Tema 1

Introducción.
Estadística descriptiva

Introducción. Estadística descriptiva

- Introducción
- Distribución de frecuencias
- Representaciones gráficas
- Medidas descriptivas

Estadística

Engloba tres concepciones gramaticales diferentes:

- **Colección** de datos presentada de forma ordenada y sistemática.
- **Técnica** de recoger, organizar, resumir, presentar, analizar, generalizar y contrastar resultados de una investigación.
- **Ciencia** que busca las leyes generales del comportamiento de colectivos en los aspectos que dependen del azar.

Objeto de la Estadística

- La Estadística estudia **fenómenos de naturaleza aleatoria**, es decir, en los que interviene el azar y no es posible predecir el resultado aunque se den las mismas condiciones iniciales.
- Se distingue así de otras ciencias que estudian **fenómenos deterministas,** en los que la misma causa provoca siempre el mismo efecto.

Conceptos I

- Se llama **población** al colectivo sobre el que se estudia el fenómeno aleatorio. Su tamaño es **N**.
- Cada subconjunto de la población elegido en representante de ésta se denomina **muestra**. Su tamaño es **n**.
- Los miembros concretos de la población se llaman **individuos** o unidades elementales.

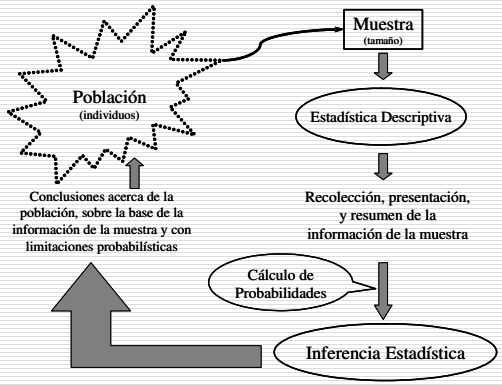
Conceptos II

- Si se estudian todos y cada uno de los individuos de la población se habla de un **censo**. En caso contrario se habla de **muestreo**.
- Se llama **parámetro** a la cantidad que recoge algún aspecto relevante de la población.
- Se llama **estadístico** a la cantidad que recoge algún aspecto relevante de la muestra.

Partes de la Estadística

- La **Estadística Descriptiva** tiene por objeto describir la información de la muestra o de la población.
- La **Inferencia Estadística** extrae conclusiones de la población a partir de las resultados obtenidos en una muestra de ella.

Esquema partes de la Estadística



Caracteres

- Se denomina **carácter** a la propiedad que se analiza en la población investigada.

Los hay de dos tipos:

- Caracteres **cualitativos o atributos**
- Caracteres **cuantitativos o variables**

Caracteres cualitativos

Los caracteres **cualitativos (atributos o factores)** no se pueden expresar numéricamente.

Las posibles respuestas se llaman modalidades.

Los atributos se clasifican en:

- **Caracteres nominales:** no se puede establecer una relación de orden natural entre sus modalidades.
- **Caracteres ordinales:** sus modalidades no son cuantificables, pero es posible establecer un orden entre ellas.

Caracteres cuantitativos

- Son el objeto de nuestro estudio.
- Los caracteres **cuantitativos o variables** se pueden expresar numéricamente. Las posibles respuestas son los valores de la variable. A su vez se clasifican en:
  - **Variable continua**: entre cada dos valores de la variable siempre se podría encontrar un valor intermedio. Toma valores en (al menos) un intervalo.
  - **Variable discreta:** toma un nº finito o numerable de valores.

Organización de datos

- El resultado de la observación de una **variable** en la población (o en una muestra) es un **conjunto de datos**.
- Denotaremos por **N** el número total de datos.
- Los resultados **diferentes** obtenidos en las N experiencias y ordenados de **forma creciente** son los números $x_1, x_2, ..., x_k$.
- x_1=**Mín(X)** y x_k= **Máx(X).**

Frecuencia absoluta

- La **frecuencia absoluta** de cada resultado x_i se denotará por $\mathbf{n_i}$ o por $\mathbf{n(x_i)}$, y es el **número** de veces que aparece dicho resultado en las N experiencias.
- Se cumple que $\sum_{i=1}^{k} n_i = N.$

