Fundamentos matemáticos para la Ingeniería

Grado en Arquitectura Técnica

Tema 8
Estadística

José Barrios García

Departamento de Análisis Matemático

Universidad de La Laguna

jbarrios@ull.es

Índice

Tema 8. Estadística descriptiva

Introducción

Un gran porcentaje de la información que recibimos diariamente consiste en datos numéricos recogidos, organizados y presentados en tablas numéricas o gráficos, mediante *técnicas estadísticas*.

Es difícil dar una definición que cubra todos los campos que abarca la Estadística. Una posible definición sería: *ciencia encargada de recoger, describir e interpretar los datos relativos a un conjunto de elementos, permitiendo hacer inferencias científicas a partir de ellos*.

Hay dos grandes tipos de Estadísticas. La *Estadística descriptiva* tiene por objeto el estudio y descripción de determinadas características de una muestra o población dada. La *Estadística inferencial* tiene por objeto estudiar la extensión de los datos de una muestra a toda la población, así como la evolución previsible de los caracteres estudiados. En este tema solo nos ocuparemos de la estadística descriptiva

Estadística descriptiva

Los conceptos generales que utiliza son los siguientes.

- Población: conjunto de personas, cosas, etc., objeto específico de nuestro estudio.
- Individuo: cada uno de los elementos que componen la población.
- Muestra: subconjunto (representativo) de la población.
- Tamaño de la población o de la muestra: número de individuos que componen la población o la muestra elegida.
- Carácter estadístico: propiedad observable que poseen todos los individuos de la población objeto de estudio. Por ejemplo, si la población son un grupo de personas, los caracteres a observar pueden ser: sexo, edad, estatura, etc.
- Censo: conjunto de datos recogidos al observar todos y cada uno de los individuos que componen la población.
- Estadística: conjunto de datos recogidos al observar todos y cada uno de los individuos que componen una muestra.

Caracteres estadísticos

Los caracteres estadísticos pueden ser:

- Cualitativos: el carácter no se puede medir (sexo, nacionalidad...).
- Cuantitativos: el carácter es medible o cuantificable (edad, estatura...).

A su vez, los caracteres cuantitativos pueden ser:

- Discretos: solo puede tomar un conjunto finito o numerable de valores (generalmente, números enteros) sin que puedan tomar valores intermedios (por ejemplo, número de trabajadores en una empresa).
- Continuos: pueden tomar cualquier valor en un intervalo, finito o infinito (por ejemplo, la altura de los alumnos de un instituto).

Modalidades

Los valores que puede tomar un carácter estadístico se denominan: *modalidades*. Por ejemplo: El carácter sexo puede tomar las modalidades: hombre, mujer. El carácter edad puede tomar como modalidades cualquier número entero comprendido entre, digamos, 0 y 150 años. El carácter estatura puede tomar como modalidades cualquier número decimal comprendido entre, digamos, 0 y 3 m.

Ejemplo. *Elaboración de un censo de los institutos existentes en la isla de Tenerife con vistas a elaborar un plan de mantenimiento.*

Deseamos conocer en qué municipio se encuentran, cuántos alumnos tienen matriculados y la superficie construida de cada una de ellas. Para ello es necesario localizar y recabar los datos de todas y cada una de las unidades existentes en la isla.

- Población = institutos existentes en Tenerife.
- Individuo = cada uno de estos institutos.
- Tamaño de la población = a determinar por el censo.
- Caracteres estadísticos observados:
- X = municipio al que pertenece el instituto.
- Y = número de alumnos matriculados.
- Z = superficie construida.
- El carácter X es cualitativo. Puede adoptar como modalidad cualquier municipio de la isla.
- El carácter Y es cuantitativo (variable estadística). Puede adoptar como modalidad cualquier número entero comprendido entre, digamos, 1 y 500.
- El carácter Z es cuantitativo (variable estadística). Puede adoptar como modalidad cualquier número decimal comprendido entre, digamos, 0 y 10.000 m^2.

