Conceptos Basicos de Estadistica

Felipe

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:

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

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:*

$$\sum_{i=1}^{5} 2 = 2 + 2 + 2 + 2 + 2 \tag{8}$$

Si c es una constante que multiplica las observaciones:

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{9}$$

$$\sum_{i=1}^{n} cx_i = c \cdot x_1 + c \cdot x_2 + c \cdot x_3 + \cdots + c \cdot x_n \tag{10}$$

$$\sum_{i=1}^{n} cx_i = c\,(x_1 + x_2 + x_3 + \cdots + x_n) \tag{11}$$

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{12}$$

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

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{13}$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) + \cdots + (x_n + y_n) \tag{14}$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) \tag{15}$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{16}$$

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

$$\sum_{i=1}^{n} (x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \tag{17}$$

5.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{18}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \tag{19}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \frac{n}{n} \sum_{i=1}^{n} x_i - n\bar{x} \tag{20}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = n\bar{x} - n\bar{x} \tag{21}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{22}$$

6.

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

Si $y_i = a + bx_i$ siendo a y b constante

$$\bar{y} = a + b\bar{x} \tag{23}$$

En efecto:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + bx_i) \tag{24}$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i \tag{25}$$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{26}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{na}{n} + b\frac{\sum_{i=1}^{n} x_i}{n} \tag{27}$$

$$\bar{y} = a + b\bar{x} \tag{28}$$

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

Uno de los problemas es que la unidad de medida queda al cuadrado, e.g. si se miden cm, la varianza tiene unidades de cm^2:

- Varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{29}$$

- Varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{30}$$

$$s^2 = \sum_{i=1}^{n} \frac{x_i^2 - 2x_i\bar{x} + \bar{x}^2}{n-1} \tag{31}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} \tag{32}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{n}{n}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} \tag{33}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2}{n-1} \tag{34}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}{n-1} \tag{35}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \tag{36}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n\bar{x}^2}{n-1} \tag{37}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \tag{38}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{1}{n-1}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{39}$$

En algunos casos puede ser más conveniente calcular la varianza de esta forma.

Coeficiente de variación

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

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.

Se ordena la muestra y se toma la posicion que corresponde.

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

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

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{40}$$

Media ponderada

Media geometrica

Media armonica

Sirve en diseño de experimentos par aproximar el numero de replicas en todo el expermineto cuando el diseño es desbalanceado.

Ejemplo en excel

Tabla de pesos de mujeres en una empresa

1. construcción de la tabla de frecuencias

- Definición de los intervalos: podemos utilizar la siguiente ecuación:

$$k = 1 + 3.322log_{10}(n) \tag{41}$$

Para esta muestra de 50 pesos de mujeres:

$$k = 1 + 3.322log_{10}(50) \tag{42}$$

$$k = 6.6 \approx 7 \tag{43}$$

Vamos a usar 7 intervalos, para saber la longitud dividimos el rango en el numero de intervalos.

Para hallar el rango podemos importar los datos a R:

```
tablaPesos <- read.table("TABLAPESOS.txt", header = T)
```

Y luego preguntar sobre el valor maximo y el minimo:

```
max(tablaPesos)
```

```
## [1] 72
```

```
min(tablaPesos)
```

```
## [1] 53
```

Podemos entonces calcular la longitud de cada intervalo:

```
(max(tablaPesos)-min(tablaPesos))/7
```

```
## [1] 2.714286
```

Obtenemos una longitud de intervalo de $2.71 \approx 3$

Los intervalos entonces serían:

```
##   intervalo valores
## 1         1   53 55
## 2         2   56 58
## 3         3   59 61
## 4         4   62 64
## 5         5   65 67
## 6         6   68 70
## 7         7   71 73
```

En excel:

- Usamos los limites superiores de los intervalos
  - Datos, analisis de datos, histograma, aceptar, seleccionar rango de entrada, rango de clases son los limites superiores. Al hacer esto, excel genera una tabla como la siguiente:

Clase	Frecuencia
55	2
58	5
61	9
64	15
67	12
70	5
73	2
y mayor...	0

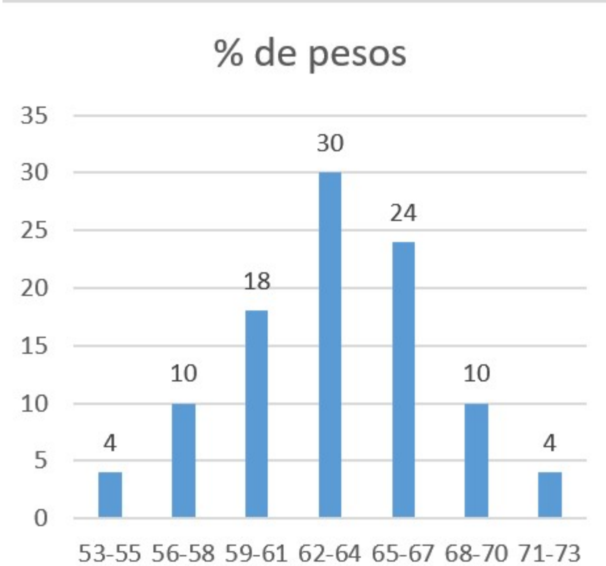
Figure 1: tabla de frecuencias en excel

Luego, a partir de esta tabla podemos calcular todas las frecuencias, la frecuencia absoluta (f_i), recuencia relativa (f_r), la frecuencia absoluta acumulada (F_i) y la frecuencia relativa acumulada (F_r):

	absoluta	relativa	abs. acumul.	rel. acumul.
intervalo	fi	fr	Fi	Fr
53-55	2	0.04	2	0.04
56-58	5	0.1	7	0.14
59-61	9	0.18	16	0.32
62-64	15	0.3	31	0.62
65-67	12	0.24	43	0.86
68-70	5	0.1	48	0.96
71-73	2	0.04	50	1

AF Andrés Felipe Beltrán Rodri... G44
El 62% de las mujeres de la muestra
pesan entre 53 y 64 kg
10/26/2021 6:51 AM

Luego a partir de esta tabla de frecuencias, utilizando las columnas de intervalo y % de frecuencia podemos construir un histograma como el siguiente:



Para hacer una operación análoga en R podemos crear los intervalos de la siguiente manera:

```
mins <- seq(53,71, by = 3)
maxs <- seq(55,73, by = 3)
```

Luego, podemos juntar las dos columnas de limites inferiores y superiores de intervalo en la tabla `TDF`:

```
TDF <- data.frame(min = mins,
                                          max = maxs )
TDF
```

```
##   min max
## 1  53  55
## 2  56  58
## 3  59  61
## 4  62  64
## 5  65  67
## 6  68  70
## 7  71  73
```

Luego podemos iterar a lo largo de las filas de `tablasDefrecuencias` buscando cuantos elementos de `tablaDePesos` están dentro del intervalo definido por cada fila:

```
for(i in 1:nrow(TDF)) {
  TDF$Freq[i] <-
      length(
          which(
mins[i] <=  tablaPesos & tablaPesos<= maxs[i]
}
```

Podemos luego calcular la frecuencia relativa dividiendo por el total de observaciones:

```
TDF$fr <-TDF$Freq/
    sum(TDF$Freq)

TDF
```

```
##   min max Freq   fr
## 1  53  55    2 0.04
## 2  56  58    5 0.10
## 3  59  61    9 0.18
## 4  62  64   15 0.30
## 5  65  67   12 0.24
## 6  68  70    5 0.10
## 7  71  73    2 0.04
```

Luego podemos calcular la frecuencia absoluta y relativa acumuladas para cada intervalo de manera ascendiente:

```
for(i in 1:nrow(TDF)){

TDF$Fi[i] <- sum(TDF$Freq[1:i])

}
```

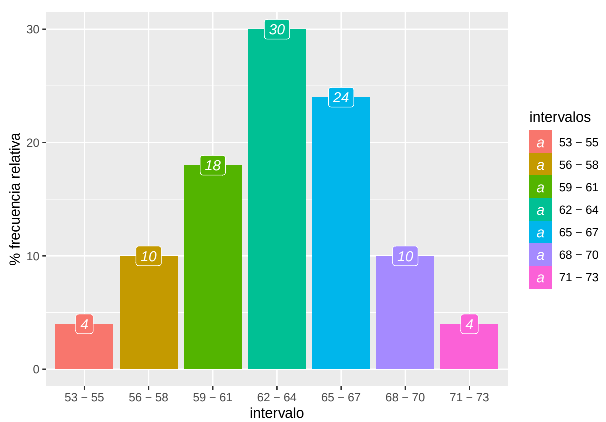
Podemos hacer lo mismo para la frecuencia relativa acumulada (F_r):

```
for(i in 1:nrow(TDF)){

TDF$Fr[i] <- sum(TDF$fr[1:i])

}
```

Una vez tenemos la tabla de frecuencias completa, podemos hacer la gráfica de frecuencias porcentuales:

```
library(ggplot2)
intervalos <- factor(paste(TDF$min,'-',TDF$max))

ggplot(TDF,
       aes(x = intervalos,
           y = fr*100,
           fill=intervalos,
           label = round(fr*100,2)
           )
       ) +
    geom_bar(stat="identity") +
  xlab("intervalo") +
    ylab('% frecuencia relativa')+
```

```
  geom_label(aes(fill = intervalos),
             colour = "white",
             fontface = "italic")
```



