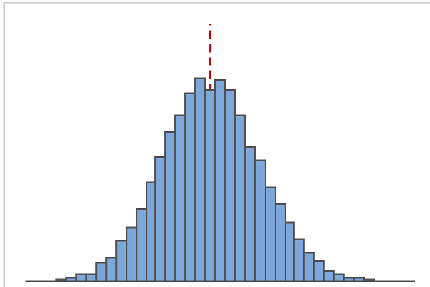
La distribución normal ($\bar{N}$), es la distribución más importante en probabilidad y estadística. Fue desarrollada por Gauss, por lo que se le llama *curva de Gauss*, *distribución gaussiana, etc.*, una de sus principales características es la asintotía. Se dice que la distribución muestral es asintótica porque el techo nunca toca el piso (abscisa). En una distribución de probabilidad, la ordenada es *p(x)*; si el techo pudiera tocar el piso, existiría *p(X)* = 0, y por axioma sabemos que esto es un *evento imposible*.

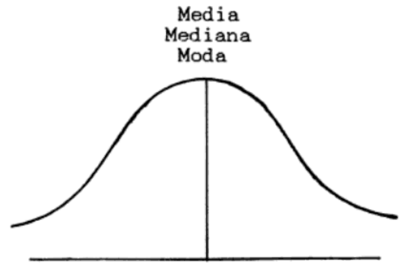


4.1. Medidas de Tendencia Central

El centro de los datos es el área donde se aglomera la mayoría de los valores de un conjunto de datos. La tendencia central se describe mediante varios estadísticos diferentes, como la media, la media recortada, la mediana o la moda. El conocer la tendencia central de los datos permite identificar los valores más representativos de los datos, de acuerdo a la manera como se concentran.

En la $\bar{N}$, se cumple la siguiente relación $M = Md = Mo$, y los tres parámetros se ubican exactamente en el centro. Puesto que el modo queda en el centro, las frecuencias tienen al disminuir hacia los extremos. Esto determinar la forma que campana.





4.1.1. Media:

Es la suma de todos los valores observados, dividido por el número total de observaciones.

Cuando los valores representan una población la ecuación se define como:

$$\bar{\mu} = \frac{X_1 + X_2 + X_2 + \cdots + X_n}{N} = \frac{\sum_{i-1}^{n} X_i}{N}$$

Donde:

$\bar{\mu}$ ≡ media,

N ≡ tamaño de la población,

X_i ≡ cada uno de los valores de la población.

La media aritmética para una muestra está determinada como

$$\bar{X} = \frac{X_1 + X_2 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i-1}^{n} X_i}{n}$$

Donde

$\bar{X}$ ≡ Media para la muestra,

n ≡ tamaño de la muestra,

X ≡ cada uno de los valores observados.

Esta fórmula únicamente es aplicable si los datos se encuentran desagrupados; en caso contrario debemos calcular la media mediante la multiplicación de los diferentes valores por la frecuencia con que se encuentren dentro de la información; es decir,

$$\bar{X} = \frac{\sum_{i-1}^{n} Y_i n_j}{n}$$

Donde:

$Y_i \equiv$ representa el punto medio de cada observación,

$n_j \equiv$ es la frecuencia o número de observaciones en cada clase

$n \equiv$ es el tamaño de la muestra siendo igual a la suma de las frecuencias de cada clase.

4.1.2. Mediana

Esta medida indica que la mitad de los datos se encuentran por debajo de este valor y la otra mitad por encima del mismo.

La posición de la mediana utiliza la fórmula:

$$Md = \frac{n+1}{2}$$

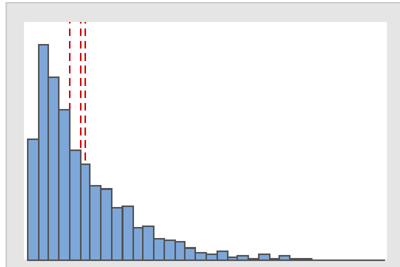
La mediana indica el valor que separa los datos en dos fracciones iguales con el 50% de los datos cada una. Las muestras que cuentan con un número impar de observaciones o datos, la mediana dará como resultado una de las posiciones de la serie ordenada; mientras que las muestras con un número par de observaciones se debe promediar los valores de las dos posiciones centrales.

