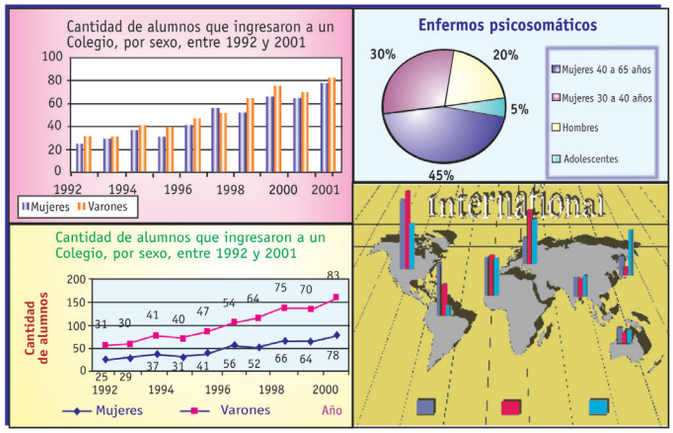
GUÍA 2

CONCEPTOS DE ESTADÍSTICA: LOS DATOS AGRUPADOS Y LAS MEDIDAS DE TENDENCIA CENTRAL



La representación de la gráfica de datos agrupados permite, de una ojeada, calcular aproximadamente ciertos elementos estadísticos, amén de ser llamativos, independientemente del tipo de gráfico que se utilice. En el mosaico, a la izquierda, se ven un histograma y un polígono de frecuencia para el mismo conjunto de datos; a la derecha, se ven un diagrama circular o un pastel y un mapograma, similar a un diagrama de barras, pero mejor decorado.

INDICADORES DE LOGRO

- Analiza la necesidad de organizar y presentar los datos en tablas de distribución de frecuencias
- Usa adecuadamente la técnica para elaborar tablas de distribución de frecuencias, porcentajes y porcentajes acumulados
- Elabora gráficas para representar los datos estadísticos, como histogramas y polígonos de frecuencia
- Maneja adecuadamente la aplicación EXCEL para elaborar diversos tipos de gráficas con datos estadísticos
- Define y calcula correctamente algunas medidas de tendencia central (media, mediana y moda) y de dispersión (varianza y desviación estándar)
- Usa con suficiencia la aplicación EXCEL para calcular los elementos mencionados antes
- Analiza instrumentos de evaluación, comparación y analiza datos para tomar decisiones. REFERENCIACIÓN COMPETITIVA
- Formula indicadores claros, que permitan medir el desempeño de sus acciones
- Reconoce las etapas del ciclo gerencial básico (PHVA)
- Reconoce procesos exitosos de otros
- Identifica las debilidades de sus procesos y los compara con los de otros
- Aprende y aplica en forma continua las mejores prácticas desarrolladas por otros
- Asume una posición positiva al cambio, que permite ajustar sus prácticas habituales

Las matemáticas son la base de la mayoría de las demás ciencias. La estadística es una de las varias formas en que se aplican los conceptos matemáticos.

Los profesionales de casi todas las ramas de la ciencia deben tener un conocimiento básico en estadística, que les sirva de base y apoyo en sus decisiones, en sus investigaciones y, en general, en todas las actividades propias de su profesión.

Las consideraciones anteriores, justifican plenamente la inclusión de estos apuntes de estadística básica que iniciaremos en esta guía.

Con el fin de iniciar la revisión de conceptos estadísticos, desarrollo la siguiente actividad:

Dados los valores 6, 3, 8, 15, 12, 4, 6, 12, 3, 6:

a) Los ordeno de menor a mayor y luego de mayor a menor.

b) Calculo su sumatoria.

c) Hallo su promedio.

d) Busco la sumatoria de sus cuadrados.

e) Determino las diferencias entre cada dato y su promedio.

f) Digo cuál es el dato que más se repite.

B

Con un compañero, leemos, interpretamos e interiorizamos los conceptos expuestos en la guía. Analizamos los ejercicios desarrollados y si es necesario los volvemos a resolver.

El vocablo estadística designa un grupo de métodos cuya finalidad es recoger, organizar y analizar datos numéricos significativos.