Frecuencia relativa

- La **frecuencia relativa** asociada al resultado x_i se denotará $\mathbf{f_i}$ o por $\mathbf{f(x_i)}$, y es la **proporción** de veces que aparece dicho resultado en las N experiencias.
- Se verifica $f_i = f(x_i) = \frac{n_i}{N}.$
- Además $\sum_{i=1}^{k} f_i = 1.$

Porcentaje

- El **porcentaje** asociado al resultado x_i se denotará $\mathbf{p_i}$ o $\mathbf{p(x_i)}$, y es el **porcentaje** de veces que aparece dicho resultado en las N experiencias.
- Se verifica $p_i = p(x_i) = f(x_i) \times 100.$
- Además $\sum_{i=1}^{k} p_i = 100.$

Variable estadística

- El conjunto de pares (x_i, n_i) , (x_i, f_i) o (x_i, p_i) será desde ahora la **variable estadística** y con ellos se tiene la **distribución de frecuencias** (representada en **la tabla estadística**).
- Las variables estadísticas se denotan por **X** (o Y, Z, …) y representan la propiedad de estudio (si es preciso con unidades de medida) en cada uno de los individuos de la muestra.

Ejemplo

- Un fabricante de componentes electrónicos se interesa en determinar el tiempo de vida (duración) de cierto tipo de batería. La que sigue es una muestra en horas de vida:

  123, 116, 122, 110, 135, 126, 125, 111, 118, 116

  - X = tiempo de vida (en horas) de cada una de las baterías. Variable de tipo continuo
  - N=10

Tabla estadística del ejemplo

x_i	$n(x_i)$	$f(x_i)$	$p(x_i)$
110	1	0.1	10%
111	1	0.1	10%
116	2	0.2	20%
118	1	0.1	10%
122	1	0.1	10%
123	1	0.1	10%
125	1	0.1	10%
126	1	0.1	10%
135	1	0.1	10%
	10	1	100%

Porcentaje acumulado

- El **porcentaje acumulado**, **P(x)**, es el porcentaje de valores menores o iguales que el considerado.
- En general

$$P(x) = \sum_{\{i / x_i \le x\}} p(x_i).$$

- En particular

$$P(x_i) = \sum_{j=1}^{i} p(x_j).$$

Ejemplo

x_i	$n(x_i)$	$f(x_i)$	$p(x_i)$	$P(x_i)$
110	1	0.1	10%	10%
111	1	0.1	10%	20%
116	2	0.2	20%	40%
118	1	0.1	10%	50%
122	1	0.1	10%	60%
123	1	0.1	10%	70%
125	1	0.1	10%	80%
126	1	0.1	10%	90%
135	1	0.1	10%	100%
	10	1	100%	

- **P(126) =** porcentaje de baterías con duración no superior a 126 horas = 90%.
- **P(120) =** porcentaje de baterías con duración no superior a 120 horas = p(X≤120) = P(118) = 50%.

Representaciones gráficas

Según los datos que manejemos se pueden realizar las gráficas siguientes:

- Diagrama de sectores
- Diagrama de barras
- Histograma
- Diagrama de cajas

Diagrama de sectores

- Se divide un círculo en tantos sectores circulares como valores toma la variable. A cada sector se le asigna un ángulo central proporcional a su frecuencia absoluta (relativa o porcentaje).
- Es la gráfica más adecuada para representar caracteres nominales.

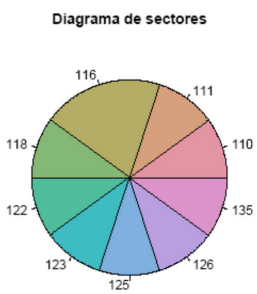
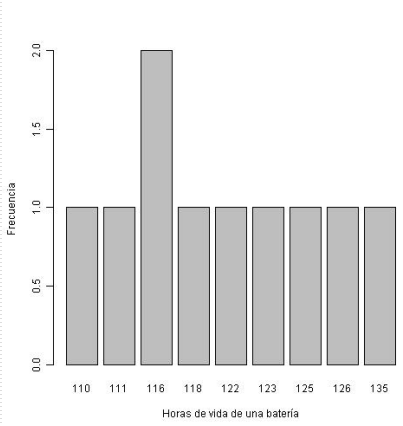


Diagrama de barras

- Se construye levantando barras horizontales de la misma anchura sobre cada valor de la variable. La altura de cada barra es su frecuencia absoluta (relativa o porcentaje).
- Es la gráfica más adecuada para representar caracteres ordinales o variables discretas.