4.1.3. Moda

Indica el valor que más veces se repite dentro de los datos. En algunas ocasiones se presente dos valores con la mayor frecuencia, lo cual se denomina *bimodal* o en otros casos más de dos valores, lo que se conoce como *multimodal*.

"Si los datos son simétricos, la media y mediana, serán aproximadamente iguales. Si los datos son asimétricos, las medidas pudieran desplazarse hacia las observaciones más extremas. Entre estas medidas, la media se ve más afectada por los valores extremos que la mediana".

Es necesario considerar que *a medida que las distribuciones se desvían de la normalidad, estos estadísticos comienzan a diferenciarse*. Por lo que, cuando las distribuciones se desvían de la normalidad y se vuelven más asimétricas, la desviación estándar se diferencia más de la distancia entre la media y un valor típico de los datos.



En estos casos el *rango intercuartil* es una mejor medida de dispersión que la desviación estándar cuando se trata de datos muy asimétricos, porque el rango intercuartil no se ve afectado por los rangos extremos.

4.2. Medidas de ubicación o posición

Las **medidas de posición** dividen un conjunto de datos en grupos con el mismo número de individuos.

Para calcular las **medidas de posición** es necesario que los **datos** estén ordenados de **menor a mayor**.

4.2.1 Límites exactos

En variables continuas cualquier puntaje empírico corresponde al punto medio de un rango. Es decir, un valor entero por ejemplo 110, señala que este valor se encuentra entre 109.50 y110.50. Por otra parte, en distribuciones intervalares se consideran los límites del intervalo, por ejemplo un punta que se ubique entre el intervalo 91-95 señalará que sus límites exactos serán 90.50 y 95.50.

4.2.2 Fractil

Es el dato o puntaje X que limita por la derecha una proporción cualquiera de una distribución. El cálculo de fractiles permite calcular cualquier medida de ubicación al determinar la proporción de la distribución que limita.

Símbolo	*Nombre*	*Proporción*
Md	Mediana	0.50
Q_1	Cuartil 1	0.25
$Q_2 = Md$	Cuartil 2	0.50
Q_3	Cuartil 3	0.75
D_1	Decil 1	0.10
D_7	Decil 2	0.70
P_{89}	Percentil 89	0.89

4.2.3 Rango percentil

Es la proporción de la distribución que queda limitada por un dato determinado y permite evaluar la posición del resultado de una prueba en particular entre los resultados generales de la misma.

Los rangos percentiles de una calificación se definen como el porcentaje de personas del grupo normativo que obtienen las calificaciones más bajas. Indica la clasificación relativa de la persona en porcentajes.

Ejemplo:

X	*f*	*fX*	*FA*	*FApm*	*PA*	*RP*	*x*	*z*	*Z*
20	1	20	0	0.5	0.003	< 1	-6.54	-2.37	26
21	3	63	1	2.5	0.014	1	-5.54	-2.01	30
22	12	264	4	10	0.056	6	-4.54	-1.64	34
23	14	322	16	23	0.130	13	-3.54	-1.28	37
24	14	336	30	37	0.209	21	-2.54	-0.92	41
25	22	550	44	55	0.311	31	-1.54	-0.56	44
26	15	390	66	73.5	0.415	42	-0.54	-0.20	48
27	28	756	81	95	0.537	54	0.46	0.17	52
28	18	504	109	118	0.667	67	1.46	0.53	55
29	22	638	127	138	0.780	78	2.46	0.89	59
30	17	510	149	157.5	0.890	89	3.46	1.25	63
31	7	217	166	169.5	0.958	96	4.46	1.61	66
32	4	128	173	175	0.989	99	5.46	1.98	70
Σ	177	4698							

4.2.4. Cuartiles

Divide a los datos ordenados en cuartos o cuatro partes de igual valor porcentual, se conforma con la letra *Q*: (Q_1, Q_2 y Q_3)

Q_1: representa un valor por debajo del cual quedan un cuarto (25%) de los valores ordenados.