Supongamos que se recolectaron los siguientes datos, correspondientes a las edades de un grupo de 46 personas: 30, 31, 31, 32, 12, 36, 38, 39, 33, 28, 27, 38, 31, 27, 34, 35, 34, 38, 26, 33, 33, 27, 35, 36, 35, 32, 30, 34, 35, 10, 38, 27, 30, 37, 30, 35, 34, 26, 35, 38, 35, 32, 18, 34, 10, 39.

A simple vista, estos datos son poco significativos, pero podemos organizarlos en una distribución de frecuencias no agrupadas o agrupadas mediante el siguiente proceso:

1. Para las distribuciones de frecuencia no agrupada, se ordenan los datos de mayor a menor (o de menor a mayor) y se llenan, la primera columna con las edades X y la segunda con el número de veces en que se repite cada dato (frecuencia f), como se ve en la gráfica, aunque repartiendo los datos en 3 columnas para comprimir un poco la información:

X	f	X	f	X	f
10	2	32	3	36	2
10		32		36	
12	1	32		37	2
18	1	33	3	37	
26	2	33		38	5
26		33		38	
27	3	34	5	38	
27		34		38	
27		34		38	
28	1	34		39	2
30	4	34		39	
30		35	7		
30		35			
30		35		n =	46
31	3	35			
31		35			
31		35			
		35			

2. Para las distribuciones de frecuencia agrupada se ordenan los datos de mayor a menor y en la columna de datos X se escriben, entre los extremos, los otros elementos; se calcula el rango R, que es la diferencia entre el valor más grande $X_{Máx}$ y el más pequeño $X_{Mín}$, o sea: $R = X_{Máx} - X_{Mín}$. En este caso: $R = 39 - 10 = 29$

3. Se decide el número C de clases o intervalos, que por lo general no debe ser menor que 5 ni mayor que 15, dependiendo del número de datos. En este caso, tomemos 10.

4. Se determina el ancho A del intervalo, que es igual al rango R dividido entre el número C de clases y redondeando al entero más próximo, o sea: $A = \frac{R}{C} = \frac{29}{10} = 2.9 \approx 3$

5. Se especifican los límites de cada intervalo: los límites inferiores son múltiplos del ancho A (3, 6, 9, 12, 15,....39, 41, 44). El primer límite inferior debe ser menor o igual que el dato X mínimo, que es 10; y los límites superiores son iguales a los inferiores, más el ancho del intervalo, menos 1(en este caso, 3-1=2), o sea, 11, 14, 17,....41.

Lo expuesto de 2 a 5, se visualiza en la siguiente tabla:

X	Frecuencia		Intervalo C	Conteo	Frecuencia	%	% Acumulado
39	2		39 - 41	⁘	2	4.35	100.00
38	5		36 - 38	⁘	9	19.57	95.65
37	2		33 - 35	⁘ ⁘	15	32.60	76.08
36	2		30 - 32	⁘	10	21.74	43.48
35	7		27 - 29	⁘	4	8.70	21.74
34	5		24 - 26	⁘	2	4.35	13.04
33	3		21 - 23		0	0.00	8.69
32	3		18 - 20	⁘	1	2.17	8.69
31	3		15 - 17		0	0.00	6.52
30	4		12 - 14	⁘	1	2.17	6.52
28	1		9 - 11	⁘	2	4.35	4.35
27	3						
26	2				n=46	100.00	
18	1						
12	1						
10	2						
	n=46						

Cuando el conjunto de datos es grande, se usa el método de Tukey que cuenta los valores en prácticos grupos de a diez: las primeras 4 cuentas se denotan por puntos

que forman los vértices de un cuadrado; las siguientes cuatro cuentas son los lados del cuadrado; la cuentas novena y décima son las diagonales del cuadrado.

De los elementos de la columna "Conteo" se deduce la frecuencia f con sólo contar elementos.

