TEMA 3

ESTADÍSTICA DESCRIPTIVA

Introducción práctica
a la
Estadística Descriptiva
a través de
Ejercicios y Problemas.

José Álvarez Fajardo

Estadística Descriptiva

Nociones Generales

La **Estadística** es una ciencia que se dedica al estudio de ciertas técnicas numéricas que permiten conocer y analizar características de toda clase de objetos, partiendo de datos obtenidos de forma empírica: a través de experimentos o de encuestas.

La **población** es el conjunto de seres u objetos acerca de los que se desea obtener información.

Una **muestra** es una parte o subconjunto de la población que es examinada para obtener los datos, y cuyo estudio nos permitirá inferir las características que estemos estudiando de la población.

Se llama **individuo** a cada uno de los miembros de la población – o la muestra –. El **tamaño** de la población o de la muestra es el número de individuos que la forman.

Variables Estadísticas

Se denomina **variable estadística** a cada una de las características que se están estudiando en la población.

Las variables se clasifican en

- **Cualitativas**: si toman valores que no son numéricos.
- **Cuantitativas discretas**: si toman valores numéricos que pueden enumerarse.
- **Cuantitativas continuas**: si toman valores numéricos que pueden tomar cualquier cantidad de un cierto intervalo.

Los valores de las variables continuas se suelen agrupar en intervalos, denominados **clases**; al punto medio de cada intervalos se le llama **marca de clase**.

Frecuencia

La **frecuencia absoluta** – n – de un dato es el número de veces que aparece entre los valores que toma la variable estadística.

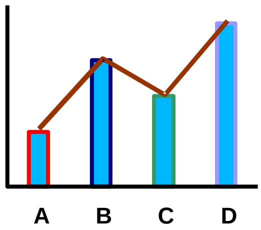
La **frecuencia relativa** – f – es el cociente de la absoluta entre el número total de datos – puede expresarse mediante porcentajes – .

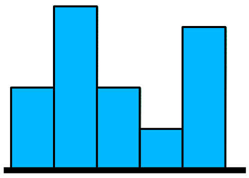
Las **frecuencias acumuladas** – N y F – son las sumas de las frecuencias de todos los datos hasta el que consideramos (incluido).

Tablas y Gráficos

Las tablas y gráficas persiguen clasificar y organizar los datos obtenidos, de modo que se visualice con facilidad la información.

Los **diagramas de barras** son adecuados para las variables cualitativas y discretas.

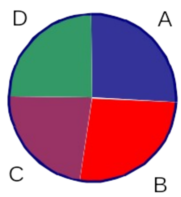




Los **histogramas de frecuencias** son adecuados para representar variables continuas.

El **polígono de frecuencias** es la poligonal que une el centro de cada barra rectangular con el de la siguiente.

Los **diagramas de sectores** sirven para comparar distintas distribuciones dadas en términos de porcentajes.



Estadística Descriptiva

Consideremos una distribución cuyos valores y frecuencias son:

x_i	x_1	x_2	...	x_l
n_i	n_1	n_2	...	n_l

Si los datos están agrupados en clases, x_i y n_i son las marcas y frecuencias de las respectivas clases.

Medidas de Centralización

Una medida de centralización es un valor que pretende resumir la diversidad de los datos, intentando ser lo más representativo posible del conjunto de los datos.

La **media aritmética** es el número dado por:

$$\bar{x} = \frac{\sum x_i n_i}{\sum n_i}$$

Si colocáramos los datos sobre una barra horizontal sería su punto de equilibrio o centro de gravedad.

La **moda** es el dato con mayor frecuencia. Con datos agrupados en clases, diremos que la moda es la marca de la clase con mayor frecuencia –la clase modal – .

Medidas de Posición

Los valores de centralización de variables cuantitativas dependen de la ordenación de los datos y no sólo del valor de ellos.

