Análisis de datos en Psicología

Grado en Psicología UMU

Apuntes

1. APROXIMACIÓN AL ANÁLISIS ESTADÍSTICO DE DATOS

1.1 CONCEPTOS GENERALES

Población: conjunto de todos los elementos u objetos que comparten una o varias características.

Muestra: subconjunto de la población de estudio.

Individuo: cada uno de los elementos que componen una población o muestra.

Parámetro: resume una determinada información referente a la población.

Estadístico: resume una determinada información referente a una muestra.

La **estadística descriptiva** se encarga de resumir y describir un conjunto de datos para su comprensión.

La **inferencia estadística** se encarga de extraer conclusiones sobre una población a partir del estudio de una muestra mediante técnicas probabilísticas.

Característica: propiedad de los individuos de una población o una muestra → variable.

Modalidad: cada una de las variantes con las que se manifiesta una característica → valores.

1.2 MEDICIÓN Y ESCALAS DE MEDIDA

Medición: proceso en el que se atribuyen números a las modalidades de una característica. La asignación obedece unas reglas. Se hace con la condición de que las relaciones que establecemos entre los números reflejen las relaciones empíricamente comprobables entre los sujetos. Stevens en 1946 estableció las diferentes **escalas de medida** en función de las relaciones entre los números:

Nominal:

Ordinal:

Intervalo:

Razón:

1.3 VARIABLES: CLASIFICACIÓN Y NOTACIÓN

Variable: característica observable que cambia entre los elementos de una población. Característica que puede ser medida o contada y que puede adoptar más de un valor.

Variables nominales: son aquellas variables que tienen un conjunto de categorías que no tienen ninguna tipo de jerarquía.

- **Dicotómicas:** tienen dos categorías.
- **Politómica:** tienen más de dos categorías.

Variables ordinales o casicuantitativas: son aquellas variables donde el conjunto de categorías sí tienen una jerarquía u orden.

Variables cuantitativas o numéricas: variables que recogen como información una cantidad numérica de lo que se está observando. *Ej.: Edad, peso, tensión arterial, número de hijos, hermanos…*

- **Discretas:** tienen un conjunto finito de valores y en caso infinito solamente toman los números enteros
- **Continuas:** el conjunto de posibles valores entre dos números fijos es infinito.

Variable dicotomizada y politomizada: que se categoriza de forma artificial en 2 o más valores.

No es posible conocer el valor exacto de una **variable continua**, así que tendremos que trabajar con su valor informado o aparente (aproximado).

NOTACIÓN DE VARIABLES Y SÍMBOLO SUMATORIO.

$$Variables \rightarrow (x_i, y_i, z_i \dots) \quad Constantes \rightarrow (a, b, c \ \dots)$$

$$Sumatorio: \sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \qquad \rightarrow \sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

$$Reglas\ del\ sumatorio \begin{cases} 1a\ propiedad: \sum k \cdot x_i = k \cdot \sum x_i \\ 1a\ propiedad: \sum k = n \cdot k \\ 3a\ propiedad: \sum (x_i + y_i + z_i) = \sum x_i + \sum y_i + \sum z_i \end{cases}$$

$$Consecuencias \quad \rightarrow \quad \begin{array}{l} \sum (x_i + k) = \sum x_i + n \cdot k \\ \sum (x + k)^2 = \sum x_i^2 + 2k \sum x_i + n \cdot k^2 \\ \sum (x_i + y_i)^2 = \sum x_i^2 + 2 \sum x_i \cdot y_i + \sum y_i^2 \end{array}$$

2. ORGANIZACIÓN Y REPRESENTACIÓN DE DATOS

DISTRIBUCIÓN DE FRECUENCIAS

Frecuencias absolutas (n_i): es el número de veces que aparece cada observación.

Frecuencias absolutas acumuladas (n_a): suma de las frec. absolutas menores o iguales al dato x_i.