Q_2: Es el segundo cuartil y se considera la mediana

Q_3: El tercer cuartil representa el valor por debajo del que quedan el 75% de los datos.

$$\boldsymbol{Q_1 = L_{ri}\frac{\frac{n}{4} - Fa}{f}(Ac)}$$

$$\boldsymbol{Q_3 = L_{ri}\frac{\frac{3n}{4} - Fa}{f}(Ac)}$$

Q_1 = Primer cuartil

Q3 = Tercer cuartil

Lri = Límite inferior de la clase que contiene Q_1
N= Tamaño de la muestra
Fa = número acumulado de observaciones precedente a la clase Q_1
F = Frecuencia de la clase
Ac = Amplitud de la clase (Ac = $Q_3 - Q_1$)

4.2.5. Deciles

Dividen los datos ordenados en diez partes equitativas porcentualmente hablando, y se denotan por D1, D2, D3,… D9.

$D_1 = C_{10}$	$D_4 = C_{40}$	$D_7 = C_{70}$
$D_2 = C_{20}$	$D_5 = C_{50}$	$D_8 = C_{80}$
$D_3 = C_{30}$	$D_6 = C_{60}$	$D_9 = C_{90}$

$$F_p = L_i + \left(\frac{(p * N) - F_a}{fp} * i\right)$$

$$p = 0.1000, 0.2000, 0.3000, \ldots, 0.9000$$

X	*F*	*FA*	*fX*	X^2	fX^2
6	8	8	48	36	288
7	19	27	133	49	931
8	23	50	184	64	1472
9	10	60	90	81	810
10	3	63	30	100	300
$\Sigma f =$	63	$\Sigma fX =$	485	$\Sigma fX^2 =$	3801

p ≡ Proporción de la distribución dada
F_p ≡ Fractil que limita a *p*
Li ≡ Límite inferior exacto del dato o intervalo que contiene *p*
Fa ≡ Frecuencia acumulada hasta antes del dato o intervalo que contiene a *p*
fp ≡ Frecuencia simple del dato o intervalo que contiene a *p*
i ≡ Tamaño del intervalo. Nótese que en una distribución puntual *i* = 1 y, por tanto, puede eliminarse en la formula

Para obtener los valores de cada literal en la fórmula, comenzamos para identificar el dato o intervalo que contiene a (*p* * *N*) buscando en la columna de frecuencias acumuladas en la distribución de frecuencias.

4.3. Medidas de dispersión (rango, desviación media, rango intercuartilar, desviación estándar y varianza.)

4.3.1. Rango

Intervalo entre el valor mínimo de un conjunto de datos. Es la diferencia entre el dato con el valor más grande (X_{max}) y el valor del dato más pequeño (X_{min}) obtenidos después de la aplicación de la escala. El rango representa el intervalo más pequeño que contiene todos los valores de los datos.

4.3.2. Rango potencial de la escala

Se da en relacionan los puntajes mínimos y máximos que probablemente se pueden obtener en una escala

4.3.3. Rango intercuartilar

Es una estimación estadística de la dispersión de una distribución de datos. Consiste en la diferencia entre el **tercer** y el **primer** cuartil. Mediante esta medida se eliminan los valores extremadamente alejados. El rango intercuartílico es recomendable cuando la medida de tendencia central utilizada es la mediana

$$IQR = Q_3 - Q_1$$

4.3.3. Suma (Σ)

Es el total de todos los valores de los datos.

4.3.4. Desviación media

Es la media aritmética de los valores absolutos de las desviaciones respecto a la media, se representa por D_x.

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N}$$

Ejemplo: Desviación media de: 6, 4, 7, 7, 5, 7, 6, 16

$$\bar{x} = \frac{6 + 4 + 7 + 7 + 5 + 7 + 6 + 16}{8} = 7.25$$

$$D_{\bar{x}} = \frac{\lceil 6 - 7.25 \rceil + |4 - 7.25| + |7 - 7.25| + |7 - 7.25| + |5 - 7.25| + |7 - 7.25| + |6 - 7.25| + \lceil 16 - 7.25 \rceil}{8} = 3.5$$

4.3.4. Desviación estándar

Es una medida de dispersión de los datos con respecto al valor del promedio, es decir, es la variación esperada con respecto a la media aritmética.

4.3.5. Varianza

Es la media de las desviaciones cuadráticas de una variable aleatoria, considero a la media de ésta. A mayor valor de la dispersión, mayor variabilidad, a menor valor más homogeneidad.