Procedimiento de cálculo:

1. Calculamos las frecuencias relativas acumuladas F_i y las expresamos en porcentajes.
2. Llamamos "**centil** k" o "**percentil** k" al valor p_k que, tras ordenar los datos de menor a mayor, deja antes de él al $k\,\%$ de ellos.
3. En las variables discretas, buscamos el primer dato x_i que cumpla $F_i > k$ y entonces pueden darse dos casos:

   a) $F_i > k$ y $F_{i-1} < k \longrightarrow p_k = x_i$

   b) $F_i > k$ y $F_{i-1} = k \longrightarrow p_k = \dfrac{x_{i-1} + x_i}{2}$

4. Si están los datos agrupados en clases, dicho percentil p_k se encuentra en la primera clase $[a_i\,, b_i)$ con $F_i > k$ y usaremos la fórmula:

$$p_k = a_i + (b_i - a_i)\,\frac{k - F_{i-1}}{F_i - F_{i-1}}$$

Se llaman **cuartiles** a los siguientes percentiles:

$$Q_1 = p_{25}\ ,\ Q_2 = p_{50}\ ,\ Q_3 = p_{75}$$

Y a p_{50} se le llama **mediana**.

Medidas de Dispersión

Las **medidas de dispersión** tratan de medir el grado de cercanía de los datos a su media.

El **recorrido** o **rango** es la diferencia entre el mayor y el menor de los datos.

La **desviación típica** viene dada por:

$$\sigma = \sqrt{\frac{\sum x_i^2 n_i}{\sum n_i} - \bar{x}^2}$$

Intenta cuantificar si los datos se acumulan en las proximidades de la media o no.

A su cuadrado se le llama **varianza**. Así:

$$\mathrm{VAR} = \sigma^2 \quad \text{ó} \quad \sigma = \sqrt{\mathrm{VAR}}$$

Interpretación Conjunta

En una distribución:

- La media $\bar{x}$ nos dice dónde está su centro,
- La desviación típica σ nos señala la dispersión de los datos.

Para poder comparar la dispersión de dos distribuciones distintas se usa el llamado **coeficiente de variación**:

$$\mathrm{C.V.} = \frac{\sigma}{\bar{x}}$$

Problemas de variables discretas

EJERCICIO 1: [RESUELTO]

En un grupo de 1º de Bachillerato se ha preguntado cuántos teléfonos móviles hay en sus hogares. Las respuestas han sido las siguientes:

2 3 4 5 6 7 2 3 3 7 7 5 3 5 5 5 4 4 4 5 4 4 4 4 6 6 6 6 6 6

a) ¿Cuál es la variable estadística estudiada? ¿De qué tipo es? Indica cuál es la población y determina su tamaño.

b) Construye la tabla en la que aparezcan todas las frecuencias. Señala en dicha tabla dónde obtener qué porcentaje de tiene cinco o menos.

c) Representa su diagrama de barras y el polígono de frecuencias.

d) Calcula la media y la moda.

e) Halla los cuartiles.

f) Obtén el recorrido y la desviación típica.

EJERCICIO 2: [RESUELTO]

En una feria ganadera los toros tenían un peso medio de 600 kg y una desviación típica de 50 kg., mientras que las ovejas tenían un peso medio de 60 kg con una desviación típica de 10 kg.

¿Qué variable diríamos que presenta mayor dispersión?

EJERCICIO 3:

En una granja de 1000 gallinas se toma una muestra para estudiar su puesta. La tabla siguiente siguiente nos muestra el número de aves correspondientes a los huevos de su puesta:

N.º huevos	3	4	5	6	7
N.º aves	6	10	20	9	5

a) ¿Cuál es la variable estadística estudiada? ¿De qué tipo es? Indica cuál es la población y determina su tamaño. Representa su diagrama de barras y el polígono de frecuencias.

b) Construye la tabla en la que aparezcan todas las frecuencias. Señala en dicha tabla dónde obtener qué porcentaje de gallinas que pone hasta cinco huevos.

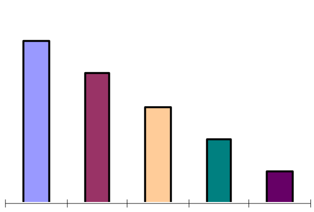
c) Calcula la media y la varianza.

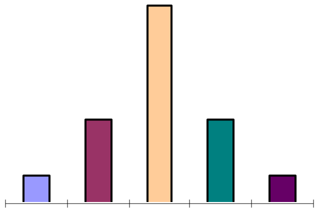
d) Halla la moda y los cuartiles.

e) Se ha medido también el peso de las gallinas, obteniendo un peso medio de $\overline{y} = 1.75\ \ \mathrm{kg}$ y una varianza de $\sigma_y^2 = 0.45\ \ \mathrm{kg}^2$. ¿Qué distribución es más dispersa, la de los huevos o la de los pesos de las gallinas?