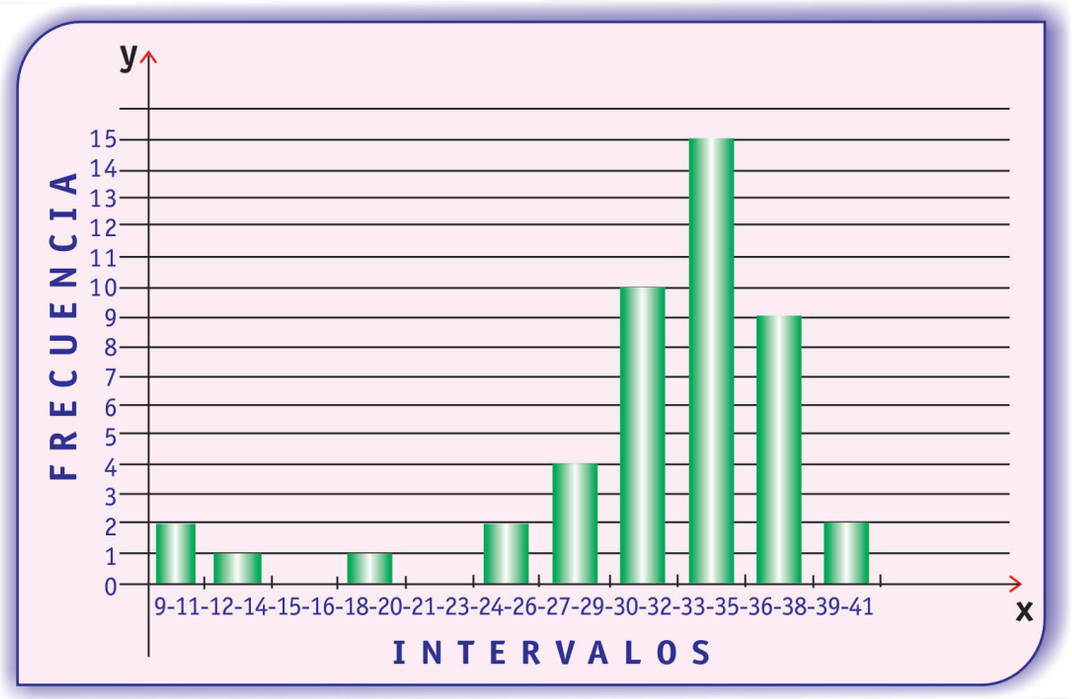
Para obtener la columna % es suficiente calcular, mediante una regla de tres, qué porcentaje es cada frecuencia con respecto al total de 46 datos (para este caso). Para obtener el porcentaje acumulado, basta tener en cuenta que el porcentaje del último intervalo (9 - 11) es el mismo porcentaje acumulado. Ahora, de abajo para arriba, al porcentaje acumulado se le suma el siguiente porcentaje (Por ejemplo: al último porcentaje acumulado 4.35 se le suma el porcentaje 2.17 para obtener un porcentaje acumulado de 6.52; ahora, a 6.52 de porcentaje acumulado se le suma el porcentaje 0.00 para obtener un porcentaje acumulado de 6.52, y así sucesivamente hasta llegar al 100.00).

GRÁFICAS DE DISTRIBUCIONES DE FRECUENCIAS

Las distribuciones de frecuencias agrupadas pueden analizarse más fácilmente si se representan gráficamente (una imagen vale más que mil palabras), pues se puede obtener información importante mirando simplemente su representación geométrica. Las gráficas más comunes son los histogramas o diagramas de barras y polígonos de frecuencia o diagramas de líneas.

