Tema 10. ESTADÍSTICA BÁSICA Resumen

Caracteres y escalas de medición

Al hacer un trabajo estadístico hay que decidir los caracteres (las propiedades) que desean estudiarse. Un carácter puede ser cuantitativo o cualitativo.

Los valores que toma un carácter pueden medirse en distintas escalas: nominal, ordinal, de intervalo, o de proporción.

- La escala nominal consiste en situar a cada individuo o elemento en una u otra clase dada (por ejemplo, hombre/mujer; lugar de nacimiento). Pertenecer a una u otra clase no significa ser mejor o peor, indica que son distintos.
- La escala ordinal sitúa los posibles valores en orden (primero, segundo, ...), sin que la *distancia* entre dos posiciones consecutivas sea necesariamente constante, fija. En esta escala puede distinguirse, además, entre mayor y menor. Por ejemplo, la posición de los equipos de fútbol en el Campeonato de Liga; o las categorías profesionales en una empresa.

Las escalas nominal y ordinal son apropiadas para caracteres cualitativos.

- La escala de intervalo permite asignar a cada individuo un número para así indicar su posición exacta a lo largo de una escala continua. Por ejemplo, la temperatura medida en grados Celsius, donde 10 ºC significa más calor que 5 ºC, pero no el doble de calor.
- La escala de proporción (o proporcional) es la más perfecta. En ella existe un cero absoluto y, además, tiene sentido hablar de doble o mitad (un ejemplo de esta medida sería la longitud).

Las escalas de intervalo y proporcional se usan para medir caracteres cuantitativos.

Tablas de frecuencia

Se utilizan para facilitar la lectura e interpretación de grandes conjuntos de datos. Los datos suele agruparse, indicando su frecuencia absoluta o relativa; simples o acumuladas. La agrupación puede hacerse también en intervalos de clase. El punto medio de cada uno de esos intervalos sería el valor que representa a todos; s se llama marca de clase.

Ejemplo:

Intervalo	M.c.	f_i	%	F_i	%a
[1, 3]	2	0	0	0	0
[4, 6]	5	8	6,2	8	6,2
[7, 9]	8	31	23,8	39	30
[10, 12]	11	51	39,2	90	69,2
[13, 15]	14	33	25,4	123	94,6
[16, 18]	17	7	5,4	130	100
Totales		130	100	130	100

Diagrama de tallo y hojas

Es otro método de organización y visualización de un conjunto numérico de datos. Cada número del conjunto se representa por una *hoja* y un *tallo*. La hoja es la cifra de las unidades; el tallo es la cifra de las decenas. Una raya vertical separa el tallo de las hojas.

Ejemplo: El diagrama asociado a los datos

45 47 24 42 43 51 58 60 45 47 47 67 82 51 55

es el adjunto.

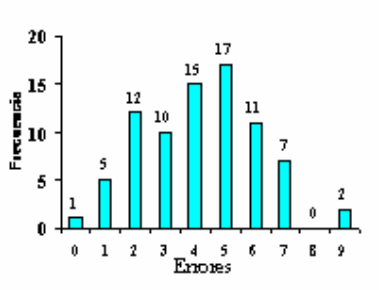
2 | 4
4 | 2 3 5 5 7 7 7
5 | 1 1 5 8
6 | 0 7
8 | 2

2 | 4 representa 24 corbatas

Gráficos estadísticos

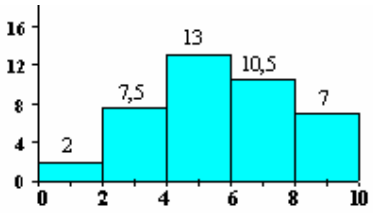
Diagramas de barras

Son gráficos que representan cada valor de la variable mediante una barra proporcional a la frecuencia con que se presenta. Las barras deben estar separadas, como en la figura del ejemplo.
Los diagramas de barras son apropiados para datos medidos en escala nominal u ordinal.