Diagrama de barras del ejemplo



Histograma

- Sobre cada uno de los intervalos en los que se subdividen los datos de la variable se levanta una barra cuya altura es su frecuencia absoluta (relativa o porcentaje) del intervalo. También se puede utilizar como altura la densidad (frecuencia del intervalo dividida por la amplitud del intervalo).
- Es una gráfica adecuada para representar variables continuas.

Histograma del ejemplo

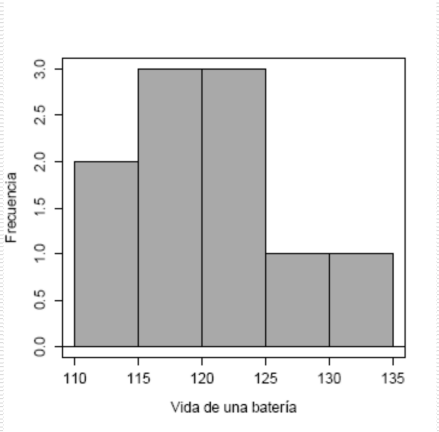


Diagrama de cajas

- Se verá asociada al concepto de mediana y cuartiles.
- Es una gráfica adecuada para representar variables continuas.

Medidas descriptivas

- Las medidas descriptivas son aquellas que tratan de condensar la información proporcionada por la distribución de frecuencias de la característica en estudio.
- Se subdividen en:
  - Medidas de centralización
  - Medidas de dispersión
  - Medidas de posición

Medidas de centralización

Su objetivo es **resumir** la información contenida en la tabla de frecuencias **en un solo número**.

Manejaremos:

- **Media aritmética**
- **Mediana**
- **Moda**

Media aritmética

- Se calcula como el **promedio** de los datos:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i = \frac{1}{N} \sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} \frac{x_i n_i}{N}.$$

- A la cantidad $\sum_{i=1}^{k} x_i n_i$ se le denomina **total**.
- La media se ve muy afectada por valores extremos.
- La media es un valor único.

Propiedades de la media aritmética

1. $\sum_{i=1}^{k} (x_i - \bar{x}) f_i = 0$ (la **media de las desviaciones** con respecto a la media **es 0**). Se interpreta como que es el centro de gravedad o punto de equilibrio de la distribución de frecuencias.
2. Si la variable X se transforma de la **forma lineal** Y=aX+b, entonces,

$$\bar{y} = a\bar{x} + b.$$

Ejemplo

- La siguiente tabla muestra los resultados obtenidos en diferentes mediciones (en mm.) del diámetro interno de unos anillos para pistones de automóviles.
  - X=diámetro interno (en mm.) de un pistón.
  - N=8

x_i	n_i	f_i	$P(x_i)$
74.000	1	0.125	12.5%
74.001	1	0.125	25.0%
74.015	1	0.125	37.5%
74.017	1	0.125	50.0%
74.018	1	0.125	62.5%
74.020	1	0.125	75.0%
74.024	1	0.125	87.5%
74.029	1	0.125	100.0%
	8	1	

Cálculo media del ejemplo

- El diámetro interior medio de los anillos es 74.0155 mm. El total es 592.124 mm.
- Si las especificaciones del fabricante dicen que el diámetro interior de los anillos es 74.000 mm. y por cada *milésima de mm. de error* (respecto a lo especificado) supone un coste de 5€. el coste por pieza es

$$C=5000\times|X-74.000|=5000(X-74.000)$$

y por tanto, el coste medio por pieza

$$\overline{C}=5000\times(\overline{X}-74.000)=5000\times(74.0155-74.000)=77.5€.$$

Mediana

- La mediana, Me, es el valor central de los valores de una variable, una vez que éstos han sido ordenados de forma creciente.
- Divide a la distribución en dos partes con la misma frecuencia: cada parte al menos N/2 en términos absolutos (½ en términos relativos, un 50%).
- Al menos el 50% de los datos son menores o iguales que Me (al menos el 50% mayores o iguales que él).