Fases en la elaboración de una estadística

Las fases que se siguen normalmente en la elaboración de una estadística son las siguientes:

- Si la población es demasiado numerosa, elección de una muestra representativa.
- Recolección de los datos.
- Tablas de frecuencias y representaciones gráficas.
- Medidas de centralización y dispersión.
- Presentación e interpretación de los datos.

A continuación, trataremos cada uno de estos puntos.

Elección de muestras representativas

Si una población es demasiado numerosa, se hace necesario limitar el estudio a una muestra. Dicha muestra deberá elegirse cuidadosamente para que represente lo más fielmente posible a la población total. Para ello deberá tenerse en cuenta tanto su tamaño como la inclusión en proporción adecuada de representantes de todos los sectores de la población.

La elección de la muestra representa a menudo un problema delicado pues un error en esta fase falsea todo el proceso. Además, esta elección proporciona una oportunidad inmejorable de sesgar los resultados de la investigación a favor (o en contra) de alguna de las opciones.

Pero ¿cómo se puede saber si una muestra representa o no a la población total si, precisamente, hemos elegido una muestra porque no resultaba posible observar a toda la población?

Para resolver este problema se utiliza la *Teoría de probabilidades*. Su estudio desborda las posibilidades de este curso por lo que sólo añadiremos que esta teoría nos permite establecer el grado de fiabilidad de la muestra dentro de un margen de error determinado.

Recolección de datos

La recolección de los datos se lleva a cabo mediante múltiples procedimientos: encuestas directas, consultas de archivos y documentos, realización de experimentos, etc.

Tablas de frecuencias

Las tablas de frecuencias sirven para ordenar y resumir los datos obtenidos y las frecuencias con que aparecen. Nos permiten hacernos una idea general de la situación de forma rápida y cómoda. Normalmente vienen acompañadas de una *ficha técnica* donde se proporciona el título de la tabla, fuentes en que se basa la información, notas aclaratorias, etc.

La configuración de la tabla dependerá del tipo de carácter estudiado, así como de los objetivos finales deseados. En general, distinguiremos dos tipos de situaciones:

- La variable toma un número razonable de valores distintos.
- La variable toma un número excesivo de valores distintos.

Cada uno de estas dos situaciones recibe un tratamiento estadístico diferente. Veamos antes algunas definiciones generales.

Definiciones

Sea una población formada por N individuos descrita según un carácter X que toma los valores $x_1, x_2, x_3, \dots, x_k$ (ordenados de menor a mayor). Para cada modalidad x_i $(i = 1, 2, \dots, k)$:

- La frecuencia absoluta (n_i) de la modalidad x_i será el número total de veces que se repite el valor x_i en el conjunto de la población.
- Frecuencia relativa (fi) de la modalidad x_i será el cociente entre su frecuencia absoluta y el tamaño de la población. Es decir, $f_i = n_i/N$.

Si el carácter es cuantitativo (o cualitativo ordenable):

- La frecuencia absoluta acumulada (N_i) de la modalidad x_i será la suma de las frecuencias absolutas de las modalidades inferiores o iguales a ella. Es decir, $N_i = n_1 + n_2 + \cdots + n_i = \sum_{j=1}^{i} n_j$.
- La frecuencia relativa acumulada (F_i) de la modalidad x_i será la suma de las frecuencias relativas de las modalidades inferiores o iguales a ella. Es decir, $F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$.

Tabla de frecuencias de un carácter cualitativo

Se ordenan las modalidades x_i del carácter X por el criterio que se desee (alfabético, frecuencias, etc.), y se completa la tabla siguiente:

X	FA	FR	%	Sector (º)
x_1	n_1	f_1	c_1	g_1
x_2	n_2	f_2	c_2	g_2
...	...	...	...	...
x_k	n_k	f_k	c_k	g_k
	$\Sigma = N$	$\Sigma = 1$	$\Sigma = 100$	$\Sigma = 360º$

Nota: $c_i = 100 \cdot f_i$ y $g_i = 360 \cdot f_i = 3.6 \cdot c_i$.