Frecuencia relativa (p_i): cociente de la frecuencia absoluta entre el total de observaciones $p_i = \frac{n_i}{n}$

Frecuencia relativa acumulada (p_a): suma de las frecuencias relativas menores o iguales al dato x_i

Porcentaje (P_i): frecuencia relativa multiplicada por cien $P_i = 100 \cdot p_i$

Porcentaje acumulado (P_a): suma de los porcentajes menores o iguales al dato x_i.

Ejemplo: clase a la que va cada alumno B, B, A, C, B, B, B, C, B, A.

x_i	n_i	n_a	p_i	p_a	P_i	P_a
A						
B						
C						

Ejemplo: notas de un examen 3,4; 5,3; 6,8; 9,1; 7,2; 8,3; 1,9; 5,1; 4,4; 3,2.

x_i	n_i	n_a	p_i	p_a	P_i	P_a
[0-2)						
[2-4)						
[4-6)						
[6-8)						
[8-10)						

En una variable nominal no tiene sentido acumular. La última categoría siempre tiene $n_a = n \;\; p_a = 1 \; y \; P_a = 100$

Cuando la variable es cuantitativa y tiene numerosos valores, debemos agrupar por intervalos. Al medir una variable cuantitativa continua, el valor que se lee en el instrumento no es exacto, es aparente o virtual:

$$límites\ exactos = valor\ informado \pm 0{,}5 \cdot precisión\ del\ instrumento$$

Podemos distinguir entre límites informados y exactos de los intervalos en una distribución de frecuencias.

Límites informados aparentes o virtuales: son los valores mayor y menor, teniendo en cuenta el nivel de precisión del instrumento de medida.

Límites reales o exactos: son los valores mayor y menor, si el instrumento tuviera una precisión perfecta.

Amplitud del intervalo: diferencia entre los límites exactos superior e inferior

DISTRIBUCIÓN DE VALORES DENTRO DE UN INTERVALO

Cuando agrupamos los valores de una variable en intervalos, solo informamos del número de sujetos en cada intervalo, perdiendo así el valor concreto de cada uno. Para poder analizar los datos agrupados en intervalos, se suelen formular dos supuestos acerca de la distribución de los valores en el intervalo.

El **primer supuesto** establece que la distribución de valores en un intervalo es homogénea.
El **segundo supuesto** es que la concentración de los valores se da en el punto medio del intervalo.

En la práctica, aplicaremos uno u otro en función de la técnica estadística aplicada.

Error de agrupamiento: error cometido al aceptar cualquiera de los dos supuestos.

REPRESENTACIONES GRÁFICAS

Para variables cualitativas:

Diagrama de sectores

Diagrama de barras

Para variables cuantitativas:

Diagrama de barras

Histograma/Polígono de frecuencias

PROPIEDADES DE LA DISTRIBUCIÓN

La forma de una distribución de frecuencia se caracteriza por tres propiedades:

- Propiedad 1: Tendencia central. El valor de la variable que denominamos promedio y que se sitúa en el centro de la distribución. Se trata de un valor que resume, sintetiza y representa a todos los valores.

- Propiedad 2: Variabilidad. Medida del grado de concentración de los valores de una variable en torno a su promedio. Si los valores están muy cerca del promedio, es una distribución homogénea. Si los valores están muy lejos del promedio, es una distribución heterogénea.

- Propiedad 3. Asimetría. Una distribución es simétrica cuando la forma de la distribución a la izquierda del promedio coincide con la forma de la distribución a la derecha. Distribución asimétrica positiva predomina en los valores bajos. Distribución asimétrica negativa predomina en los valores altos.

3. ÍNDICES DE TENDENCIA CENTRAL Y DE POSICIÓN

ÍNDICES DE TENDENCIA CENTRAL

Nos centraremos en tres índices o medidas que permiten cuantificar la tendencia central.

Moda: es el valor de la variable con mayor frecuencia. Si tenemos intervalos, se obtiene del punto medio del intervalo con mayor frecuencia.