Cálculo de Me

Se calcula utilizando los porcentajes acumulados, $P(x_i)$.

Se puede estar en dos situaciones:

- Hay un valor de la variable x_i con $P(x_i)=50\%$, entonces la mediana sería cualquier valor de $[x_i, x_{i+1}]$ y el representante más habitual:

$$Me = \frac{x_i + x_{i+1}}{2}.$$

- En caso contrario, la mediana es el primer valor x_i con porcentaje acumulado superior al valor 50%.

Cálculo Me del ejemplo

Hay un valor x_i de la variable con porcentaje acumulado 50% ($P(x_4)=50\%$), entonces

$$Me = \frac{x_4 + x_5}{2} = \frac{74.017 + 74.018}{2} = 74.0175 \text{ mm.}$$

x_i	n_i	f_i	$P(x_i)$
74.000	1	0.125	12.5%
74.001	1	0.125	25.0%
74.015	1	0.125	37.5%
74.017	1	0.125	50.0%
74.018	1	0.125	62.5%
74.020	1	0.125	75.0%
74.024	1	0.125	87.5%
74.029	1	0.125	100.0%
	8	1	

Moda

La moda, Mo, es el valor de **máxima frecuencia,** sea ésta frecuencia la absoluta, la relativa o la porcentual.

- Puede no ser única.
  - **Ejemplo:** En la distribución anterior **todos** los valores son moda, ya que todos tienen la frecuencia máxima: 1.

Medidas de posición

- Son medidas que informan de la posición de una dato en la muestra.
- La idea es la misma que en la **mediana**, pero descomponiendo la distribución en cuatro, diez, o cien partes con la misma frecuencia (**cuartiles, deciles, percentiles**).

Cuartiles, deciles

- **Cuartiles**: dividen la distribución en cuatro intervalos con la misma frecuencia (N/4, 0.25 ó 25%). Los cuartiles son 3: P_{25}, P_{50}, P_{75}.
- **Deciles**: dividen la distribución en diez intervalos con la misma frecuencia (N/10, 0.10 ó 10%). Los deciles son 9: P_{10}, P_{20}, …, P_{90}.

Percentiles

- **Percentiles (centiles):** dividen la distribución en cien intervalos con la misma frecuencia (N/100, 0.01 ó 1%). Los percentiles son 99: P_1, P_2, …, P_{99}.
- El **percentil de orden k**, $\mathbf{P_k}$, es aquel valor de la variable que (una vez ordenados los datos de forma creciente) hace que al menos el k% de los datos sean menores o iguales que él (al menos el 100-k% mayores o iguales que él).

Cálculo de P_k

Se calcula utilizando los porcentajes acumulados, $P(x_i)$.

Se puede estar en dos situaciones:

- Hay un valor de la variable x_i con $P(x_i)=k\%$, entonces P_k sería cualquier valor de $[x_i, x_{i+1}]$ y el representante más habitual:

$$\mathbf{P_k} = \frac{x_i + x_{i+1}}{2}.$$

- En caso contrario, P_k es el primer valor x_i con porcentaje acumulado superior al valor k.

Cálculo de P_k en el ejemplo

- Primer cuartil P_{25}:

$P(x_2) = P(74.001)=25\%$. Por tanto,

$$P_{25} = \frac{74.001 + 74.015}{2} = 74.008 \quad \text{mm.}$$

- Tercer cuartil P_{75}:

$P(x_6) = P(74.020)= 75\%$. Por tanto,

$$P_{75} = \frac{74.020 + 74.024}{2} = 74.022 \quad \text{mm.}$$

Cálculo de P_k en el ejemplo

- Tercer decil P_{30}:

no existe x_i con $P(x_i) = 30\%$, el primer valor que supera 30% es x_3 .Por tanto:

$P_{30}= x_3 =74.015$ mm.

- Percentil 82 P_{82}:

no existe x_i con $F(x_i) = 82\%$, el primer valor que supera 82% es x_7 .Por tanto:

$P_{82}= x_7 =74.024$ mm.