Tabla de frecuencias de una variable que toma pocos valores distintos

Cuando la variable toma un número razonable de valores distintos, se ordenan los valores de menor a mayor $(x_1 < x_2 < \cdots < x_k)$ y se completa la siguiente tabla:

X	FA	FR	FAA	FRA	%	Sector (º)
x_1	n_1	f_1	N_1	F_1	c_1	g_1
x_2	n_2	f_2	N_2	F_2	c_2	g_2
...	...	...	...	...	...	...
x_k	n_k	f_k	N_k	F_k	c_k	g_k
	$\Sigma = N$	$\Sigma = 1$	$N_k = N$	$F_k = 1$	$\Sigma = 100$	$\Sigma = 360º$

Variable agrupada en intervalos de clase

Cuando la variable estadística toma un número excesivo de valores distintos no resulta conveniente manejar todos los valores por separado. Lo que se hace en estos casos es recubrir el intervalo que recorre la variable por un cierto número de intervalos, llamados *intervalos de clase,* y agrupar las observaciones que caen dentro de un mismo intervalo. En lo que sigue, los intervalos de clase serán siempre semiabiertos por la izquierda, es decir, del tipo $(a, b]$. Se asume que las observaciones agrupadas en un intervalo de clase se encuentran uniformemente distribuidas por dicho intervalo.

Elección de los intervalos de clase

Una vez determinado el intervalo cerrado $[a, b]$, cuyos extremos son los valores máximos y mínimos de la variable, procederemos a recubrirlo por un número adecuado de intervalos semiabiertos por la izquierda que no se solapen.

La elección del número y amplitud de estos intervalos debe hacerse de forma cuidadosa, intentando guardar el equilibrio entre la simplicidad y la pérdida de información. Por comodidad, se tiende a trabajar con intervalos de la misma longitud, salvo que las circunstancias del caso lo desaconsejen.

- La frecuencia absoluta (n_i) del intervalo de clase $(a_{i-1}, a_i]$ será el número total de valores de la variable que caen dentro de ese intervalo.
- La marca de clase (x_i) del intervalo $(a_{i-1}, a_i]$ será el punto medio de dicho intervalo. La marca de clase de un intervalo es un representante de la información que contiene el intervalo.

La disposición de la tabla de frecuencias queda de la siguiente manera:

Intervalo	Marca de clase	FA	FR	FAA	FRA	%
$[a_0, a_1]$	x_1	n_1	f_1	N_1	F_1	c_1
$(a_1, a_2]$	x_2	n_2	f_2	N_2	F_2	c_2
$(a_3, a_4]$	x_3	n_3	f_3	N_3	F_3	c_3
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(a_{k-1}, a_k]$	x_k	n_k	f_k	N_k	F_k	c_k
		$\Sigma = N$	$\Sigma = 1$	$N_k = N$	$F_k = 1$	$\Sigma = 100$

Representaciones gráficas

A menudo, la información contenida en una tabla de frecuencias puede verse de manera más expresiva y directa reflejándola en un gráfico, especialmente si se desea comparar los datos. El tipo de gráficas que se utilice dependerá del tipo de carácter a representar. Los tipos de gráficos son muchos y variados. A continuación, veremos los más básicos.

Caracteres cualitativos

Gráfico de barras. Se representan las modalidades en el eje X y sobre cada una de ellas se levanta una barra cuya altura sea la frecuencia absoluta (o relativa) de la modalidad.