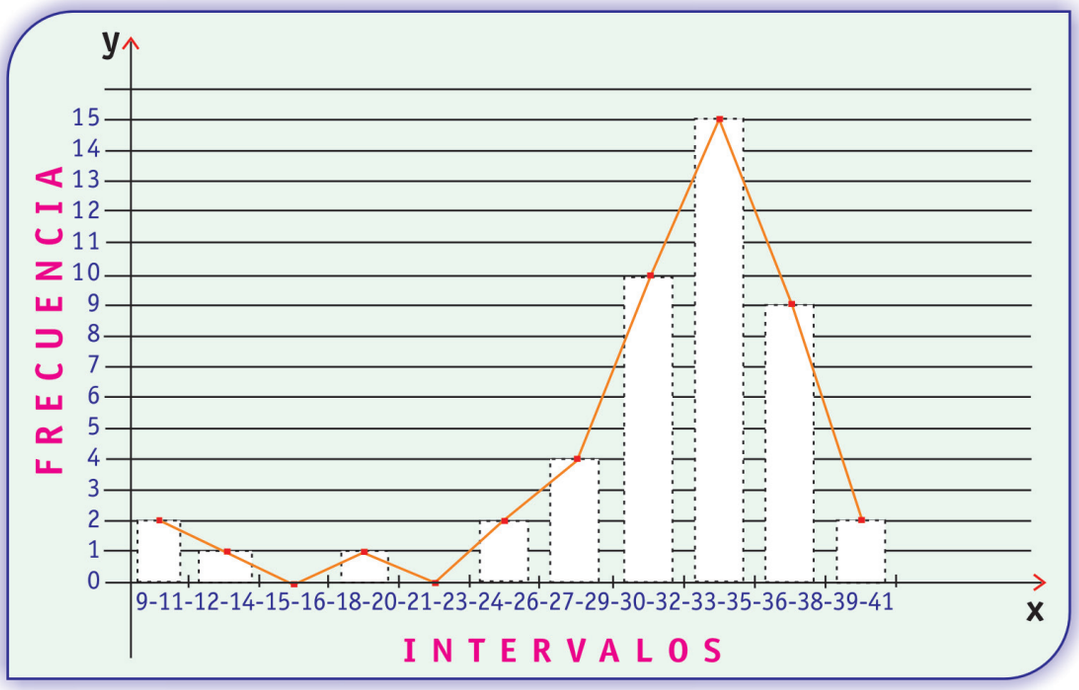
El histograma

Un histograma es una gráfica de una distribución de frecuencias en la que se utilizan barras cuyas alturas son proporcionales a la frecuencia de observaciones para cada clase o intervalo. En un sistema coordenado, en el eje de las abscisas se toman los intervalos y en el eje de las ordenadas la frecuencia, que se muestran en la gráfica anterior, así:



El polígono de frecuencia

El polígono de frecuencia se puede construir dibujando primero un histograma y conectando luego, por medio de segmentos de recta, los puntos medios de la parte superior de cada una de las barras o rectángulos, como se ve en la siguiente figura:



En un gráfico, con una ojeada puede determinarse, por ejemplo, cuál es la mayor frecuencia.

Cabe resaltar que la representación gráfica de los datos llama la atención, evitando la monotonía de la descripción numérica; ayuda a esclarecer el significado de los datos y facilita la retención de los mismos. El inconveniente es que la organización de los datos consume mucho tiempo, pero en la actualidad los computadores facilitan esta labor de una manera rápida y sencilla, usando una aplicación tan común como EXCEL. Se empieza por digitar los datos en columna (aunque no necesariamente) y luego se pueden realizar las siguientes acciones, por ejemplo:

a. Ordenar los datos: para ello seleccione el rango, pulse en Datos (en la barra de menús), Ordenar (en forma ascendente o descendente).

b. Contar los datos: seleccione el rango. Pulse Insertar, Función (en seleccionar una categoría, escoja Matemáticas y trigonométricas), Contar. El resultado es el valor de n.

c. El valor más grande de entre los datos: seleccione el rango. Pulse Insertar, Función (en categoría, opte por Estadísticas), Máx. El resultado es $X_{Máx}$.

d. El valor más pequeño de entre los datos: seleccione el rango. Pulse Insertar, Función (en categoría, opte por Estadísticas), Mín. El resultado es $X_{Mín}$.

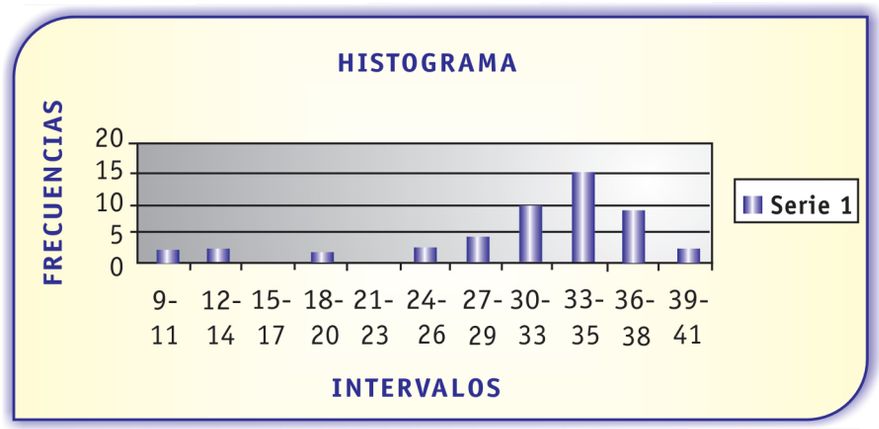
e. Cálculo del rango R. Ubíquese en la celda que desee y como ya conoce El resultado $X_{Máx}$ y $X_{Mín}$, digite: $=X_{Máx} - X_{Mín}$. La máquina mostrará el valor R del rango.