Medidas de dispersión

Estudian la dispersión de los datos:

- **Dispersión global:**
  - Recorrido
  - Recorrido intercuartílico
- **Dispersión en torno a la media**:
  - Absoluta:
    - **Varianza**
    - **Desviación típica**
  - Relativa: **Coeficiente de variación**

Recorridos

- **Recorrido o Rango:** Es la diferencia entre el máximo y mínimo de los valores de la variable:

$$\mathbf{R = x_k - x_1.}$$

- **Recorrido intercuartílico**: es la diferencia entre el tercer y el primer cuartil:

$$R_I = \mathbf{P_{75} - P_{25}.}$$

  - Ejemplo: R=74.029-74.000=0.029 mm.

    R_I=74.022-74.008=0.014 mm.

Varianza

La varianza se calcula a partir de los cuadrados de las desviaciones de los datos respecto a la media:

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

- $S_X^2 \geq 0.$
- **<u>Inconveniente</u>**: maneja **unidades^2**.
- Una varianza grande indica mucha dispersión de datos alrededor de la media.

Desviación típica

La desviación típica o desviación estándar es la **raíz cuadrada positiva** de la varianza, se representa por $\mathbf{S_X}$**:**

$$S_X = +\sqrt{S_X^2}.$$

- **<u>Ventaja</u>:** maneja las unidades de partida.
- Una desviación típica grande indica mucha dispersión de datos alrededor de la media.

Cambio lineal

Si Y=aX+b, entonces:

$$S_Y^2 = a^2 S_X^2,$$

$$S_Y = |a|\, S_X,$$

es decir, la varianza y la desviación típica son invariantes frente a traslaciones.

Cálculo varianza del ejemplo

- En los datos del diámetro de los anillos, la varianza es: $S_X^2 = 0.0001\ \text{mm.}^2$
- Y la desviación típica es:

$$S_X = +\sqrt{S_X^2} = 0.0102\ \text{mm.}$$

- Mientras que para la variable coste en €:

$$C = 5000 \times |X - 74.000| = 5000(X - 74.000)$$

$$S_C^2 = 5000^2 \times S_X^2 = 2621.4286\ €^2,$$

$$S_C = 5000 \times S_X = 51.1999\ €.$$

Coeficiente de variación

El coeficiente de variación es una medida de dispersión relativa **adimensional** y se define por:

$$CV(X) = \frac{S_X}{|\bar{x}|}.$$

- El coeficiente de variación de **una transformación lineal Y = aX + b** es:

$$CV(Y) = \frac{|a|\, S_X}{|a\bar{x} + b|}.$$

Utilidad del coeficiente de variación

El CV se utiliza para:

- **Comparar la dispersión u homogeneidad** de los datos de dos distribuciones:

  Es menos dispersa o más homogénea la distribución con un CV menor.

- **Comparar la representatividad de la media** en dos distribuciones:

  Es más representativa la media de la distribución con un CV menor.

CV en el ejemplo

- En el problema del diámetro interior de los anillos se dispone de una **segunda muestra** de anillos en los que el diámetro interior medio es **25.97 mm**. y su desviación es **0.11 mm.**

- Si comparamos la dispersión de las dos distribuciones (o la representatividad de la media):
  - **1ª muestra**, CV(X)=0.0001, es decir tiene una dispersión relativa del **0.01%.**
  - **2ª muestra**, CV(Y)=0.0042, es decir tiene una dispersión relativa del **0.42%.**

➔ Por tanto, ***en la primera muestra el diámetro medio es más representativo que en la 2ª***, o ***los datos son más homogéneos (menos diferentes) o menos dispersos relativamente que en la 2ª***.

Diagrama de cajas

- Representación gráfica que muestra características de posición y dispersión de una variable.
- Para construirlo se traza una caja con límite inferior P_{25}, límite superior P_{75} y dentro de la caja una línea para indicar la Me. El recorrido intercuartílico R_I es la longitud de la caja.

  Además se trazan dos bigotes:
  - desde el límite inferior de la caja hasta el valor Máx(x_1,P_{25}-1.5R_I)
  - desde el límite superior de la caja hasta el valor Mín(x_k,P_{75}+1.5R_I)

  Los datos que se encuentran por fuera de los bigotes superior o inferior se consideran **valores atípicos**.

Diagrama de cajas del ejemplo