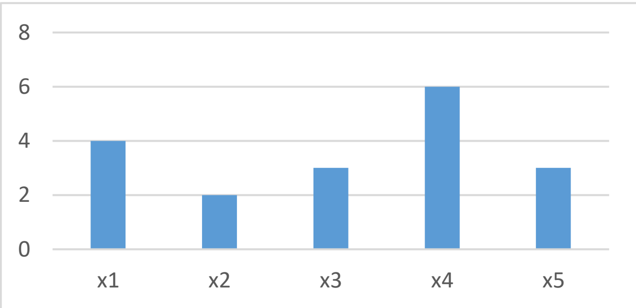


Gráfico de líneas. Se representan las modalidades en el eje X y las frecuencias absolutas (o relativas) en el eje de ordenadas. Los puntos correspondientes se unen con una línea.

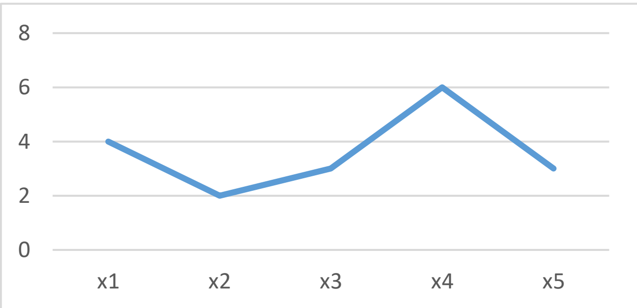
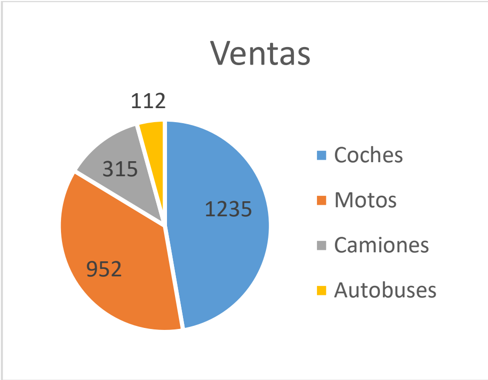


Gráfico circular. En un círculo se asigna un sector circular a cada una de las modalidades, siendo la amplitud del sector proporcional a la frecuencia absoluta o relativa de la modalidad.



Pictograma. Es una variante del diagrama de barras en la que en vez de rectángulos se dibujan figuras de tamaño proporcional a la frecuencia absoluta (o relativa) de cada modalidad, o bien se coge una figura estándar y se repite un número de veces proporcional a cada frecuencia absoluta (o relativa).

Cartograma. Consiste en la utilización de un mapa, señalando sobre él la frecuencia absoluta (o relativa) que corresponde a cada zona geográfica.

Variables que toman pocos valores distintos

Gráfico de barras. Se representan los valores que toma la variable en el eje X y sobre cada uno de ellos se levanta una barra cuya altura sea la frecuencia absoluta (o relativa) de cada modalidad.

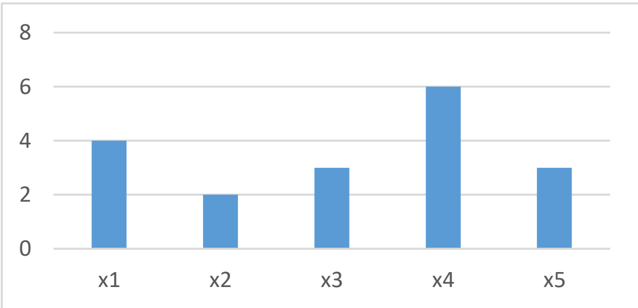


Gráfico de líneas. Se representan los valores que toma la variable en el eje X y las frecuencias absolutas (o relativas) en el eje de ordenadas. Los puntos correspondientes se unen con una línea.