f. Ancho del intervalo: para los datos del ejemplo, ubíquese en una celda adecuada y digite = (valor de R) / 10. Se visualiza el ancho, que debe ser aproximado o por exceso o por defecto a un entero.

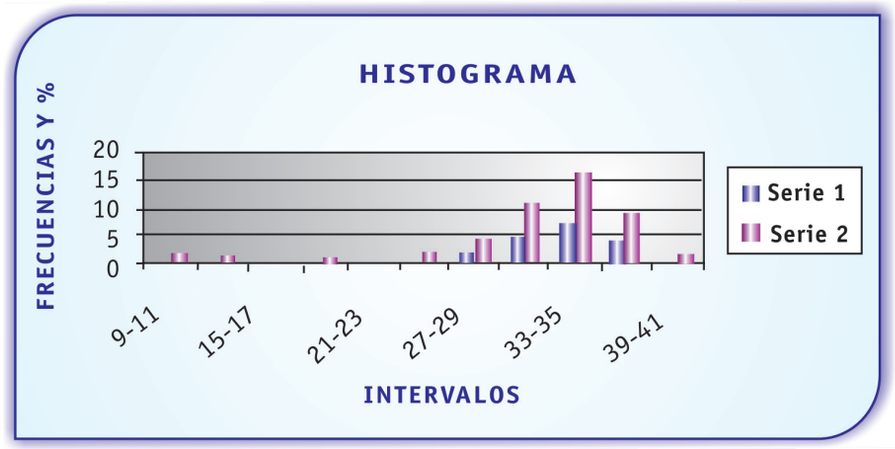
g. Para llenar la columna de las clases o intervalos, recuerde que los límites inferiores son múltiplos del ancho del intervalo (en este caso 9, 12, 15,...,39); los límites superiores se hallan sumando 2 al límite inferior respectivo (11, 14, 17,... 41).

h. En la columna que sigue a la de los intervalos, digite las respectivas frecuencias.

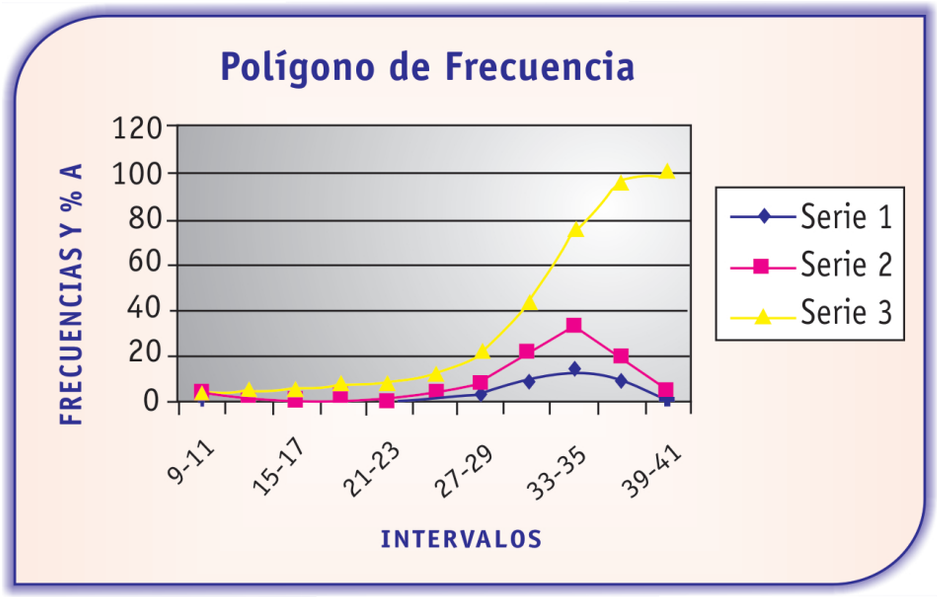
i. Para construir el histograma seleccione los datos de los intervalos y los de las frecuencias. Pulse Insertar, Gráfico, Columnas (pulse en la primera opción de las 7 que se muestran), Siguiente (Ya se visualiza el histograma), Siguiente, (en título del gráfico) escriba HISTOGRAMA, (en eje de categorías (X)) escriba INTERVALOS, (en eje de valores (Y) escriba FRECUENCIAS, Leyenda (Escoja la ubicación del cuadro Serie 1, Siguiente (si lo desea, active mostrar tabla de datos), Siguiente, Finalizar. El gráfico lo puede arrastrar a la posición que desee; mediante los elementos haladores (los 8 cuadraditos que resaltan) lo puede ampliar o reducir. He aquí el resultado:



En una misma gráfica se pueden ilustrar varias situaciones, pues Excel siempre toma la primera columna como el eje X y las demás columnas como valores de Y. Por ejemplo, se pueden graficar tanto la frecuencia como el porcentaje en el problema que nos ocupa. Para obtener la columna de porcentaje (%) ubíquese en la columna contigua a la de las frecuencias y digite los porcentajes correspondientes (aunque podría pedirle a Excel que lo haga por usted). Seleccione tanto la columna de los intervalos como la de la frecuencia y la de los porcentajes. Reitere los pasos para construir la gráfica anterior. El resultado es como éste:



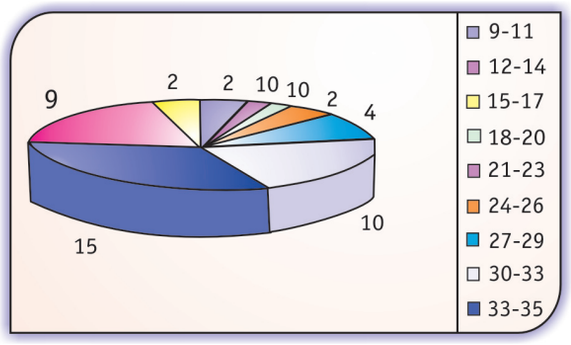
Ahora, elaboremos un gráfico de líneas usando las tres columnas anteriores y otra más que contenga los porcentajes acumulados % A. Digite entonces los porcentajes acumulados, seleccionemos todos los datos, introduzcámonos en gráficos, pero seleccionando Líneas y escogemos la cuarta opción de las 7 que aparecen. Siguiendo las instrucciones, se obtendrá algo como esto:



La serie 1 es la gráfica de las frecuencias, la serie 2 es la de los porcentajes y la serie 3 es la de porcentajes acumulados, conocida como OJIVA DE GALTON.

Otro tipo de gráfica bastante usada, y que se hace fácilmente en Excel, es la de tipo pastel.

Usemos la columna de intervalos y la de frecuencias del ejercicio sobre las edades. Seleccionamos todos los datos, entremos a gráficos. Se pulsa sobre Circular y se escoge la opción 2 de las 6 que se muestran. Se siguen las instrucciones del programa para obtener una gráfica como ésta:



Además de los tipos de gráficos usados antes, Excel presenta otra gama, al gusto y acorde con las necesidades. Es importante, pues, ensayar.

LAS MEDIDAS DE TENDENCIA CENTRAL

Como su nombre lo indica, son ciertos valores de datos que son centrales o típicos en un conjunto de datos agrupados.

Entre las medidas de tendencia central se tienen:

1. La media aritmética o promedio

Si dado un conjunto de datos se suman cada uno de los valores que toma la variable X, y el total se divide por el número de datos, se obtiene un indicador estadístico llamado media aritmética o promedio, que se simboliza y define así:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Por ejemplo, en las edades en años de las 46 personas indicadas al principio: $\overline{X} = \frac{30 + 31 + 31 + ...34 + 10 + 39}{46} = \frac{1448}{46} \approx 31.47$. Luego, la edad promedio es de 31.47 años aproximadamente.