4.4. Medidas de forma (asimetría o sesgo y agrupamiento o curtosis)

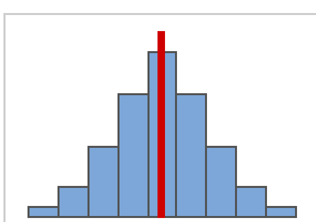
4.4.1. Asimetría y sesgo

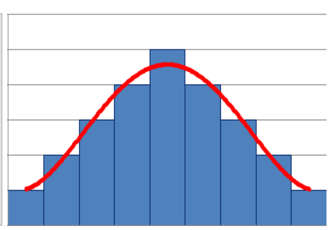
Es el grado en que los datos no son simétricos. El hecho de que el valor de la asimetría sea 0, positivo o negativo, revela información sobre la forma de los datos.

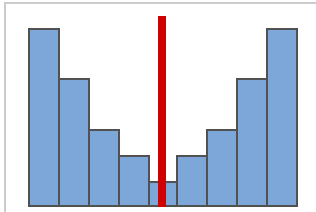
La $\overline{N}$ es simétrica, es decir, tomando el centro como referencia ambas mitades son idénticas pero inversas. Si *I(X)* es la imagen inversa de *X*, entonces A = I(B); B = I(A).

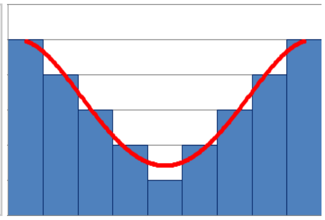
A medida que los datos son más simétricos, el valor de su asimetría se acerca a cero.

Pero, una distribución en la que ambos extremos cuentan con valores elevados siguen siendo un reflejo el uno del otro pero dista mucho de ser normal.





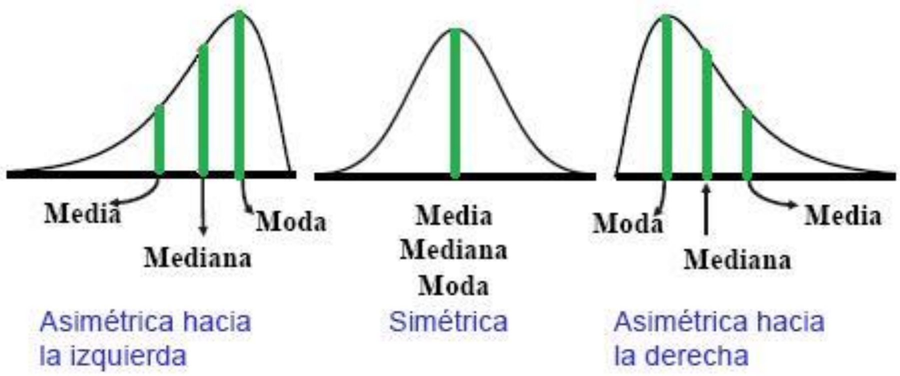




El sesgo, es el rasgo de *asimetría* de una distribución, y puede medirse por:

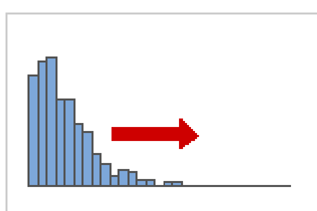
$S \equiv$ Sesgo $\qquad S = \frac{3(M-Md)}{\sigma} \qquad$ Para una distribución normal: $\overline{N}|S^0 = 0$

Si el sesgo tiene signo positivo, se dice que es un *sesgo derecho o positivo*; si tiene signo negativo, se dice que es *sesgo izquierdo o negativo*.



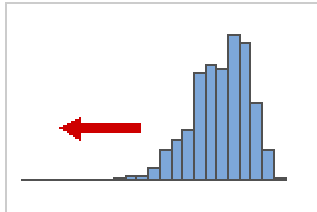
Distribuciones asimétricas hacia la derecha o positivas

En los datos con *asimetría positiva* o *asimétricos* hacia la derecha (*sesgo derecho o positivo)* la *cola* de la distribución apunta hacia la derecha y su valor de asimetría es positivo (mayor que 0). Ejemplo: Salarios de empleados.