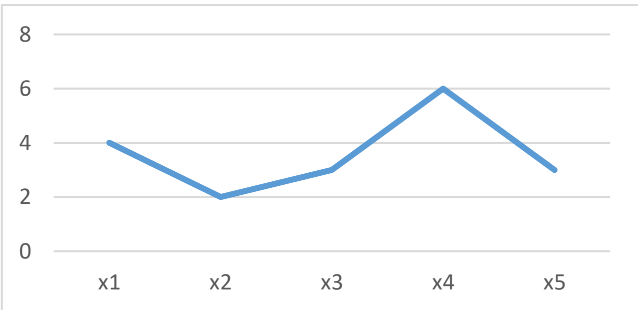
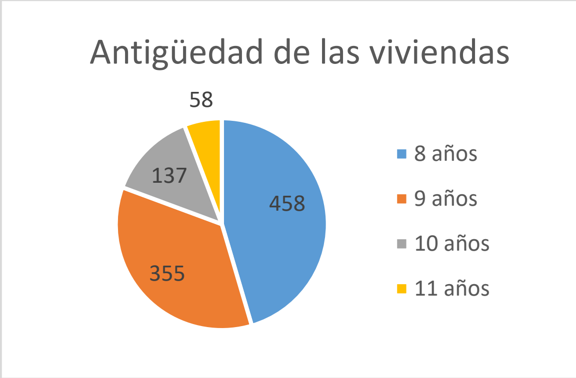
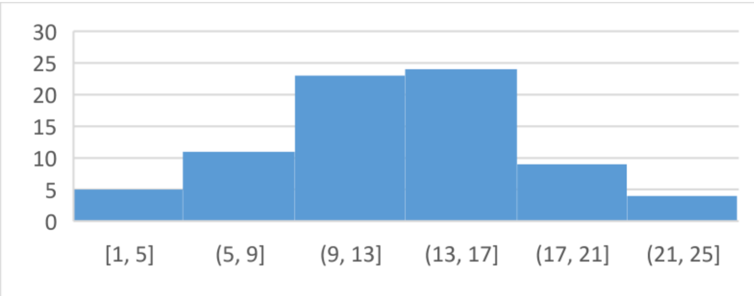


Gráfico circular. En un círculo se asigna un sector circular a cada una de los valores que toma la variable, siendo la amplitud del sector proporcional a la frecuencia absoluta o relativa de la modalidad.



Variables agrupadas en intervalos

Histograma. En un sistema de ejes coordenados, se colocan en el eje X los distintos intervalos y se levanta sobre cada uno de ellos un rectángulo de área proporcional a la frecuencia absoluta (o relativa) del intervalo. Si los intervalos tienen la misma amplitud, puede tomarse como altura del rectángulo la frecuencia absoluta (o relativa) correspondiente.



Gráficos múltiples o combinados

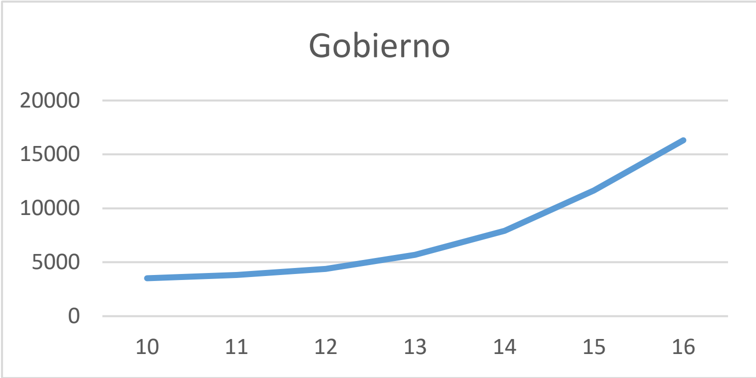
Se utilizan cuando se desean comparar los datos de diferentes caracteres estudiados en una misma población, o bien los datos de un mismo carácter estudiado en distintas poblaciones. En las hojas de problemas pueden verse algunos ejemplos.

Gráficos que inducen a error

Debemos advertir que las representaciones gráficas pueden ser manipuladas fácilmente para acentuar una determinada interpretación de los hechos. Esta manipulación puede llevarse a cabo de diversas maneras, por ejemplo, la que se presenta a continuación.

