Medidas resumen con datos agrupados

Información agrupada

Cuando se tiene un gran número de observaciones, la distribución de frecuencias condensa la información contenida en los datos de la muestra.

El hecho de agrupar datos hace que se pierda alguna información de la muestra original. Los cálculos que se hacen a partir de datos agrupados son matemáticamente convenientes, pero son menos precisos que los que se obtienen con todo el conjunto de datos.

Muchos datos originales se presentan en forma de tablas de frecuencias, por ejemplo, cuando se recopila información en una encuesta que pregunta por la edad o el nivel de ingreso, datos que de antemano se encuentran enmarcados en un intervalo específico.

La media aritmética de una muestra agrupada

Fórmulas de uso

Si la distribución de frecuencia tiene k intervalos de clase con puntos medios (marcas de clase) m_1, m_2, ..., m_k, se asume que cada dato en el i-esimo intervalo es igual a la marca de clase con sus correspondientes frecuencias absolutas f_1, f_2, ..., f_k, entonces:

Formulas a utilizar

$$Media\ Agrupada = \bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

$$Media\ Agrupada = \bar{x} = \frac{\sum_{celdas}(Punto\ medio \text{ x } Frecuencia)}{Frecuencia\ total}$$

Varianza de datos agrupados

Fórmulas de uso

Se asume que todas las observaciones en una clase se localizan en el punto medio del intervalo, es decir, su marca de clase **m_i**

Las desviaciones cuadráticas respecto a la media son: $(m_1 - \bar{x})^2$,

$(m_2 - \bar{x})^2, \ \cdots, (m_k - \bar{x})^2$ que se repiten con frecuencias f_1, f_2, ..., f_k, respectivamente.

La suma de las desviaciones cuadráticas es:

$$\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i$$

La varianza muestral se obtiene dividiendo por n - 1. En datos agrupados n es por lo general grande y dividir por n - 1 es virtualmente equivalente a dividir por n.

$Varianza\ de\ la\ muestra\ agrupada$

$$s^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{k} (m_i)^2 f_i}{n} - \bar{x}^2$$

Ejemplo: Un estudio con 200 mujeres menores de 40 años, respecto a la edad en que contraen matrimonio, por primera vez, reportó la siguiente distribución de frecuencias.

Intervalo de clase, años de edad			**Marca de clase**	**Frecuencia**
LI		**LS**	**mi**	**fi**
14,5	–	19,5	17	18
19,5	–	24,5	22	74
24,5	–	29,5	27	62
29,5	–	34,5	32	26
34,5	–	39,5	37	20
Total				200

Figura 1. *Primer matrimonio en mujeres < 40 años (1).*

Calcula la media y la desviación estándar de esta muestra agrupada.

Solución

Intervalo de clase, años de edad			Marca de clase	Frecuencia		
LI		LS	m_i	f_i	$m_i f_i$	$m_i^2 f_i$
14,5	–	19,5	17	18	306	5202
19,5	–	24,5	22	74	1628	35816
24,5	–	29,5	27	62	1674	45198
29,5	–	34,5	32	26	832	26624
34,5	–	39,5	37	20	740	27380
Total				**200**	**5.180**	**140.220**
			Media =	**25,9**		
				Varianza =	**30,29**	
					Desviación estándar =	**5,5**

Figura 1. *Primer matrimonio en mujeres < 40 años (2).*

$$Media\ agrupada =$$

$$\bar{x} = \frac{17\text{ x}18 + 22\text{ x}74 + 27\text{ x}62 + 32\text{ x}26 + 37\text{ x}20}{200}$$

$$Media\ agrupada = \bar{x} = 25{,}90$$

Los datos necesarios para el cálculo de la varianza se dan en la misma tabla de frecuencia.

$$s^2 = \frac{\sum m_i^2 f_i}{n} - \bar{x}^2 = 30{,}29$$

$$s = \sqrt{s^2} = \sqrt{30{,}29} = 5{,}50\ años.$$

Pensar en la forma de interpretar estos resultados.

Referencias bibliográficas

Díaz Cadavid, A. y Gutiérrez Arias, A. (1991). *Estadística General.* Universidad de Antioquia.

Wayne, W. D. (1991). *Bioestadística: base para el análisis de las ciencias de la salud.* Limusa S.A.