2. La moda

Es el dato de más alta frecuencia, es decir, el que se repite un mayor número de veces (el que está de moda). Por ejemplo, en el conjunto propuesto de las 46 edades, la moda es 35 porque es el que más se repite (7 veces).

Sugerencia: así como Excel facilita la construcción de gráficas estadísticas, de modo ágil, también permite calcular automáticamente las medidas de tendencia central. Simplemente se digitan los datos (aún sin ordenar), se observan las direcciones del rango y se ubica el cursor en la posición en donde se desea que aparezca el resultado. Se entra a Insertar, Función, Estadísticas, Mediana (se le digita el rango - digamos A1..A46 - para nuestro caso de las edades. En la celda seleccionada aparece el valor central; para la moda ó para el promedio se procede igual, buscando estos nombres en Funciones Estadísticas.

3. La mediana

La mediana Md es otra medida de tendencia central y corresponde al valor medio en un conjunto de valores ordenados, es decir, es un valor tal que antes y después de él hay igual número de valores. Si el número de observaciones (ya ordenados) es impar, la mediana corresponde al valor que está en el centro de la distribución; si el número de observaciones es par, la mediana es el promedio de los dos términos centrales de la distribución.

Por ejemplo, en el conjunto ordenado 23, 22, 21, **20**, 19, 18, 17 la mediana es 20, porque antes y después de él hay 3 observaciones.

En el conjunto de datos de las edades de 46 personas (propuesto al principio), ordenado queda: 39, 39, 38, 38, 38, 38, 38, 37, 37, 36, 36, 35, 35, 35, 35, 35, 35, 34, 34, 34, 34, **34**, **33**, 33, 33, 32, 32, 32, 31, 31, 31, 30, 30, 30, 30, 28, 27, 27, 27, 26, 26, 18, 12, 10, 10. Como los datos ordenados tienen un número par de elementos, la mediana es el promedio de los dos términos centrales, 34 y 33, es decir que $\text{Md} = \frac{34 + 33}{2} = 33.5$ y significa que antes de los valores 34 y 33 hay 22 observaciones antes y después de ellos.

4. Cuartiles, Deciles y Percentiles

Otras medidas que están muy relacionadas con la Mediana, ya que se basan también en su posición en una serie de observaciones, son los Cuartiles, Deciles y Percentiles.

Si el conjunto de observaciones se divide en cuatro partes iguales, los valores que le corresponden a cada cuarto de la serie se denominan CUARTIL. El primer Cuartil, Q_1, es el valor por debajo del cual quedan el 25% de todos los valores; el tercer Cuartil, Q_3, es el valor bajo el cual quedan el 75% de los valores; el segundo Cuartil, Q_2, es exactamente la Mediana, como se puede apreciar en la Ojiva de Galton sacada de los intervalos y el porcentaje acumulado, como se dijo antes:

Para estimar la mediana en la gráfica, se localiza "50" (Q_2) en el eje vertical; se avanza horizontalmente hasta la intersección con la curva y luego se avanza verticalmente hacia abajo hasta intersectar el eje horizontal, en donde se puede leer la mediana (33.5).

Similarmente, los Deciles dividen la distribución en 10 partes iguales y los Percentiles en 100 partes iguales.

MEDIDAS DE VARIABILIDAD

Además de las medidas típicas o de tendencia central, existen medidas de variabilidad entre los valores, es decir, qué tan grandes son las diferencias entre los valores. Estas medidas cuantifican el grado de dispersión o la extensión de las diferencias individuales evidenciadas en la distribución.

Entre las medidas de variabilidad veremos la varianza y la desviación estándar.

La varianza se utiliza en estudios poblacionales para medir la homogeneidad de una muestra X, es decir, permite saber si la muestra es representativa de toda la población, aunque su resultado se encuentra elevado al cuadrado y en consecuencia se recomienda usar la desviación estándar que es una medida de variabilidad más exacta.