Ejemplo. La siguiente tabla muestra el paro en el sector de la construcción entre 2010 y 2016.

Año	2010	2011	2012	2013	2014	2015	2016
Parados	3517	3814	4384	5691	7918	11659	16314



Medidas de centralización y dispersión

Una vez recogidos, tabulados y representados gráficamente los datos, procederemos a obtener otros números que sinteticen aún más la información, facilitándonos la comparación entre las distintas muestras o poblaciones. Estos números se denominan *parámetros estadísticos*. Veremos dos tipos de parámetros estadísticos: *parámetros de centralización* y *parámetros de dispersión*.

Parámetros de centralización

Tienen como objetivo agrupar o centralizar los datos correspondientes a toda la población en un sólo valor numérico que represente el conjunto total de valores. Los más importantes son *moda*, *mediana* y *media*.

Sea una población formada por N individuos, descrita según un carácter X que toma los valores $x_1, x_2, x_3, \dots x_k$ (ordenados de menor a mayor), con frecuencias $n_1, n_2, n_3, \dots n_k$, respectivamente. En estas condiciones:

Moda (Mo)

Es la modalidad que posee la frecuencia más alta. Si todas las modalidades tienen la misma frecuencia se dice que no hay moda. En el caso de variables agrupadas en intervalos de clase, se llama c*lase modal* al intervalo con la frecuencia más alta, tomándose como moda su marca de clase.

Este parámetro no es muy significativo, pero resulta útil en algunas ocasiones, especialmente cuando su frecuencia destaca claramente sobre el resto.

Media aritmética (solo para variables estadísticas)

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{N} = \sum_{i=1}^{k} x_i \frac{n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

En el caso de variables agrupadas en intervalos, se toma como x_i la marca de clase del intervalo.

Media geométrica (sólo para variables estadísticas)

$$\overline{x_g} = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdots x_k^{n_k}}\,.$$

En el caso de variables agrupadas en intervalos se toma como x_i la marca de clase del intervalo. La media geométrica tiene el inconveniente de ser cero desde que alguno de los valores x_i lo sea, en cuyo caso no es representativa.

Mediana (solo para variables estadísticas)

Supuestos los N valores de la variable ordenados de forma creciente, llamaremos *mediana* al valor de la variable que deja el mismo número de observaciones menores que superiores a ella.

Variable discreta

- Si N es impar la mediana será el valor central.
- Si N es par la mediana será la media aritmética de los dos valores centrales.

Variable agrupada en intervalos

Supuestos los N valores de la variable distribuidos uniformemente a lo largo de cada uno de los intervalos de clase, la mediana será el valor de la variable que parte por la mitad el conjunto de observaciones. Su valor exacto se calcula mediante una proporción sobre el número de observaciones existentes en el intervalo que contiene la observación que divide en dos el número de observaciones. El procedimiento es el siguiente:

Calculamos $N/2$ y miramos en la tabla de frecuencias la columna de las FAA para ver qué intervalo contiene la observación $N/2$. Supongamos que sea el intervalo $(a_{i-1}, a_i]$. Pueden ocurrir dos cosas:

- $N/2$ coincide con la frecuencia absoluta acumulada del intervalo. En este caso se toma como mediana su extremo superior. Es decir, $Me = a_i$.
- $N/2$ no coincide con la frecuencia absoluta acumulada del intervalo. En este caso el valor de la variable que ocupa el lugar $N/2$ caerá en el interior del intervalo $(a_{i-1}, a_i]$. Su posición exacta la hallaremos mediante una proporción. Es decir, si en el intervalo hay $(N_{i-1} - N_i)$ observaciones repartidas uniformemente, la mediana será el punto que ocupe el lugar $\frac{N}{2} - N_{i-1}$. Es decir:

$$\frac{N_i - N_{i-1}}{a_i - a_{i-1}} = \frac{\frac{N}{2} - N_{i-1}}{Me - a_{i-1}} \Rightarrow Me = a_{i-1} + \frac{\left(\frac{N}{2} - N_{i-1}\right)(a_i - a_{i-1})}{(N_i - N_{i-1})}.$$

Cuartiles (solo para variables estadísticas)

Son las tres medidas Q_1, Q_2, Q_3 que, supuestos los N valores de la variable ordenados de forma creciente, dividen las observaciones en cuatro partes iguales.

