

Unidad 1.2 Tablas de distribución e indicadores estadísticos

Módulo 1

Daniel Enrique González Gómez
Universidad Javeriana Cali

2021-08-02

AGENDA

1. Presentación guía de aprendizaje 1.2
2. Tablas de frecuencia
3. Indicadores estadísticos
4. Varios





+20.5 c
→  1049 km
km

Tablas de frecuencia variables cualitativas

Las distribuciones de frecuencia o también llamadas tablas de frecuencia nos sirven para agrupar los datos y así permitir resumir para poder tener una idea más clara de sus características.

Para las variables cualitativas la tabla posee 3 columnas :

- C1: los diferentes **valores** que toma la variable.
- C2: **frecuencia absoluta** que consiste en el conteo para cada uno de los valores distintos que toma la variable.
- C3: **frecuencia relativa** que corresponde al porcentaje la cantidad de datos para cada los valores

```
# Forma simple  
table(ventas$Tipo_Cliente)
```

```
##  
## Promocional      Regular  
##           140          60
```

```
# Forma simple  
t=table(ventas$Tipo_Cliente)  
prop.table(t)
```

```
##  
## Promocional      Regular  
##           0.7          0.3
```

```
#utilizando summarytools  
library(summarytools)  
t1=freq(ventas$Metodo_Pago, cumul = FALSE, headings = FALSE)  
t1
```

```
##  
##  
## Freq % Valid % Total  
## -----  
## American Express 4 2.00 2.00  
## Discover 8 4.00 4.00  
## MasterCard 28 14.00 14.00  
## Star Card 140 70.00 70.00  
## Visa 20 10.00 10.00  
## <NA> 0 0.00 0.00  
## Total 200 100.00 100.00
```

Nota: paquete [summarytools](#)

Tablas de frecuencia para variables cuantitativas

Para las variables cuantitativas las tablas de frecuencias tiene una presentación diferente a la vista anteriormente. Como se trata de variables con una gran numero de valores diferentes, es necesario dividirlas por intervalos .

```
library(agricolae)
h2=with(ventas,graph.freq(Edad,plot=FALSE));t2=table.freq(h2);
colnames(t2) = c(" LI ", " LS ", "marca clase'", "Frec.Abs","Frec.Rel", "Frec.Abs.Ac","Frec
t2
```

##	LI	LS	marca clase'	Frec.Abs	Frec.Rel	Frec.Abs.Ac	Frec.Rel.Ac
## 1	20.0	26.4	23.2	10	5	10	5
## 2	26.4	32.8	29.6	42	21	52	26
## 3	32.8	39.2	36.0	28	14	80	40
## 4	39.2	45.6	42.4	36	18	116	58
## 5	45.6	52.0	48.8	36	18	152	76
## 6	52.0	58.4	55.2	26	13	178	89
## 7	58.4	64.8	61.6	10	5	188	94
## 8	64.8	71.2	68.0	6	3	194	97
## 9	71.2	77.6	74.4	4	2	198	99
## 10	77.6	84.0	80.8	2	1	200	100

Frec.Abs : Frecuencia absoluta ; **Frec.Rel** : Frecuencia relativa ; **Frec.Abs.Ac** : Frecuencia Absoluta Acumuada ;
Frec.Rel.Ac : Frecuencia Relativa Acumulada

```
library(stringr)
t1=freq(Colombia$estado, cumul = FALSE, headings = FALSE)
t1
```

```
##  
##  
##          Freq   % Valid   % Total  
## ----- -----  
##    Fallecido    120998    2.523729    2.523729  
##    Grave        3564    0.074337    0.074337  
##    leve         12312    0.256799    0.256799  
##    Leve        4621455   96.392489   96.392489  
##    LEVE          2    0.000042    0.000042  
##    Moderado     19891    0.414879    0.414879  
##    N/A          16192    0.337726    0.337726  
##    <NA>            0           0.000000  
##    Total        4794414   100.000000   100.000000
```

Rango percentil

Es un número que divide la muestra en dos partes. x % de los datos de la muestra son iguales o menores que P_x y un $(100 - x)$ % por encima de el.