Propiedades de la moda:

1. Cuando los datos están agrupados por intervalos, la moda puede cambiar en función del número de intervalos elegidos y su amplitud.

2. Cuando los datos están agrupados por intervalos y alguno de los intervalos extremos es abierto, la moda se puede calcular siempre y cuando la frecuencia máxima no esté en ninguno de los intervalos abiertos.

Mediana: es el valor numérico o puntuación que deja por debajo y por encima el 50% de las observaciones. Se calcula de forma diferente si los datos están o no agrupados:

$$Pos_{Md} = \frac{n+1}{2} \qquad Me = L_{i(e)} + \frac{\frac{n}{2} - n_d}{n_c} \cdot I_i$$

Propiedades de la mediana:

1. $\sum|x_i - Md| < \sum|x_i - c| \quad si\ c \neq Md$

2. La mediana apenas se deja afectar por las puntuaciones extremas.

3. Se puede calcular la mediana cuando algunos de los intervalos extremos son abiertos y la mediana no está en ellos.

4. Gráficamente dividirá el área total del histograma en dos partes iguales.

Media aritmética: es la suma de todas las puntuaciones divididas entre el total de casos, por lo que será normalmente diferente a la moda y a la mediana. Toma en consideración todas las puntuaciones y tenemos varias fórmulas en función de si los datos los tenemos agrupados o sin agrupar:

$$\bar{X} = \frac{\sum X_i}{n} = \qquad \frac{\sum n_i \cdot X_i}{n} = \sum p_i \cdot X_i$$

Propiedades de la media:

1. $\sum(X_i - \bar{X}) = 0$
2. $\sum(X_i - \bar{X})^2 < \sum(X_i - c)^2 \ \ si\ c \neq \bar{X}$
3. $si\ Y_i = X_i + a \quad \rightarrow \quad \bar{Y} = \bar{X} + a$
4. $si\ Y_i = b \cdot X_i \quad \rightarrow \quad \bar{Y} = b \cdot \bar{X}$
5. $si\ Y_i = a + b \cdot X_i \quad \rightarrow \quad \bar{Y} = a + b \cdot \bar{X}$
6. $\bar{X}_T = \frac{n_1 \cdot \bar{X}_1 + n_2 \cdot \bar{X}_2 + \cdots}{n_1 + n_2 \ldots}$
7. *La media no se puede calcular si los datos están en intervalos y alguno es abierto*

Criterios para la elección de un índice de tendencia central:

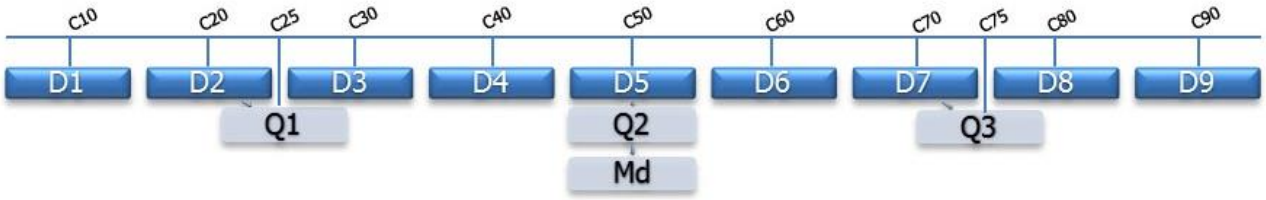
- Con variables cualitativas o medidas en escala nominal: la Mo.
- Con variables medias en escala ordinal o cuasi-cuantitativa: la moda o la Md, preferiblemente la Md.
- Con variables cuantitativas medidas en escala de intervalo o de razón: la Mo, la Md o $\bar{X}$, preferiblemente $\bar{X}$

INDICADORES DE POSICIÓN O CUANTILES

Percentiles: que deja debajo de sí un porcentaje acumulado de sujetos de la distribución.

Deciles: los 9 valores que dividen la distribución en 10 partes iguales.