Calculo de medidas descriptivas

```
mean(tablaPesos$pesos)
```

```
## [1] 63.2
```

```
quantile(tablaPesos$pesos)
```

```
##   0%  25%  50%  75% 100%
## 53.0 61.0 63.5 66.0 72.0
```

El 75% de las mujeres pesan entre 53 y 66 kg. El 25% restante pesa entre 66 kg y 72 kg

- la varianza: Es importante calcular la varianza muestral y no la varianza poblacional, dado que no se puede saber la poblacional (σ^2) si no su estimador (s^2).
- Algunos paquetes hacen calculos con la varianza poblacional

En R: 'The denominator n - 1 is used which gives an unbiased estimator of the (co)variance for i.i.d. observations.'

```
var(tablaPesos$pesos)
```

```
## [1] 17.10204
```

Para calcular la varianza poblacional:

```
pesos <- tablaPesos$pesos
sum((pesos - mean(pesos))^2)/length(pesos)
```

```
## [1] 16.76
```

para el calculo de la desviacion estandar:

```
sd(pesos)
```

```
## [1] 4.135461
```

```
sqrt(var(pesos))
```

```
## [1] 4.135461
```

Interpretación:

Varianza: Tiene unidades al cuadrado. Tendriamos que tener otro grupo de comparación, con una medición similar.

Desviacion estandar: la desviacion de las observaciones respecto al promedio es de 4.14 unidades de masa (kg).

Para calcular otros estadísticos descriptivos podemos utilizar en excel:

- Datos > Análisis de datos > Estadística descriptiva > Se selecciona rango de entrada y de salida.

Se obtiene la siguiente tabla:

Dentro de esta tabla está el error típico

Error típico - error estándar

$$\frac{s}{\sqrt(n)} \tag{44}$$

En este caso:

```
sd(pesos)/sqrt(50)
```

```
## [1] 0.5848426
```

Se utiliza para inferencia, para intervalos de confianza. En diseño de experimentos sirve para el cálculo de tamaño de muestra. Se espera que no aumente el número de réplicas si no disminuye lo suficiente el error típico.

Columna1	
Media	63.2
Error típico	0.5848426
Mediana	63.5
Moda	64
Desviación estándar	4.1354614
Varianza de la muestra	17.102041
Curtosis	-0.0713723
Coeficiente de asimetría	-0.2023508
Rango	19
Mínimo	53
Máximo	72
Suma	3160
Cuenta	50

Figure 2: Cuadro de estadística descriptiva en excel

```
library(e1071)
kurtosis(pesos)
```

```
## [1] -0.2936688
```

```
skewness(pesos)
```

```
## [1] -0.1903716
```

La distribucion de pesos tiene una curtosis < 3, lo cual indica que es mas aplanada que una distribucion normal, o tiene hombros más pesados.

Además, tiene un coeficiende de asimetría cercano a cero, lo cual indica un ligero sesgo con cola hacia valores menores de peso.

```
range(pesos)
```

```
## [1] 53 72
```

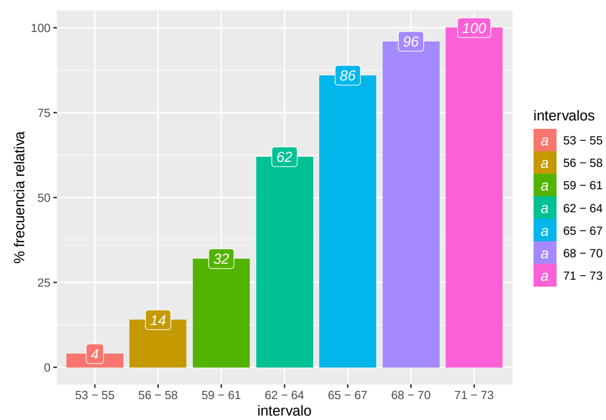
```
range(pesos)[2] - range(pesos)[1]
```

```
## [1] 19
```

La persona que mas peso tiene, tiene 19 kg mas que la persona de menos peso.

Tambien podemos graficar la frecuencia acumulada:

```
library(ggplot2)
intervalos <- factor(paste(TDF$min,'-',TDF$max))

ggplot(TDF,
       aes(x = intervalos,
           y = Fr*100,
           fill=intervalos,
           label = round(Fr*100,2)
           )
       ) +
    geom_bar(stat="identity") +
  xlab("intervalo") +
    ylab('% frecuencia relativa')+
  geom_label(aes(fill = intervalos),
             colour = "white",
             fontface = "italic")
```



el 86% de las mujeres pesan entre 53 y 67 kg.

Desviación media absoluta

Ejemplo de edades

Este ejemplo está en el código llamado `medidasdescriptivas.R`

Proporción

Es similar al promedio, para variables de tipo cualitativo