- Participé en una carrera **K10** y mi posición correspondió al percentil 30: P_{30}
- Mi nota en un examen de matemáticas y mi posición fue el percentil noventa: P_{90}
- Que significa: P_{25} ; P_{50} ; P_{75}

Diagrama de cajas

```
boxplot(ventas$Edad)
```

Características de los datos

Tendencia central

- media
- mediana
- moda
- media truncada
- rango medio
- media armónica
- media geométrica

Dispersión

- rango
- varianza
- desviación estándar
- coeficiente de variación

Forma

- sesgo o asimetría
- curtosis

Media aritmética :

$$\hat{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Es una de los indicadores estadísticos mas conocidos

Propiedades de la media :

- La suma de las desviaciones de los datos con respecto a la media es cero. $\sum(x_i - \bar{x}) = 0$.
- La suma de los cuadrados de las desviaciones de los datos con respecto a un valor a es mínimo cuando $a = \bar{x}$.
- Si $x_i = k$ para todo i , entonces, $\bar{x} = k$.
- Si todos los datos de una variable se multiplican por una constante k , es decir $y_i = kx_i$, entonces $\bar{y} = k\bar{x}$
- Si $z_i = ax_i + by_i$, donde: a, b constantes y x_i, y_i variables, entonces: $\bar{z} = a\bar{x} + b\bar{y}$.

```
mean(Colombia$edad,na.rm = TRUE)
```

```
## [1] 39.57437
```

```
mean(ventas$Edad, na.rm = TRUE)
```

```
## [1] 43.08
```

PROBLEMA

```
x=1:10
```

```
x
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

```
cat("media :",mean(x))
```

```
## media : 5.5
```

```
x[10]=20
```

```
x
```

```
## [1] 1 2 3 4 5 6 7 8 9 20
```

```
cat("media :" ,mean(x))
```

```
## media : 6.5
```

Mediana :

Me: Es el número que divide la muestra en dos partes de igual proporción (50% : 50%). Es decir que corresponde a:

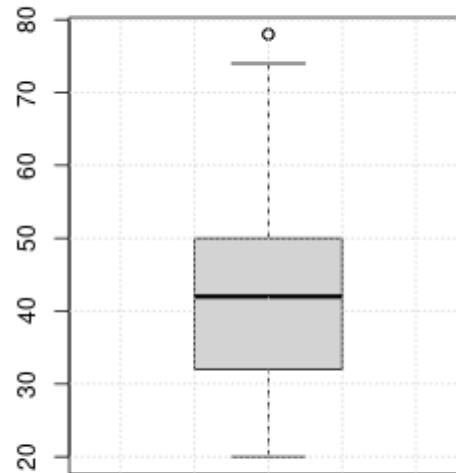
$$P_{50} = D_5 = Q_2$$

Tambien es corresponde a la linea central del diagrama de cajas.

```
median(Colombia$edad,na.rm = TRUE)
```

```
## [1] 37
```

```
boxplot(ventas$Edad)  
grid()
```



La **Me** corresponde a la linea central de a caja en el diagrama de cajas

La mediana es mas robusta a los cambio en los datos extremos. En presencia de datos atípicos es mejor utilizar la mediana en lugar que la media.

```
x=1:10
```

```
x
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

```
cat("mediana :" ,median(x))
```

```
## mediana : 5.5
```

```
x[10]=20
```

```
x
```

```
## [1] 1 2 3 4 5 6 7 8 9 20
```

```
cat("mediana :" ,median(x))
```

```
## mediana : 5.5
```

La Moda

Mo : Dato o valor que más se repite. Es utilizada como medida de tendencia central en variables cualitativas o en cuantitativas discretas con pocos valores. En una tabla o gráfico se puede distinguir fácilmente.

```
#utilizando summarytools  
library(summarytools)  
t1=freq(ventas$Metodo_Pago, cumul = FALSE, headings = FALSE)  
t1
```

```
##  
##  
## Freq % Valid % Total  
## -----  
## American Express 4 2.00 2.00  
## Discover 8 4.00 4.00  
## MasterCard 28 14.00 14.00  
## Star Card 140 70.00 70.00  
## Visa 20 10.00 10.00  
## <NA> 0 0.00  
## Total 200 100.00 100.00
```

Moda : ?

Otras medidas de centro

- Media truncada al 10%

```
mean(ventas$Ventas_Netas, na.rm = TRUE, trim = 0.10)
```

```
## [1] 827.25
```

- **Rango medio** : $\frac{1}{2}(\max(x) + \min(x))$

```
(max(ventas$Ventas_Netas,na.rm = TRUE)+min(ventas$Ventas_Netas,na.rm = TRUE))/2
```

```
## [1] 14388.5
```

- **Media geométrica** : este indicador de tendencia central se utiliza para promediar tasa de crecimiento o de interés. Para encontrar su valor se multiplican los valores de n tasas incrementadas en uno. A ese producto se le extrae la raíz n -ésima.
- **Media armónica** : Este indicador corresponde al inverso de la media aritmética

Problema reconocimiento de grupo

Grupo 1

Edades : 19, 22, 18, 21



Promedio : 20 años

Hace falta otro indicador que nos oriente de cual grupo hablamos cuando solo tenemos como información : media = 20 años.