Histogramas

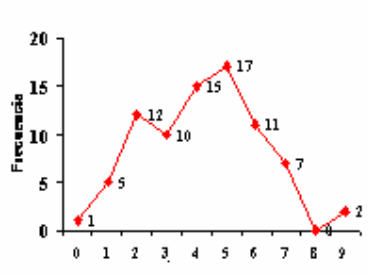
Se usan para variables agrupadas en intervalos, asignando a cada intervalo un rectángulo de superficie proporcional a su frecuencia. Por tanto, la altura de cada rectángulo se halla dividendo la frecuencia que representa entre la longitud del intervalo. (En la figura adjunta, el segundo rectángulo representa frecuenta 15; el 3º, 26; …)



Un histograma se diferencia de un diagrama de barras en que en el primero las barras no están separadas, pues la variable es continua.
Los histogramas son apropiados para variables continuas (medidas en escala de intervalo o de proporción).

Poligonal de frecuencias

Los histogramas, y algunos diagramas de barras, también se pueden representar por una poligonal de frecuencias, que es la línea que une los puntos correspondientes a las frecuencias de cada valor (extremos superiores de las barras). Pueden ser simples o de frecuencias acumuladas.



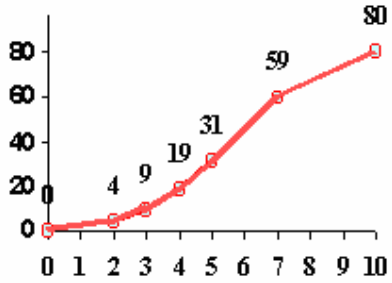
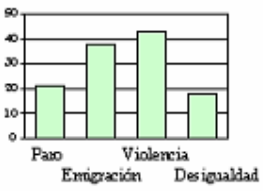
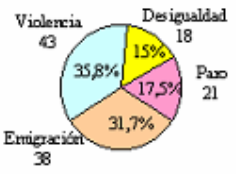


Diagrama de sectores En estos gráficos, cada suceso viene representado por un sector circular de amplitud proporcional a su frecuencia. La amplitud de cada sector se halla mediante una regla de tres.
Ejemplo: Los datos de frecuencia de la tabla de la izquierda, pueden representarse mediante un diagrama de barras o mediante un gráfico de sectores.

Paro	21
Emigración	38
Violencia	43
Desigualdad	18





Medidas de centralización. Están relacionadas con el promedio de los datos estudiados, y dan una idea de los valores más representativos para todo el conjunto.

La media aritmética
Se calcula sumando el valor de todos los datos y dividiendo por el número de ellos. Esto es:

$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$; o más breve: $\bar{x} = \frac{\Sigma x_i}{n}$

- Para datos agrupados: $\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}$, donde f_i es el número de veces que se repite el valor x_i.

Media ponderada: $\bar{x}_p = \frac{\Sigma x_i p_i}{\Sigma p_i}$, siendo p_i el peso del dato x_i.

La mediana: es el valor del dato que ocupa el lugar intermedio. Los datos deben estar ordenados.

La moda. Es el valor que se presenta con mayor frecuencia.

Medidas de posición. Indican la situación, en términos porcentuales, de algunos elementos de la distribución. Los datos estén ordenados de menor a mayor.

Amplitud, rango o recorrido. Es la diferencia entre los valores de los datos máximo y mínimo. La información que proporciona es imprecisa, pues sólo tiene en cuenta los valores extremos.

Cuartiles, deciles y percentiles
- Cuartiles: Valores de las posiciones correspondientes al 25 %, al 50 % y al 75 % de los datos.
- Deciles: Valores correspondientes al 10 %, 20 %, ... y 90 % de los datos
- Percentiles (o centiles) dan el valor de la posición correspondiente a cualquier porcentaje.

En todos los casos, su cálculo requiere aplicar la interpolación.

Medidas de dispersión. Dan una idea del *alejamiento* de los datos respecto de las medidas de centralización.

Varianza y desviación típica
- La varianza:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} \qquad s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 \qquad s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} \qquad s^2 = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2$$

- La desviación típica es la raíz cuadrada de la varianza. En consecuencia:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}; \qquad s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} \qquad s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$$

Ejemplo: La desviación típica de los conjuntos de datos: (a) 1, 4, 5, 6, 14 y (b) 1, 2, 3, 10, 14, es, respectivamente $s = \sqrt{18{,}8} = 4{,}34$ y $s = \sqrt{26} = 5{,}1$.

El coeficiente de variación. Es una medida de la *dispersión relativa* de dos conjuntos de datos.

Se define como: $CV = \frac{s}{\bar{x}}$

- El coeficiente de variación suele darse en porcentajes: $CV = \frac{s}{\bar{x}} \cdot 100$. Un CV mayor del 30 % indica que la media es poco representativa como medida del promedio.