Primer cuartil (Q_1): Es la medida que deja $1/4$ parte de las observaciones menores que él y las $3/4$ partes superiores a él.

Segundo cuartil (Q_2): Es la medida que deja $1/2$ parte de las observaciones menores que él y $1/2$ parte superiores a él. Coincide con la mediana. Es decir, $Q_2 = Me$.

Tercer cuartil (Q_3): Es la medida que deja 3/4 parte de las observaciones menores que él y $1/4$ parte superiores a él.

A diferencia de la mediana, que se calcula siempre mediante la misma la fórmula, en la literatura existen diferentes definiciones del primer y tercer cuartil.

Percentiles (solo para variables estadísticas)

Son las cien medidas que, supuestos los N valores de la variable ordenados de forma creciente, dividen las observaciones en cien partes iguales. Sus valores se calculan análogamente a los cuartiles. Nótese que los cuartiles son los percentiles 25, 50 y 75, respectivamente.

Parámetros de dispersión

En general las medidas de centralización no son suficientes para darnos una idea fiable del comportamiento de los datos. Necesitamos entonces otros parámetros que nos permitan saber cómo están distribuidos los datos en torno al valor central. Estos parámetros son los *parámetros de dispersión*. Los parámetros de dispersión principales son el *recorrido*, la *varianza* y la *desviación típica.*

Sea una población formada por N individuos, descrita según una variable estadística X que toma los valores $x_1, x_2, x_3, \dots, x_k$ (ordenados de menor a mayor), con frecuencias $n_1, n_2, n_3, \dots, n_k$.

Recorrido

Diferencia entre los valores máximo y mínimo de la variable. Es fácil de calcular, pero proporciona poca información. En el caso de variables agrupadas, el *recorrido* es la diferencia entre el extremo inferior del primer intervalo y el extremo superior del último intervalo de clase.

Varianza

Media aritmética de los cuadrados de las desviaciones de los datos respecto de la media aritmética.

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}$$

Cuando la variable está agrupada en intervalos, x_i representa la marca de clase.

Propiedades

- $\sigma^2 \geq 0$ siempre.
- $\sigma^2 = 0 \Rightarrow$ La variable toma un sólo valor para toda la población.

La varianza presenta el inconveniente de que sus unidades no son las mismas en que se han medido los datos, sino que están elevadas al cuadrado para evitar que se anulen entre sí las diferencias positivas y negativas. Por ello introducimos una nueva medida de dispersión más útil: la desviación típica.

Desviación típica

Raíz cuadrada positiva de la varianza:

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}}$$

Otros parámetros de dispersión

Desviación media respecto de la media aritmética: $D_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| \cdot n_i}{N}$.

Desviación media respecto de la media aritmética: $D_{Me} = \frac{\sum_{i=1}^{k} |x_i - Me| \cdot n_i}{N}$.

Nota: Si eliminamos los valores absolutos, las desviaciones positivas y negativas se contrarrestarían, falseando el resultado.

Coeficiente de variación de Pearson

Las medidas de dispersión que hemos definido hasta el momento vienen dadas con las mismas unidades en que vienen las variables por lo cual no son útiles para comparar las dispersiones de dos muestras expresadas en distintas unidades (por ejemplo, si en una misma población estudiamos el peso y la altura). Para ello se introducen los coeficientes de variación, pues proporcionan números abstractos comparables unos con otros. El principal es el coeficiente de variación de Pearson.

Coeficiente de variación de Pearson: $CVP = \frac{\sigma}{|\bar{x}|}$.