Cuando todas las N observaciones de la población están incluidas en el conjunto de datos, la varianza σ^2 (sigma minúscula al cuadrado), se encuentra dividiendo la suma

de los cuadrados de las diferencias entre cada dato y el promedio $\overline{X}$, o sea: $\sigma^2 = \frac{\sum (X - \overline{X})^2}{N}$. Para el ejemplo del conjunto de los datos de las 46 edades propuesto al principio, cuyo promedio es 31,47, se tiene:

Varianza = $\sigma^2 = \frac{\sum (X - \overline{X})^2}{N} = \frac{2179,48}{46} \approx 47,38$ años^2 la varianza, como suma de cuadrados, no tiene mucho significado descriptivo directo, amén de estar expresada en unidades cuadradas en lugar de las unidades originales.

La varianza se usa para calcular la desviación estándar σ, que se define como la raíz cuadrada de la varianza, o sea, $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SC}{N}}$; en el ejemplo: $\sigma = \sqrt{\frac{2179,48}{46}} = 6,88$ años.

La desviación estándar es más útil que la varianza para describir la variabilidad de un conjunto de datos y también tiene las mismas unidades originales. En una distribución normal, se espera que aproximadamente dos tercios de los valores estén dentro de + 1 ó -1 desviaciónes estándar con respecto al promedio.

Para determinar la comprensión de los elementos estadísticos vistos en B, en el subgrupo resolvemos los siguientes planteamientos. Comparamos las respuestas con las de otros subgrupos, nos ponemos de acuerdo y comentamos nuestras conclusiones con el profesor con la finalidad de aprovechar los aportes de todos.

Los siguientes datos corresponden al cociente intelectual de 150 niños de un colegio:
98, 119, 93, 99, 106, 102, 108, 109, 114, 108, 91, 91, 89, 120, 106, 127, 98, 104, 106, 114, 104, 106, 124, 101, 97, 121, 108, 113, 105, 125, 113, 120, 96, 108, 104, 116, 114, 118, 115, 121, 125, 129, 105, 118, 105, 100, 102, 110, 98, 122, 101, 120, 95, 118, 122, 95, 96, 129, 112, 117, 114, 109, 91, 113, 112, 89, 99, 124, 103, 105, 104, 106, 114, 124, 103, 108, 105, 92, 101, 112, 93, 109, 108, 115, 114, 93, 125, 88, 101, 88, 91, 121, 113, 121, 115, 107, 126, 113, 89, 104, 96, 129, 107, 120, 115, 118, 100, 100, 109, 97, 91, 122, 97, 118, 100, 106, 115, 110, 99, 85, 100, 112, 128, 111, 105, 98, 113, 101, 108, 116, 94, 92, 125, 121, 108, 119, 116, 103, 111, 113, 85, 109, 128, 88, 119, 118, 116, 113, 122, 126.

a) Ordenamos los datos de mayor a menor.

b) Hallamos la frecuencia de cada dato.

c) Calculamos el rango para la distribución agrupada.

d) Tomamos 9 clases o intervalos y buscamos la amplitud del intervalo, redondeando al entero más próximo.

e) Establecemos los límites inferior y superior de los intervalos.

f) Buscamos la frecuencia de las diversas clases.

g) Hallamos la frecuencia acumulada.

h) Encontramos el porcentaje de las diversas clases.

i) Determinamos el porcentaje acumulado.

j) Elaboramos las gráficas (en Excel si es posible) de barras y de líneas, tomando en el eje horizontal los intervalos o clases y en el vertical la frecuencia y la frecuencia acumulada o el porcentaje y el porcentaje acumulado.

k) Volvemos a los datos y calculamos el promedio, la mediana y la moda.

l) Si disponemos de los medios, damos las instrucciones a EXCEL para desarrollar los literales anteriores y poder confrontar con respuestas.

Como aplicación a los conceptos vistos, con el subgrupo desarrollamos las siguientes actividades:

1) En su institución, recopilen datos sobre el consumo de alcohol, tabaco y sustancias alucinógenas considerando, por ejemplo, edad y sexo. Tabulen, grafiquen y calculen elementos estadísticos de los indicados en la guía.
2) En su entorno, recopilen datos sobre analfabetismo con variables como edad, sexo, escolaridad. Igualmente, tabulen, grafiquen y calculen elementos estadísticos de los indicados en la guía.
3) Recorten datos estadísticos de periódicos o revistas recientes, interprétenlos y presenten sus resultados al grupo.

ESTUDIO Y ADAPTACIÓN DE LA GUÍA