EJERCICIO 4:

Los gráficos siguientes corresponden a dos distribuciones estadísticas X e Y:





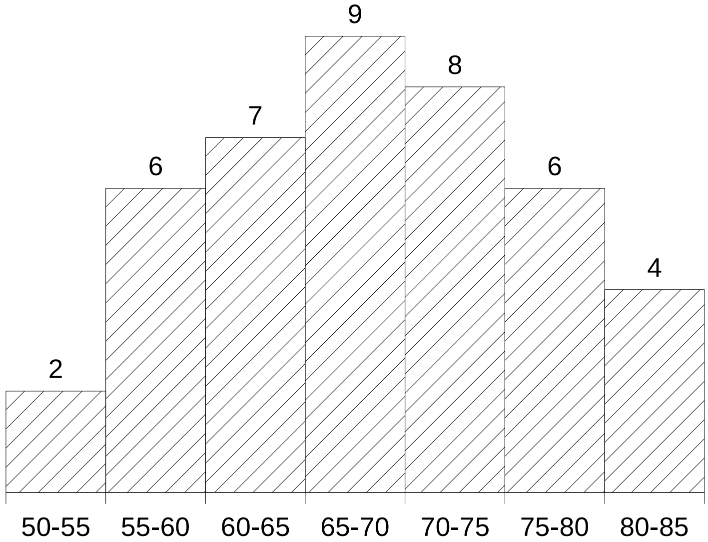
a) Ordena, razonadamente, sus medias y sus desviaciones típicas.

b) ¿Podemos afirmar que la que tiene mayor desviación típica es la más dispersa?

Problemas de distribuciones continuas

EJERCICIO 5: [RESUELTO]

Queremos estudiar el peso de los cuatrocientos alumnos de 2º de Bachillerato de un Centro. Para ello se ha realizado una medición cuyo resultado está aquí representado:



a) ¿De qué tipo de gráfico se trata? ¿Cuál es la variable estadística estudiada? ¿De qué tipo es? Indica cuál es la población y determina su tamaño. ¿Cuál es el tamaño muestral?

b) Escribe la tabla de frecuencias. ¿Qué porcentaje de alumnos pesa más de 70 kg?

c) Calcula el peso medio y la varianza.

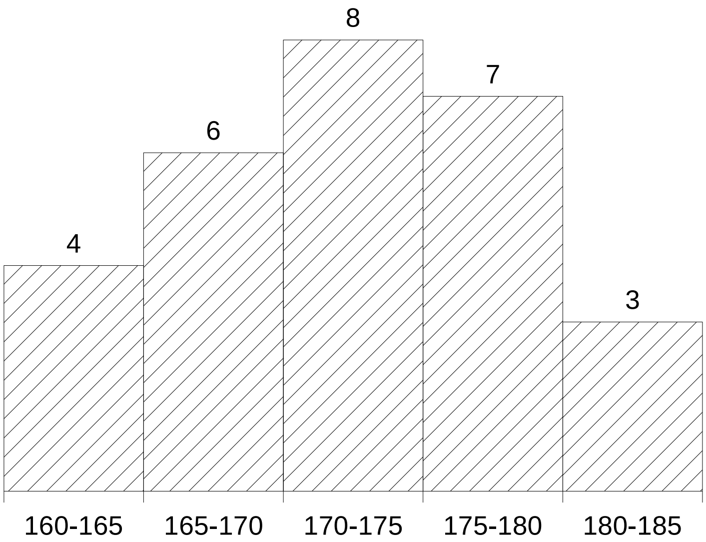
d) Obtén el peso mediano y la moda.

e) Qué porcentaje de alumnos tiene un peso inferior o igual a 71 kg?

f) En un colegio se ha obtenido un peso medio de 45.2 kg y una varianza de 68.5 kg². ¿En qué caso la distribución de los pesos presenta mayor dispersión?

EJERCICIO 6:

Queremos estudiar la estatura de los quinientos alumnos de 1º de Bachillerato de un Centro. Para ello se ha realizado una medición cuyo resultado está aquí representado:



a) ¿De qué tipo de gráfico se trata? ¿Cuál es la variable estadística estudiada? ¿De qué tipo es? Indica cuál es la población y determina su tamaño. ¿Cuál es el tamaño muestral?

b) Escribe la tabla de frecuencias. ¿Qué porcentaje de alumnos mide más de 170 cm?

c) Calcula la estatura media y la varianza.

d) Obtén la estatura mediana y la moda.