Otros coeficientes de variación son:

Coeficiente de variación media respecto de la media aritmética: $CVM_{\bar{x}} = \frac{D_{\bar{x}}}{|\bar{x}|}$.

Coeficiente de variación media respecto de la mediana: $CVM_{Me} = \frac{D_{Me}}{|Me|}$.

Nota: Cuanto más bajos sean los valores de los coeficientes de variación, menos dispersas estarán las distribuciones.

Ejemplo 1

La tabla muestra el estado civil de los empleados de una empresa de la construcción. Hacer la tabla de frecuencias. Representar la información de tres maneras distintas. Calcular la moda. ¿Qué otros parámetros de centralización y dispersión se pueden calcular?

Estado	Soltero	Casado	Separado	Divorciado	Viudo
Empleados	35	44	27	15	27

x_i	n_i	f_i	%	Sector

Diagrama combinado de barras y líneas

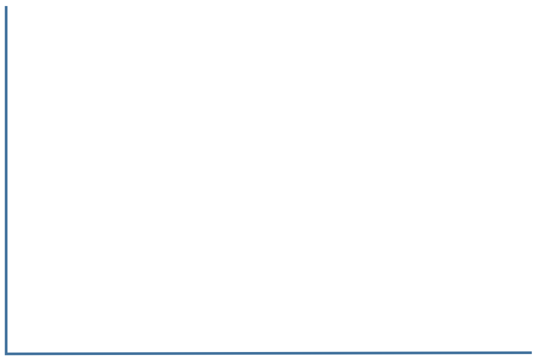
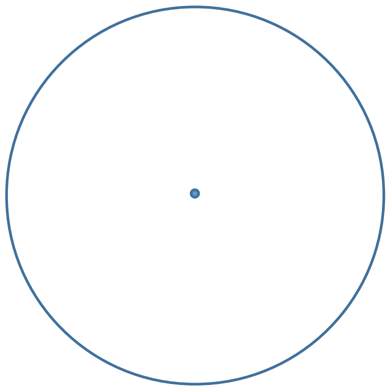


Diagrama de sectores



Parámetros de centralización y dispersión

Ejemplo 2

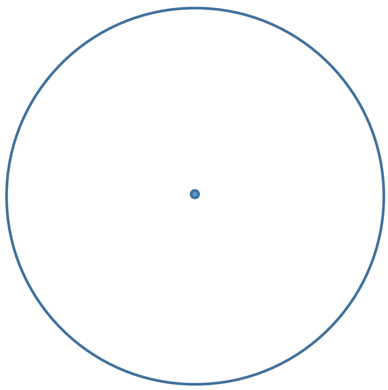
La tabla muestra los años trabajados por 40 empleados de una empresa de construcción. Hacer la tabla de frecuencias. Hacer las dos representaciones gráficas que se indican. Calcular los parámetros de centralización y dispersión.

8	8	9	9	9	10	10	11	11	12
8	8	9	9	10	10	11	11	11	12
8	8	9	9	10	10	11	11	11	12
8	9	9	9	10	10	11	11	12	12

x_i	n_i	f_i	%	Sector	$x_i \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$

Diagrama de barras

Diagrama de sectores



Parámetros de centralización y dispersión

Ejemplo 3

La tabla muestra los pesos en kg de un cierto número de bloques de cemento. Agrupar los datos en intervalos de amplitud 10 kg. Hacer la tabla de frecuencias, el histograma y calcular las medidas de centralización y dispersión.

30.0	42.1	43.9	47.8	50.3	53.4	59.4	62.9	65.0	76.0
35.5	42.6	45.6	49.5	52.1	53.8	60.2	63.5	65.3	77.8
37.8	42.6	47.6	49.7	52.5	56.3	60.8	64.5	71.0	80.0

Intervalos	M. de c.	n_i	f_i	%	$x_i\, n_i$	$(x_i - \bar{x})^2 n_i$

Histograma

Parámetros de centralización y dispersión