Cuartiles: los 3 valores que dividen la distribución en 4 partes iguales.



$$P_k = L_{i(e)} + \frac{\frac{k}{100} \cdot n - n_d}{n_c} \cdot I_i$$

Ejemplo: calcula los indicadores anteriores con los siguientes datos 2, 1, 1, 0, 1, 2, 2, 0, 1, 1

Ejemplo: calcula los indicadores anteriores con los siguientes datos:

x_i	n_i	n_a	p_i	p_a
[8-10)	*2*			
[6-8)	*2*			
[4-6)	*3*			
[2-4)	*2*			
[0-2)	*1*			

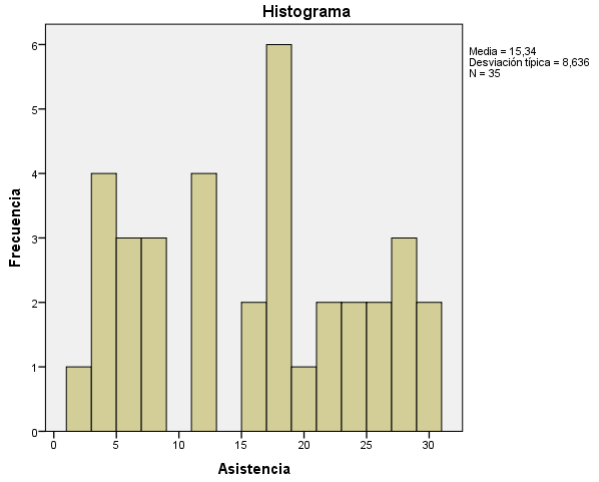
Ejemplos SPSS y Jamovi:

Descriptivos

			Estadístico	Error típ.
Asistencia	Media		15,34	1,460
	Intervalo de confianza para la media al 95%	Límite inferior	12,38	
		Límite superior	18,31	
	Media recortada al 5%		15,27	
	Mediana		17,00	
	Varianza		74,585	
	Desv. típ.		8,636	
	Mínimo		2	
	Máximo		30	
	Rango		28	
	Amplitud intercuartil		16	
	Asimetría		,044	,398
	Curtosis		-1,221	,778

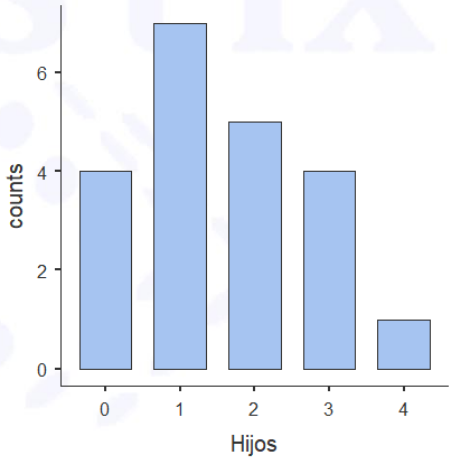
Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado (definición 1)	Asistencia	2,80	3,00	7,00	17,00	23,00	27,40	29,20



Descriptives

	Hijos
N	21
Missing	1
Mean	1.57
Median	1
Mode	1.00
Minimum	0
Maximum	4
25th percentile	1.00
50th percentile	1.00
75th percentile	2.00



4. ÍNDICES DE VARIABILIDAD Y SESGO O ASIMETRÍA

CONCEPTO DE VARIABILIDAD

Variabilidad: hace referencia al grado de concentración de las puntuaciones de una variable con respecto al promedio, es decir, refleja el grado de diferencias individuales de tal forma que si hay muchas, habrá más dispersión o variabilidad. Se puede expresar de varias maneras:

$$homogénea \leftrightarrow heterogénea \quad concentrada \leftrightarrow dispersa \quad poca\ variabilidad \leftrightarrow mucha\ variabilidad$$

AMPLITUD TOTAL, SEMI-INTERCUARTIL, VARIANZA Y DESVIACIÓN TÍPICA

Ahora veremos una serie de índices que nos indican la variabilidad que tiene una determinada variable:

Amplitud total: es la diferencia entre el máximo y el mínimo.

