Estadística descriptiva undimensional.

Introducción

Estadística descriptiva.- Su objetivo es recoger, clasificar, representar gráficamente y resumir los datos de un conjunto así como calcular determinadas medidas que proporciones información sobre dicho conjunto.

Estadística Inferencial.- Su objetivo es llegar a conclusiones (inferencias) acerca de un determinado conjunto a partir de datos.

Algunos conceptos generales:

- Población: conjunto de individuos sobre los que se realiza un estudio.
- Individuo: cada elemento de la población.
- Muestra: subconjunto de la población cuyos datos son objeto de estudio.
- Tamaño muestral: número de individuos de la muestra.

Variable.- Se trata de una propiedad o cualidad referida a los elementos de la población. Se calsifican en

- Cuantitativas: variable que se pueden expresar numéricamente; a su vez pueden ser:
  - Discretas: pueden tomar únicamente valores numéricos aislados.
  - Continuas: pueden tomar valores numéricos en un intervalo de los números reales.
- Cualitativas: no se pueden expresar numéricamente; que pueden ser
  - Ordinales: se puede establecer un orden entre su valores.
  - Nominales: no admiten ordenación.

Frecuencia absoluta f_i es el número total de veces que aparece un determinado valor.

Frecuencia absoluta acumulada F_i es la suma de las frecuencias absolutas de los valores anteriores al de lugar i (caso de que se puedan ordenar).

Frecuencia relativa es el valor $h_i = f_i/n$ que resulta de dividir la frecuencia absoluta de un determinado valor entre el número total de datos n,

Frecuencia relativa acumulada H_i es la suma de todas las frecuencias relativas de los a valores anteriores al de lugar i (caso de que se puedan ordenar).

Porcentaje se trata del "tanto por ciento" atribuible a cada frecuencia.

Se verifican:

- $f_1 + \cdots + f_n = n$
- $h_1 + \cdots + h_n = 1$
- la suma de los porcentajes ha de ser 100.

Tabulación y representación gráfica

Para expresar los datos y frecuencias con mayor claridad se utilizan tablas y gráficos.

Tabla de frecuencias. se trata de una tabla donde aparecen los diferentes valores de la variable junto con las frecuencias, por ejemplo

Variable	frec. abs.	frec. abs. acum.	frec. rel.	frec. rel. acum	porcent.
x_1	f_1	F_1	h_1	H_1	$100h_1$
x_2	f_2	F_2	h_2	H_2	$100h_2$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

En el caso de variables cuantitativas continuas, los datos se agrupan en intervalos. Se llama recorrido de la muestra R a la diferencia entre el valor máximo x_{max} y mínimo x_{min} de la variable: $R = x_{max} - x_{min}$. La amplitud de cada intervalo $I_i = [\ell_i, \ell_{i+1}]$ es la diferencia $d_i = \ell_{i+1} - \ell_i$. Por último a cada intervalo se le asigna un valor medio que se llama marca de clase $x_i = d_i/2$.

En el caso de que no se proporcionen los intervalos, hay que decidir el número y la amplitud de los mismos. Una forma de decidir el número de intervalos la proporciona la Regla de Sturges:

$$\text{n}^{\text{o}} \text{ de intervalos } k = 1 + 3{,}322 \ln n$$

Para obtener la amplitud de cada intervalo se hace R/k.

Los datos se pueden expresar en diferentes tipos de gráficos:

- **Diagrama de barras.**
- **Gráfico de sectores.**
- **Polígonos de frecuencias**
- **Histogramas**

Medidas de posición

Se trata de medidas que indican sobre qué valores se sitúa la muestra.

Moda M_o

Si los datos no se agrupan en intervalos M_o es el dato, o datos, que más se repiten.

Si están agrupados en intervalos se llama intervalos modal al intervalo con mayor frecuencia:

- Igual amplitud $M_o = \ell_i + \dfrac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}}(\ell_{i+1} - \ell_i)$, con $(\ell_i, \ell_{i+1}]$ el intervalo modal.
- Distinta amplitud $M_o = \ell_i + \dfrac{k_i - k_{i-1}}{2k_i - k_{i-1} - k_{i+1}}(\ell_{i+1} - \ell_i)$, con $(\ell_i, \ell_{i+1}]$ el intervalo modal y k_i la altura correspondiente al rectángulo correspondiente al intervalo $(\ell_i, \ell_{i+1}]$ en el histograma.

Mediana M_e

Se trata del valor de la variable que deja a su izquierda un número de datos igual al de su derecha.

Si la variable es discreta, M_e es el valor central de la muestra si el número de datos es impar. En caso de que el número de datos sea par, M_e es la media de los dos valores centrales.

Si los datos están agrupados en intervalos, se llama intervalo mediano al primer intervalo cuya frecuencia absoluta acumulada F_i es igual o superior a $n/2$; entonces la mediana es

$$M_e = \ell_i + \frac{\frac{n}{2} - F_{i-1}}{f_i}(\ell_{i+1} - \ell_i);$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo mediano.

Percentiles

Se trata valor P_r de la variable que deja a su izquierda el porcentaje $r\,\%$ que la variable ($100-r\,\%$ a su derecha).

Cuando los datos están agrupados en intervalos

$$P_r = \ell_i + \frac{\frac{nr}{100} - F_{i-1}}{f_i}(\ell_{i+1} - \ell_i);$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo que contiene a P_r

Algunos percentiles frecuentes son los cuartiles $Q_1 = P_{25}$, $Q_2 = P_{50} = M_e$ y $Q_3 = P_{75}$. También se suelen considerar los deciles P_{10}, $P_{20}, \ldots$.

Media aritmética

Se calcula como

$$\overline{x} = \frac{f_1 x_1 + \cdots + f_n x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i;$$

si los datos están agrupados en intervalos x_i representa la marca de clase x_i.

Medidas de dispersión

Son medidas que proporcionan información de la separación de los valores de la variable entre sí o con respecto a ciertas medidas de posición.

Desviación mediana

Mide la separación de los datos con respecto a la mediana.

$$D\overline{x} = \frac{1}{n}\sum_{i=1}^{n} |x_i - M_e| f_i;$$

si los datos están agrupados en intervalos x_i representa la marca de clase x_i.

Desviación media

Mide la separación de los datos con respecto a la media.

$$DM_e = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}| f_i;$$

si los datos están agrupados en intervalos x_i representa la marca de clase x_i.

Varianza y desviación típica

La varianza es

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i = \frac{\sum_{i=1}^{n} {x_i}^2 f_i}{n} - \overline{x}^2$$

y la desviación típica es la raíz cuadrada de la varianza

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i} = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2 f_i}{n} - \overline{x}^2}.$$

La desviación típica también mide la separación de los datos con respecto a la media.

Medidas de forma

Se dice que una distribución es asimétrica por la derecha, o que presenta asimetría positiva si la mayoría de los datos están agrupados a la derecha (el polígono de frecuencias tiene su máximo por la derecha) y de forma similar, asimétrica por la izquierda.

Se puede, entre otros, utilizar el coeficiente de asimetría

$$\alpha = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3 f_i}{n} : {s_x}^3$$

Si $\alpha > 0$ está sesgada a la derecha, si $\alpha < 0$ a la izquierda y si $\alpha = 0$ sería simétrica.