Estadística: Resumen teórico

Parámetros estadísticos de centralización: moda, mediana y media.

La ***moda*** de una distribución estadística es el valor de la variable que se ha observado con mayor frecuencia (el valor de la variable que más se repite).

La ***mediana*** de una distribución estadística es el valor de la variable que ocupa el lugar central de la distribución, una vez que hemos ordenado los datos. Este parámetro estadístico sólo tiene sentido en el caso de variables no numéricas ordinales y, sobre todo, en el de variables numéricas. (Si la distribución tiene un número par de datos, la mediana es la media aritmética de los dos valores centrales).

La ***media*** de un conjunto de observaciones de una variable estadística numérica se obtiene sumando todos los valores y dividiéndolos por el número total de observaciones.

$$\bar{x} = \frac{\sum x_i}{n}.$$ Si los datos están agrupados por frecuencias, $$\bar{x} = \frac{\sum_1^n x_i f_i}{\sum_1^n f_i}$$

Parámetros estadísticos de dispersión: rango, cuartilas, desviación media y desviación típica

El ***rango*** de una distribución es la diferencia entre sus valores máximo y mínimo.

Las ***cuartilas*** de una distribución son aquellos valores que, una vez ordenados todos los datos, los dividen en cuatro conjuntos que tienen el mismo número de datos. Llamaremos ***rango intercuartílico*** a la diferencia existente entre la tercera y la primera cuartila.

La ***desviación media*** de una distribución se obtiene mediante la fórmula:

$$Desviación\ media = \frac{\sum f_i \left| x_i - \bar{x} \right|}{n}$$

Observa que estamos sumando las "distancias" (por eso el valor absoluto) de cada uno de los datos a la media y dividiendo este resultado por el número total de datos.

La ***desviación estándar*** o ***desviación típica*** de una distribución es la raíz cuadrada positiva de la varianza y se representa por la letra griega σ.

La fórmula que nos da su valor es $\sigma = \sqrt{\frac{\sum f_i \left(xi - \bar{x}\right)^2}{n}}$. Esta fórmula suele ser bastante laboriosa de calcular, con lo que podemos dar otra expresión de la desviación típica más sencilla $\sigma = \sqrt{\frac{\sum f_i xi^2}{n} - \bar{x}^2}$

Para calcular todos estos parámetros y poder luego repasar nuestras cuentas de forma ordenada, es recomendable construirse una tabla en la que aparezcan los siguientes datos:

x_i	f_i	$f_i\left(x_i - \bar{x}\right)$	$f_i\left(x_i - \bar{x}\right)^2$

Relaciones entre variables estadísticas.

Una ***distribución estadística bidimensional*** es el resultado de recoger, para cada individuo de una población, dos variables. (estatura y peso, horas de estudio y nota conseguida, …). A cada una de estas variables de las suele llamar x e y

El ***punto medio*** de una distribución de este tipo es el dado por $\left(\bar{x}, \bar{y}\right)$. Llamaremos ***momento producto*** al valor $\sigma_{xy} = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$, que también puede calcularse más fácilmente mediante la fórmula $\sigma_{xy} = \frac{\sum x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y}$

Cuando se estudian conjuntamente dos variables, lo esencial suele ser ***buscar relaciones*** entre ellas que nos permitan hacer predicciones sobre los valores de una de ellas conocidos valores de la otra.

Cuando trabajamos con variables numéricas, el coeficiente de correlación más empleado es el de Pearson, r, que mide con valores positivos la tendencia a una asociación positiva (cuando una variable crece, la otra también) y con valores negativos

la tendencia negativa (cuando una variable disminuye, la otra también). El coeficiente de correlación de Pearson está comprendido entre los valores -1 y 1.

$$-1 \le r \le 1$$

y se calcula mediante la siguiente fórmula: $r = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

Como ya he mencionado en párrafos anteriores, el objetivo final más frecuente de los estudios científicos experimentales es la predicción de resultados. Hay muchos tipos de relaciones entre las variables (lineal, parabólica, logarítmica, exponencial, potencial…), pero la que contemplamos aquí es la lineal. Si el coeficiente de correlación lineal r es lo suficientemente cercano a 1 o a -1 como para decir que hay relación entre las variables, entonces podremos calcular la llamada ***recta de regresión de y sobre x*** (conocido un determinado valor de la variable x, puedo conocer la predicción para la variable y) con la siguiente fórmula:

La recta es $y = ax + b$, siendo $a = \dfrac{\sigma_{xy}}{\sigma_x^2}$ y $b = \overline{y} - a\overline{x}$

Cuando se trata de trabajar con variables bidimensionales, es bastante práctico hacerse una tabla en la que figuren los siguientes datos:

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$

que nos facilita luego en gran medida tanto el cálculo de los diferentes parámetros estadísticos, como el repaso de las mismas.