Conceptos Basicos de Estadistica

makita

10/14/2021

Conceptos básicos de estadística

Población (universo o colectivo)

Es el conjunto total de ELEMENTOS de la misma naturaleza cualquera que sea, que son de interés para un problema dado

- N = Representación de el tamaño de la población

muestra

Variable aleatoria:

Son fenómenos o características de los elementos de la población.

Función de valor real que tiene como dominio el espacio muestral de un experimento aleatorio.

Variables sobre las cuales tenemos un grado de icertidumbre respecto a los valores que puede tomar

Datos

Son los resultados observados de las variables aleatorias (Cuando se hace una medición)

Parámetro

Es la medición global de cualesquer característica de los elementos de la población.

Es un valor teórico asociado a la población.

Ejemplos

Población: Los niños y niñas de 0 a 5 años de edad localozados en Bogotá

Variables: género, edad, peso, talla, estrato, localidad, fecha, lugar de nacimiento..

Parámetros: IMC

Clasificacion de variables

Cualitativas (categóricas)

Cuantitativas

Los valores de las observacione so niméricas y en conseciencia, ordenables.

Discreta

Recorridos finitos numerables sin tomar valores intermedios e.g. conteos.

Continua

Recorridos infinitos no numerables e.g. la distribución normal

Escalas de medición

Cualitativas

Nominales: Clasificación de objetos o fenómenos mediante símbolos o signos (No hay orden o dirección). e.g.

- Nombre
- Número de la cédula
- Tipo de sangre
- Color de los ojos
- Número de camiseta de los jugadores

Los números en la lista anterior no pueden ser sometidos a operaciones matemáticas

Ordinales

Categorías ordenadas (Rangos, órdenes, escalamientos)

- Sabor de un yogurt

Cuantitativas

Intervalo

Los datos medidos en una escala orrdinal para los cuales pueden clasificarse las distancias entre valores pero no existe un cero absoluto o no exista ausencia total de la característica

- Temperatura: a 0°C no deja de existir la temperatura
- Notas: se corre la escala e inicia desde 3.

Razón

Tiene todas las características de un intervalo, y además tiene un cero absoluto

Resumen y descripción de datos de una variable

Datos en bruto en forma de listas (o bases no son fáciles de usar para tomar decisiones)

- Se necesita algún tipo de organización

Para esto podemos utilizar gráficos de barras, graficos de torta, o tablas de frecuencias.

Como agrupar los datos: Sturgues

Si n no es demasiado grande, intervalos = $\sqrt{n}$

En caso contrario:

$$k = 1 + 3.322log(n) \tag{1}$$

k = intervalos de clase

Para la longitud de los intervalos:

$$L = \frac{Dato\ mayor - Dato\ menor}{n} \tag{2}$$

- A menudo es prueba y error

Tipos de frecuencias

- Absoluta: Conteo de observaciones que cae en cada intervalo.
- Relativa: $\frac{Absoluta}{n}$.
- Acumulada: Suma de las frecuencias absolutas
- Relativa acumulada: Suma de las frecuencias relativas.

Caracteristicas a revisar de las distribuciones

- Distribucion
- Localizacion (sesgo)
- Dispersion

Medidas de localización

Media aritmética:

Si $x_1, x_2, x_3, ...x_n$ es una muestra de una poblacion de tamaño N entonces la media es N

Media poblacional

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \tag{3}$$

Estimador muestral

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Caracteristicas:

- Es facil de obtener
- Medida no robusta: Afectada por valores extremos o datos atípicos.

Propiedades de la media aritmetica:

Si $x_1, x_2, x_3, ...x_n$ es una muestra de una poblacion de tamaño N entonces la media es N, entonces

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n \tag{5}$$

Si $x_i = c$ y a su vez c es constante, entonces

$$\sum_{i=1}^{n} x_i = \sum_{i=1} c = c + c + c + \cdots \tag{6}$$

Entonces

$$\sum_{i=1}^{n} x_i = nc \tag{7}$$

- *Ejemplo:*

Si c es una constante que multiplica las observaciones:

$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$

Si$x_1, x_2, x_3, ...x_n$ y $y_1, y_2, y_3, ...y_n$ son sucesiones de numeros;

$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

Si$x_1, x_2, x_3, ...x_n$ y $y_1, y_2, y_3, ...y_n$ son sucesiones de numeros;

$\sum_{i=1}^{n}(x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i$

5.

$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

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

6.

promedio de y en funcion de promedio de x en regresion lineal simple

Si A

La mediana

Es el valor central (es el dato de la variable que esta en el centro de la misma). Deja por encima y por debajo mitad y mitad de las observaciones.

Calculo de la mediana

Depende si el conjunto es par o impar:

Si$x_1, x_2, x_3, ...x_n$ Son los valores ordenados en una muestra de una poblacion de tamaño N:

$\hat{x} = \frac{x_{n/2} + x_{n+1/2}}{2}$ si n es par

$\hat{x} = x_{n=1/2}$ si n es impar

Es un estimador robusto, no se ve afectado por valores extremos

Ejemplo

Edad de ninos

```
x1 <- c(6,7,8,9,10)
```

n es impar, entonces $\hat{x} = x_{n+1/2} = x_{6/2} = x_3 = 8$

De la muestra analizada la mitad de los ninos tienen entre 6 y 8 años, y la otra mitad entre 8 y 10 años.

Moda

- El valor que más se repite
- Usada para valores numéricos o categóricos

e.g. Cual es el color más frecuente en los ojos.

Medidas de dispersión o variación

Varianza

- Varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{8}$$

- Varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{9}$$

$$s^2 = \sum_{i=1}^{n} \frac{(X_i^2 - 2x_i\bar{x} + \bar{x}^2}{n-1} \tag{10}$$

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2}{n-1} \tag{11}$$

Coeficiente de variación

$CV = \frac{s}{\bar{x}} \cdot 100\%$

Si CV es igual o menor a 5% hay homogeneidad

si esta entre 5% y 20% los datos son medianamente homogeneos

Si CV mayor a 20% hay heterogeneidad ## Rango

medida no robusta, si hay datos atipicos se ve muy afectado

rango intercuartilico

boxplot

Cuartiles

Se divide en cuatro partes porcentiales el conjunto de observaciones

Se calcula de la sigueinte manera

$Q_k = k \cdot \frac{n}{4}\ k = 1, 2, 3$

Deciles

Se divide en diez partes porcertialmente iguales

$D_k = \frac{n}{10}\ 1, 2, 3, ..., 9$

Percentiles

es mas detallado, nos da mas acceso a distintos puntos de la distribucion

$P_k = \frac{n}{100}\ 1, 2, 3, ..., 99$

Coeficientes de asimetría de Fisher

Permite interpretar la forma de la distribución, respecto a ser o no asimétrica

Coeficiente de curtosis

Mide el grado de aplastamiento o apuntamiento de la gráfica de la distribución.

Otras medidas de centralización

Desviacion media absoluta

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{12}$$