Distribuciones asimétricas hacia la izquierda o negativas

En los datos asimétricos hacia la izquierda o con asimetría negativa (*sesgo izquierdo o negativo)* la *cola* de la distribución apunta hacia la izquierda y su valor de asimetría negativo (menor que 0). Ejemplo defectos en la fabricación masiva de productos)



4.4.2. Curtosis o agrupamiento (*K*)

Indica la manera en que el pico y las colas de una distribución difieren de la distribución normal. Ayuda a entender las características generales de la distribución de datos, señalando el grado de apuntamiento de una distribución:

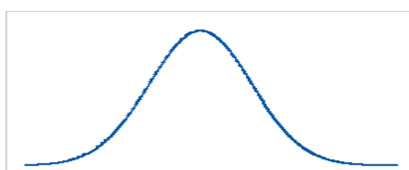
$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{Ns^4} - 3$$

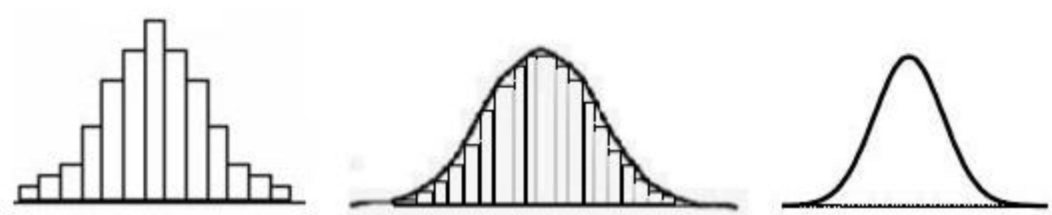
Siendo: $\bar{x}$ media, *S* desviación estándar,

En cuanto a su curtosis, existen tres tipos de distribuciones:

Distribucuón Mesocúrtica o Normal

Presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal), es decir: ***g2* = 0**

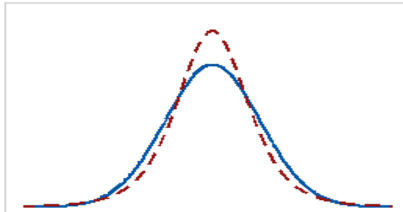


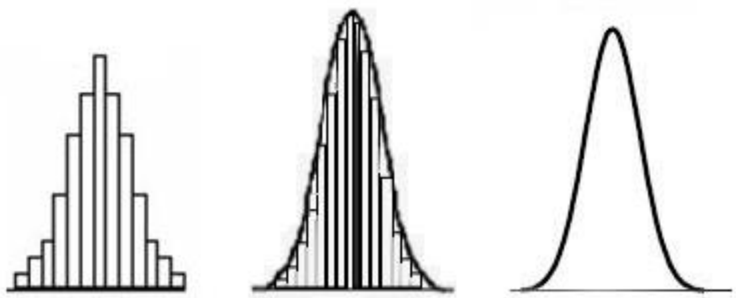


La línea de base de valor de curtosis *0*, señala que los datos que siguen una distribución normal y tienen un valor de curtosis de 0. Una curtosis de muestra que se desvía *significativamente* de 0 puede indicar que los datos no están distribuidos normalmente.

Distribución Leptocúrtica

Presenta un elevado grado de concentración alrededor de los valores centrales de la variable, es decir: ***g2* > 0**

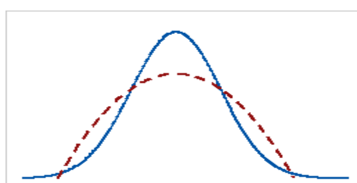


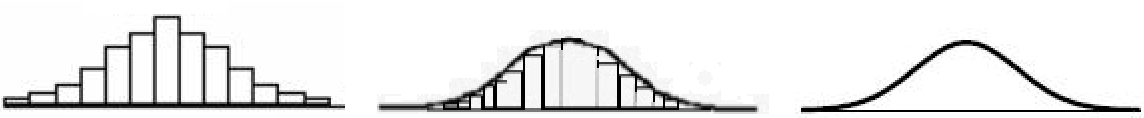


La distribución con valor positivo de curtosis indica que la primera, tiene colas más pesadas y un pico más pronunciado que la distribución normal. La línea continua indica la distribución normal y la línea de puntos indica una distribución con un valor positivo de curtosis.

Distribución Platicúrtica

Presenta un reducido grado de concentración alrededor de los valores centrales de la variable, es decir: ***g2* < 0**





La distribución con un valor negativo de curtosis indica que la primera tiene colas más ligeras y un pico más plano que la distribución normal. La línea continua indica la distribución normal y la línea de puntos indica una distribución con un valor negativo de curtosis.