EJERCICIO 7:

La siguiente tabla muestra el número de respuestas correctas (X) en una prueba de 60 cuestiones realizadas a 456 personas:

X	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)
n_i	40	60	75	90	105	86

a) ¿Cuál es la variable estadística estudiada? Clasifíquela y escriba su tabla de frecuencias completa.

b) Represéntela gráficamente. ¿Cómo se llama el gráfico adecuado?

c) Calcule la media y la desviación típica.

d) Calcule la mediana y la varianza de la distribución.

e) Una persona ha cometido 35 errores. ¿Cuántos han cometido más que él?

f) Otro cuestionario arroja los siguientes resultados:

Y	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)
n_i	20	50	60	15	100	70

Decida cuál de los dos presenta una mayor dispersión.

Soluciones

EJERCICIO 1

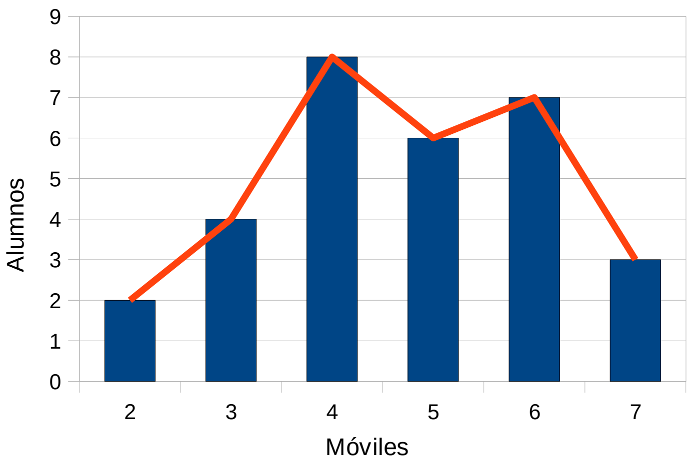
a) La variable estadística estudiada es el número de teléfonos móviles que hay en cada hogar. Es cuantitativa discreta y la población es el grupo de 1º de Bachillerato. Su tamaño es 25.

b) He aquí la tabla de frecuencias:

x_i	n_i	N_i	$f_i(\%)$	$f_i(\%)$
2	2	2	6,67	6,67
3	4	6	13,33	20
4	8	14	26,67	46,67
5	6	20	20	**66,67**
6	7	27	23,33	90
7	3	30	10	100

En la tabla vemos que $F(5) = 67,67\%$.

c) Diagrama de barras y polígono de frecuencias:



d) La moda es $\text{Mo} = 4$ móviles porque es el dato de mayor frecuencia.

$$\overline{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{141}{30} = 4.7 \text{ móviles}$$

e) Los cuartiles son:

$Q_1 = p_{25} = 4$ móviles Porque es el primer dato con $F > 25\%$.

$Q_2 = p_{50} = 5$ móviles Porque es el primer dato con $F > 50\%$ (es la mediana).

$Q_3 = p_{75} = 6$ móviles Porque es el primer dato con $F > 75\%$.

f) El recorrido o rango es $7 - 2 = 5$ móviles.

$$\sigma = \sqrt{\frac{\sum x_i^2 n_i}{\sum n_i} - \overline{x}^2} = \sqrt{\frac{721}{30} - 4.7^2} = 1.3940\ldots \approx 1.4 \text{ móviles}$$

EJERCICIO 2

Para comparar la dispersión de dos distribuciones estadísticas usamos el coeficiente de variación:

Pesos de los toros: $\mathrm{CV}_1 = \dfrac{\sigma_1}{\overline{x}_1} = \dfrac{50}{600} = 0.0833\ldots$

Pesos de las ovejas: $\mathrm{CV}_2 = \dfrac{\sigma_2}{\overline{x}_2} = \dfrac{10}{60} = 0.1666\ldots$

Esto nos indica que en el caso de los toros sus pesos presenta una desviación del 8% respecto de su peso medio, mientras que en los pesos de las ovejas sube casi al 17%. Concluimos que la distribución de los pesos de las ovejas es la más dispersa al ser mayor su coeficiente de variación.

EJERCICIO 5:

a) El gráfico se trata de un histograma de frecuencias.