$$A_t = X_{máx} - X_{mín}$$

Amplitud semi-intercuartil: es la amplitud promedio entre Q1 y Q2 y entre Q2 y Q3.

$$Q = \frac{Q_3 - Q_1}{2}$$

Varianza: es la media aritmética de los cuadrados de las diferencias de los valores de la variable con respecto a su media aritmética y coincide con el momento de segundo orden respecto a la media.

$$S_x{}^2 = \frac{\sum(X_i - \bar{X})^2}{n} = \frac{\sum(X_i - \bar{X})^2 \cdot n_i}{n} \quad = \frac{\sum X_i{}^2}{n} - \bar{X}^2 = \frac{\sum X_i{}^2 \cdot n_i}{n} - \bar{X}^2$$

La desviación estándar/típica: es la raíz de la varianza.

$$S_x = \sqrt{S_x{}^2}$$

Consideraciones sobre la varianza y la desviación típica:

1. La varianza viene expresada en unidades cuadráticas.

2. Tanto la varianza como la desviación típica solo pueden tomar valores positivos, siendo 0 el valor mínimo.

Propiedades de la varianza y la desviación típica:

1. Si $y_i = x_i + a$ *entonces* $S_y^2 = S_x^2$ *y* $S_y = S_x$*. Es decir, no cambian.*

2. Si $y_i = b \cdot x_i$ *entonces* $S_y^2 = b^2 \cdot S_x^2$ *y* $S_y = |b| \cdot S_x$

3. Como se calculan a partir de la media, presentan todos los inconvenientes de ella. Es decir, les afecta las puntuaciones extremas y no se pueden calcular si los intervalos extremos son abiertos.

ÍNDICES DE ASIMETRÍA

Explicación de clase:

Índices de asimetría: viendo el gráfico de frecuencias, podemos saber si la distribución es simétrica o no. Además, podemos cuantificar el grado de asimetría con los diferentes índices de asimetría.

$$As_{Intercuartil} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

$$As_{Pearson} = \frac{\bar{X} - Mo}{S_x}$$

$$As_{Fisher} = \frac{\sum(X_i - \bar{X})^3}{n \cdot {S_x}^3} = \frac{\sum(X_i - \bar{X})^3 \cdot n_i}{n \cdot {S_x}^3}$$

As>0 asimetría positiva

As=0 simetría

As<0 asimetría negativa

Ejemplo: calcula los indicadores anteriores con los siguientes datos 2, 1, 1, 0, 1, 2, 2, 0, 1, 1

Ejemplo: calcula los indicadores anteriores con los siguientes datos:

x_i	n_i	n_a	p_i	p_a
[8-10)	*2*			
[6-8)	*2*			
[4-6)	*3*			
[2-4)	*2*			
[0-2)	*1*			

Ejemplos SPSS y Jamovi:

Descriptivos

		Estadístico	Error estándar
Media		5,514	,3652
95% de intervalo de confianza para la media	Límite inferior	4,772	
	Límite superior	6,257	
Media recortada al 5%		5,516	
Mediana		5,000	
Varianza		4,669	
Desviación estándar		2,1608	
Mínimo		1,0	
Máximo		10,0	
Rango		9,0	
Rango intercuartil		3,0	
Asimetría		,137	,398
Curtosis		-,595	,778

Descriptives

	Notas_final
N	35
Mean	5.54
Median	5
Mode	4.00
Variance	6.26
Range	8
Minimum	2
Maximum	10
Skewness	0.233
Std. error skewness	0.398
Kurtosis	−1.04
Std. error kurtosis	0.778
25th percentile	4.00
50th percentile	5.00
75th percentile	7.00

Este dosier está hecho para seguir la clase de prueba.

Si te apuntas al curso te enviaremos el dosier entero con todos los temas que faltan, ejercicios y exámenes de años anteriores

Más información en:

www.estadistix.com

Y si tienes cualquier consulta,
escríbenos un whatsapp al 644310902