Grupo 2

Edades : 39, 38, 2, 1

Promedio : 20 años



Indicadores de Dispersion

Rango

$$r = \max(x) - \min(x)$$

En caso de los dos grupos:

Grupo 1:

$$\bar{x} = 20 \text{ años}$$

$$r = 4 \text{ años}$$

Grupo 2:

$$\bar{x} = 20 \text{ años}$$

$$r = 38 \text{ años}$$

Indicador muy útil cuando se deben realizar cálculos rápidos

Varianza s^2

Es la medida de dispersión más utilizada en estadística y está definida por

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

Propiedades de la varianza

- $s^2 = \frac{1}{n} \sum x_i^2 - (\bar{x})^2$
- La varianza es siempre no negativa $s^2 \geq 0$
- La varianza de una constante es cero $s_k^2 = 0$
- Si $y_i = kx_i$, entonces $s_y^2 = k^2 s_x^2$
- Si $y_i = x_i + k$, entonces $s_y^2 = s_x^2$
- Si $z_i = ax_i + by_i$, entonces $s_z^2 = a^2 s_x^2 + b^2 s_y^2 + 2ab \text{cov}(xy)$

La varianza se puede interpretar como el promedio de las diferencias cuadradas entre cada uno de los datos y la media

El problema de la varianza es su **interpretación**

Sus unidades son al cuadrado y en la mayoría de los casos no es posible interpretarlos. Por esta razón se optó por utilizar otra medida de dispersión

Desviación estándar

Es la raíz cuadrada de la varianza

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

Nota : no aplican todas las propiedades de la varianza

```
cat( "Varianza" : " , var(ventas$Edad), "\n" )
```

```
## Varianza : 152.7172
```

```
cat( "Desviación estándar :" , sd(ventas$Edad) )
```

```
## Desviación estándar : 12.35788
```

Aunque la desviación estándar reduce el problema debido a tener las mismas unidades de la variable, es útil para comparación de dos grupos

Coeficiente de variación

Nos indica que tan grande o que tan pequeña es la desviación estándar con respecto a su media

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

Existen diferentes reglas empíricas para la interpretación del coeficiente de variación. Una de ellas establece como límite el 20% para separar los grupos homogéneos de los heterogéneos. Por lo general se utiliza un valor hasta el 20% para determinar que un grupo de datos son homogéneos, de lo contrario se calificará como heterogéneo.

```
cat("Coeficiente de variación :",sd(ventas$Edad)/mean(ventas$Edad)*100)
```

```
## Coeficiente de variación : 28.68589
```

Indicadores de forma

Curtosis

Se mide a través del coeficiente de curtosis que mide cuan **puntiaguda** es una distribución respecto a la curva de la distribución normal entandar.

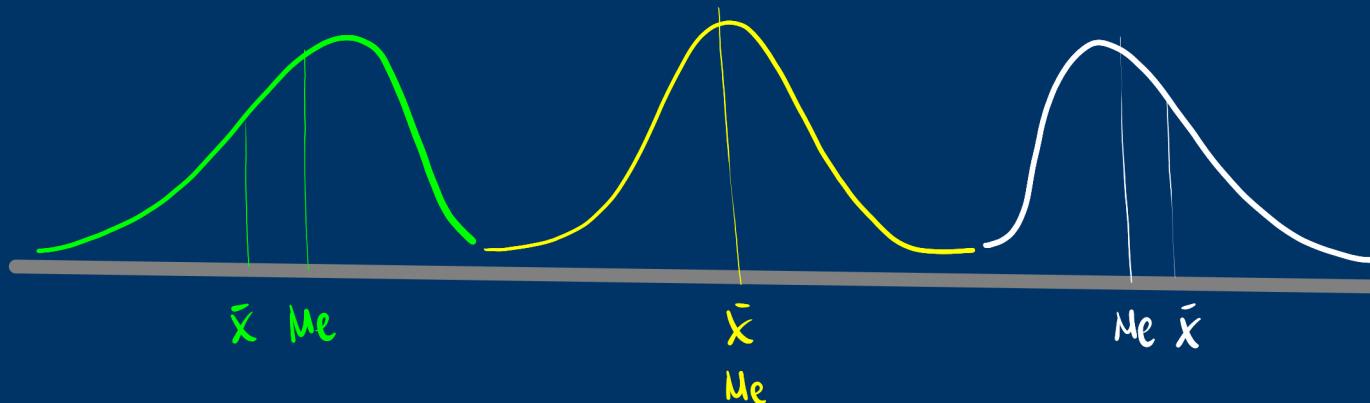
De acuerdo con su valor, la puntudez de los datos puede clasificarse en tres grupos:

- **Leptocúrtica**, con valores grandes para el coeficiente ($CA>0$)
- **Mesocúrtica**, con valores medianos para el coeficiente ($CA=0$)
- **Platicúrtica**, con valores pequeños para el coeficiente ($CA<0$)

Asimetría o sesgo

Mide que tanto la forma de la distribución de frecuencias de los datos es simétrica o no con respecto a la media. Esta característica de los datos se mide a través del coeficiente de asimetría o sesgo.

- Es **simétrica** si el valor del indicador es 0 ($\bar{x} = Me$)
- Es **asimétrica a la izquierda** si el valor del indicador es negativo ($\bar{x} < Me$)
- Es **asimétrica a la derecha** si el valor del indicador es positivo ($\bar{x} > Me$)



- **Asimetría negativa** : Poco con poco, mucho con mucho
- **Simétrica** : Poco con poco, poco con mucho, mucho al rededor de un centro

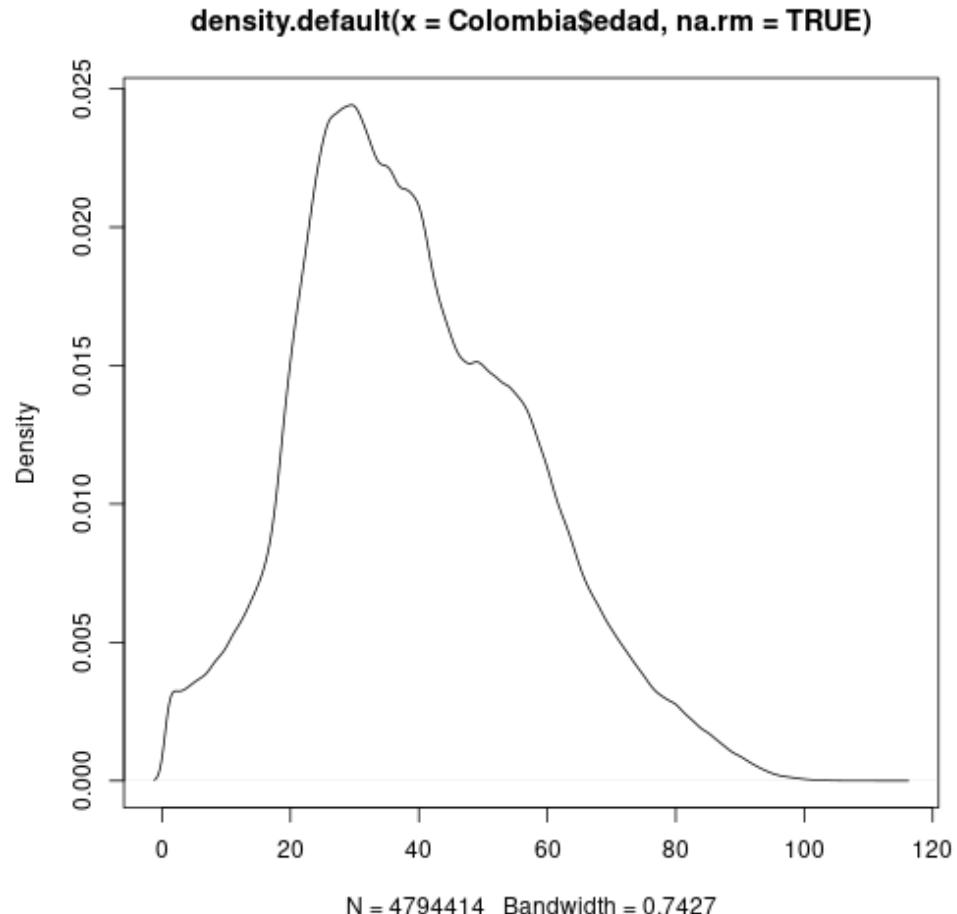
```
summarytools::descr(Colombia$edad)
```

```
## Descriptive Statistics
## Colombia$edad
## N: 4794414
##
##                               edad
## -----
##          Mean      39.57
## Std.Dev   17.89
##          Min      1.00
##          Q1      27.00
##          Median  37.00
##          Q3      52.00
##          Max     114.00
##          MAD     17.79
##          IQR     25.00
##          CV      0.45
##          Skewness 0.41
## SE.Skewness 0.00
##          Kurtosis -0.17
##          N.Valid  4794414.00
## Pct.Valid 100.00
```

