Tema 1. Estadística descriptiva de una variable

1.Variables estadísticas

1.1 Variables cualitativas

Denominamos **variable cualitativa** a aquella variable que sólo puede clasificarse en categorías no numéricas, como, por ejemplo, color de ojos, equipo de fútbol, etc.

1.2 Variables cuantitativas

Denominamos **variable cuantitativa** a aquella variable que toma valores numéricos. Se clasifican en dos tipos:

- **Discretas**: sólo pueden tomar un conjunto finito o numerable de valores. Por ejemplo, número de televisores en un domicilio.
- **Continuas**: pueden tomar cualquier valor dentro de un intervalo determinado. Por ejemplo, altura de una persona

1.3 Notación

La asignatura trabajará principalmente con variables cuantitativas y utilizará la siguiente notación:

n	Tamaño de la muestra, número de elementos observados o cuantificados
$x_1,...,x_n$	Observaciones de la muestra
$A_1,...,A_k$	Clases o conjuntos en los que podemos dividir la muestra
$n_1,...,n_k$	Número de observaciones en cada clase (frecuencias absolutas)
$f_1,...,f_n$	Frecuencias relativas en cada clase $f_i = n_i / n$

2. Diagramas de tallos y hojas

El diagrama de tallos y hojas es una de las múltiples formas de representar gráficamente las variables cuantitativas. El procedimiento es sencillo y lo seguiremos con el ejemplo del libro:

Representar mediante un diagrama de tallos y hojas los siguientes datos expresados en cm.;
11.357 12.542 11.384 12.431 14.212 15.213 13.300 11.300 17.206 12.710 13.455 16.143 12.162 12.721 13.420 14.698

a) Redondeamos los datos a un número conveniente de cifras significativas.
En nuestro caso redondearemos a tres cifras y trabajaremos en milímetros porque pensamos que será más cómodo para trabajar.

114 125 114 124 142 152 133 113 172 127 135 161
122 127 134 147

b) Colocamos en una tabla dos columnas separadas por una línea. Escribimos todas las cifras menos la última en la columna de la izquierda, ordenadas y sin repetir valores (tallo)

11
12
13
14
15
16
17

y a continuación escribimos la última cifra a la derecha, ordenados y pudiendo repetir valores (hojas)

11	344
12	24577
13	345
14	27
15	2
16	1
17	2

c) Observamos que la representación nos da una idea visual de la zona con mayor frecuencia de observaciones ya que el número de hojas representa la frecuencia de dicha clase.

3. Medidas de centralización

Las medidas de centralización nos dan una idea del valor central de una variable cuantitativa, es decir, el valor alrededor del cual se reparten los valores de la muestra que hemos obtenido.

3.1 Media muestral

Se representa por $\bar{x}$ y se define como

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3.2 Mediana

Es el valor de la muestra que deja a su izquierda y a su derecha el mismo número de observaciones, una vez ordenadas.

Por lo tanto, para calcular la mediana **siempre** hemos de empezar ordenando las observaciones. Entonces puede ocurrir dos cosas:

- si el número de observaciones es impar, la mediana es el valor central
- si el número de observaciones es par, la mediana es la media de los dos valores centrales

3.3 Moda

La moda de una muestra de una variables estadística discreta es el valor (o valores) que aparecen más veces repetidos en la muestra.

4. Medidas de dispersión

Las medidas de centralización, por sí solas, no nos dan mucha información de la muestra. Por ejemplo, supongamos dos parejas de ratones a los cuales medimos. La pareja A produce unas mediciones de 17 y 15 cm, mientras que la pareja B produce unas mediciones de 20 y 12 cm. Es evidente, que ambas parejas poseen la misma media muestral, por lo que podíamos pensar que son idénticas, pero resulta obvio que no son tan iguales y que seríamos capaces de diferenciarlas a simple vista.

Así pues, necesitamos de algún conjunto de medidas más, aparte de las de centralización, para obtener más datos de la muestra: las medidas de dispersión.

4.1 Varianza muestral

Se representa por v_x y se define como

$$v_x = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

aunque se suele utilizar otra fórmula de cálculos más sencillos

$$v_x = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

4.2 Desviación típica

La desviación típica (o desviación estándar) de la muestra es la raíz cuadrada positiva de la varianza muestral. Aunque existen diferentes notaciones el libro de texto no utiliza ninguna.

La idea intuitiva de la desviación típica es la media de desviaciones (diferencias) respecto de la media.

Por ejemplo, en nuestro caso de los ratones tenemos:

Pareja A $v_A = \frac{1}{2}\sum_{i=1}^{n}(x_i - 16)^2 = \frac{1}{2}\left[(17-16)^2 + (15-16)^2\right] = 1$ Desv. típica 1

o utilizando la fórmula alternativa

$$v_A = \frac{1}{2}\sum_{i=1}^{n} x_i^2 - 16^2 = \frac{1}{2}\left[17^2 + 15^2\right] - 16^2 = 1$$

Pareja B $v_B = \frac{1}{2}\sum_{i=1}^{n}(x_i - 16)^2 = \frac{1}{2}\left[(20-16)^2 + (12-16)^2\right] = 16$ Desv. típica 4

Así pues, tenemos que

Pareja A	$\bar{x} = 16$	Desv. típica	1
Pareja B	$\bar{x} = 16$	Desv. típica	4

Lo que nos indica que son fáciles de diferenciar, pues en la pareja A, los datos están desviados de la media un valor igual a 1 (por lo que son próximos a la media), y en cambio en la pareja B, los datos están desviados un valor de 4 respecto de la media, por lo que están más dispersos.