$\alpha = 0.05$

$\overline{N}$|K = 0.263 curtosis

(0.263 – (0.263 x 0.05)) ≤ K ≤ (0.263 + (0.263 x 0.05))

0.250 ≤ K ≤ 0.276

5%

Ejemplo: Se tiene una distribución con los siguientes parámetros

$\bar{X}$ =	56.8	RSIC =	21.4
Md =	55.3	P_{90} =	92.3
S =	4.9	P_{10} =	11.6

Ensaye la normalidad de esta distribución y concluya:

Sesgo: $S = \frac{3(56.8-55.3)}{4.9} = 0.92$

Curtosis: $K = \frac{21.3}{92.3-11.6} = 0.265$

Conclusiones: La distribución tiene sesgo derecho con 0.92 unidades. La curtosis no es normal, se tiene una distribución leptocurtica con 0.002.

La curva normal comprende para su entendimiento de una serie de elementos.

4.5. Medidas de proporción

4.5.1. Redondeo

Permite ajustar la fracción decimal a un número de dígitos preestablecido. Para el redondeo aunque no existen criterios claramente establecidos se considera que si la cifra representa una proporción, la fracción decimal se redondeará a cuatro dígitos, así al transformarla en porcentaje la fracción decimal quedará con cuatro dígitos. De no ser así solo se redondeará a dos dígitos.

5.4.2. Proporción

Es el resultado de dividir la cardinalidad o frecuencia de un conjunto *a* entre la cardinalidad o frecuencia de un conjunto *b*. Al multiplicarla por 100, se obtiene un porcentaje.

$$Proporción = \left(\frac{a}{b}\right)100$$
$$Proporción = \left(\frac{25}{75}\right)100 = 33.33$$

4.5.3. Distribución de proporciones

Se da a partir de la clasificación o distribución de frecuencias. En la distribución de proporciones la suma de sus partes siempre es 1.0

4.5.4. Ajuste de sumas

Al calcular la distribución de proporciones el redondeo de cifras puede provocar imprecisiones por lo que es necesario ajustar los datos para que su resultado final sea igual a 1.0, considerando la siguiente regla: si la suma es mayor que 1.000, restar un diezmilésimo a la proporción más grande, otro a la siguiente y así sucesivamente hasta ajustar la suma a 1.000. Si la suma es menor que 1.0000, aumentar un diezmilésimo a la proporción más pequeña, otro a la siguiente y así sucesivamente hasta ajustar la suma a 1.000.

4.6. Calificación estándar

Una calificación estándar (z) es la desviación que tiene una calificación brutal de la media, en unidades de desviación estándar:

$$z = \frac{X - \bar{X}}{s}$$

Puesto que la unidad básica de la escala es la desviación estándar, se dice que las calificaciones son estándar. Las propiedades de las calificaciones estándar (z) son:

a) se expresan en una escala que tiene una media de cero y una desviación estándar de uno.
b) El valor absoluto de una calificación z indica la distancia a la que se encuentre la calificación bruta de la media de la distribución. El signo de las calificaciones z indica si la calificación cae por encima o por debajo de la media; las calificaciones por encima de la media tendrá signos positivos; los que se encuentren por debajo de la media, negativos.
c) Al expresarse las calificaciones estándar en una escala de intervalos, se puede someter a las operaciones aleje básicas comunes.
d) La transformación de las calificaciones brutas a estándar es lineal. Por esto, la forma de la distribución de las calificaciones z es similar a la distribución de las calificaciones brutas.
e) Si la distribución de las calificaciones brutas es normal, los rangos de las calificaciones z ira de aproximadamente -3 a +3.

Para evitar decimales y valores negativos, las calificaciones z se transforma por lo común a otra escala. Esta transformación que se la forma:

$$Z = A + Bz$$

En donde Z esta calificación estándar transformada y A y B la son constantes. Puesto que la suma o la multiplicación de o por una constante no destruye las relaciones en la escala, las relaciones entre las calificaciones z serán las mismas que entre las calificaciones brutas.

Aunque se puede utilizar constantes de cualquier tipo, el procedimiento recomendado (APA, 1974) es el de utilizar la transformación:

$$Z = 50 + 10z$$

Esta transformación cambia las calificaciones a una escala que tiene una media de 50 puntos y una desviación estándar de 10 puntos.