La variable estadística estudiada es el peso de los alumnos. Es cuantitativa continua y la población son los alumnos de 1º de Bachillerato. El tamaño de la población es 500 y el tamaño de la muestra es 42

b) He aquí la tabla de frecuencias:

I	x_i	n_i	N_i	$f_i(\%)$	$f_i(\%)$
50-55	52,5	2	2	4,76	4,76
55-60	57,5	6	8	14,29	19,05
60-65	62,5	8	16	19,05	38,10
65-70	67,5	9	25	21,43	59,52
70-75	72,5	7	32	16,67	76,19
75-80	77,5	6	38	14,29	90,48
80-85	82,5	4	42	9,52	100

El porcentaje de alumnos que pesa más de 70 kg será 32 + 38 + 42 = 40,48%

c) $\overline{x} = \dfrac{\sum x_i n_i}{\sum n_i} = \dfrac{2860}{42} = 68.0952\ldots \approx 68.1 \text{ kg}$

$$\sigma^2 = \frac{\sum x_i^2 n_i}{\sum n_i} - \overline{x}^2 = \frac{197662,5}{42} - \left(\frac{2860}{42}\right)^2 = 69,2885\ldots \approx 69.3 \text{ kg}^2$$

d) La clase modal es 65 – 70 (kg), al ser la de mayor frecuencia. Diríamos que el peso modal es de 67,5 kg.

La clase mediana es 65 – 70 (kg), porque es la primera con $F > 50\%$.

$$\mathrm{Me} = p_{50} = 65 + (70-65)\cdot\frac{50-38.10}{59.52-38.10} = 67.7777\ldots \approx 67.8 \text{ kg}$$

e) Deseamos calcular el porcentaje k conociendo el valor del percentil, que es $p_k = 71$:

$$71 = 70 + (75-70)\cdot\frac{k-59.52}{76.19-59.52} \to 71 = 70 + 5\cdot\frac{k-59.52}{16.67} \to k = \frac{71-70}{5}\cdot 16.67 + 59.52$$

Obtenemos que es $k = 62.854\%$, con lo que es aproximadamente un 63%.

b) Para comparar la dispersión de dos distribuciones estadísticas usamos el coeficiente de variación:

	$\overline{x}$	σ^2	$CV = \dfrac{\sigma}{\overline{x}}$
Instituto	68.1	69.3	$\dfrac{\sqrt{69,3}}{68.1} = 0.12\dots$
Colegio	45.2	68.5	$\dfrac{\sqrt{68.5}}{45.2} = 0.18\dots$

Concluimos que la distribución de los pesos en el Colegio es más dispersa al ser mayor su coeficiente de variación.

EJERCICIO 6:

a) El gráfico se trata de un histograma de frecuencias.

La variable estadística estudiada es la estatura de los alumnos. Es cuantitativa continua y la población son los alumnos de 1º de Bachillerato. El tamaño de la población es 500 y el tamaño de la muestra es 28

b) He aquí la tabla de frecuencias:

I	x_i	n_i	N_i	$f_i(\%)$	$F_i(\%)$
160-165	162,5	4	4	14,29	14,29
165-170	167,5	6	10	21,43	35,71
170-175	172,5	8	18	28,57	64,29
175-180	177,5	7	25	25	89,29
180-185	182,5	3	28	10,71	100

El porcentaje de alumnos que mide más de 170 kg será 28,57+25+10,71 = 64,28%.

También podríamos calcularlo así: 100 – 35.71 = 64.29%. Obsérvese la diferencia de una centésima en el porcentaje provocado por el redondeo

c) Media: $\overline{x} = \dfrac{\sum x_i n_i}{\sum n_i} = \dfrac{4825}{28} = 172.2321\ldots \approx 172.2 \text{ cm}$

Varianza: $\sigma^2 = \dfrac{\sum x_i^2 n_i}{\sum n_i} - \overline{x}^2 = \dfrac{832.475}{28} - \left(\dfrac{24825}{28}\right)^2 = 36,5752\ldots \approx 36.6 \text{ cm}^2$

d) La clase modal es 170 – 175 (cm), pues es la de mayor frecuencia. Diríamos que la moda es $Mo = 172.5$ cm.

La clase mediana es 170 – 175 (cm), porque es la primera con $F > 50\%$.

$$\text{Me} = p_{50} = 170 + (175 - 170) \cdot \frac{50 - 35.71}{64.29 - 35.71} = 172.5 \text{ cm}$$