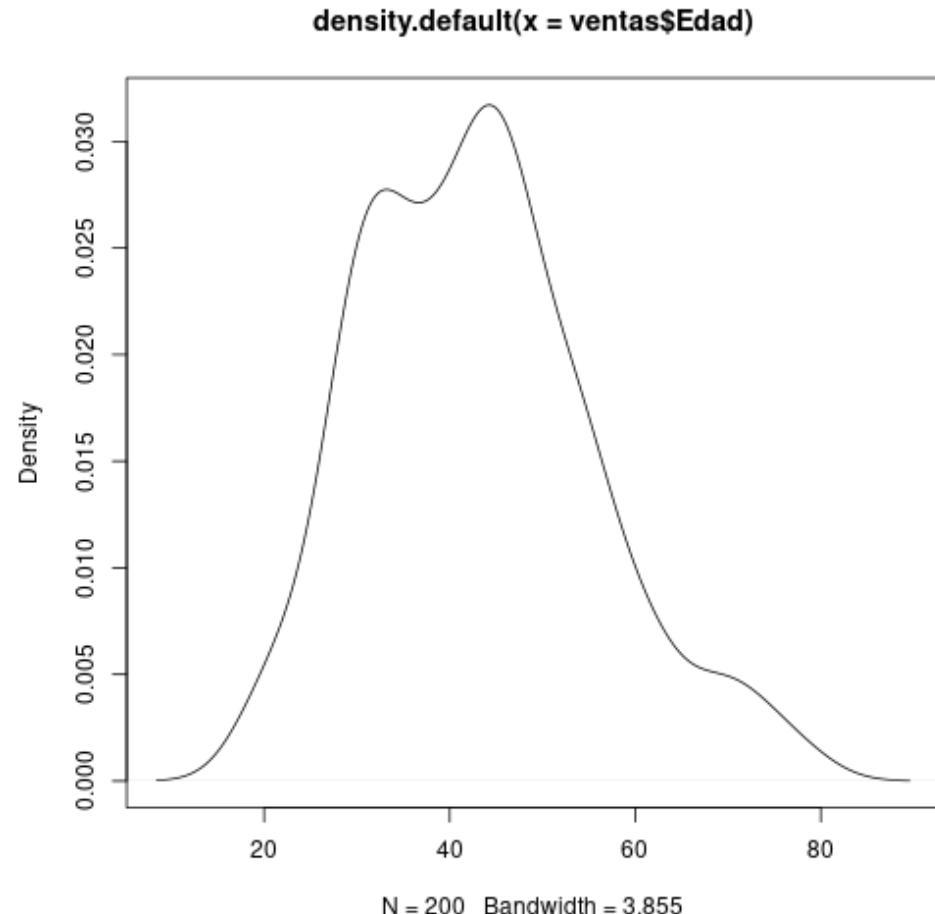
```
summarytools::descr(ventas$Edad)
```

```
## Descriptive Statistics
## ventas$Edad
## N: 200
##
##                               Edad
## -----
##          Mean      43.08
## Std.Dev   12.36
##          Min      20.00
##          Q1      32.00
##          Median  42.00
##          Q3      50.00
##          Max     78.00
##          MAD     13.34
##          IQR     18.00
##          CV      0.29
##          Skewness 0.51
## SE.Skewness 0.17
##          Kurtosis -0.03
##          N.Valid  200.00
## Pct.Valid 100.00
```

```
d1=density(Colombia$edad, na.rm=TRUE)  
plot(d1)
```



```
d2=density(ventas$Edad)  
plot(d2)
```



Actividades

- **Actividad 1 :** Solución del caso 101
 - **Nota:** RMarkdown permite realizar el trabajo fácilmente
- **Actividad 2 :** A partir de la información contenida en la base de datos seleccionada en la **Unidad 1.1**, realice un análisis de al menos una variable cualitativa y una cuantitativa teniendo como soportes las tablas de frecuencia y los indicadores estadísticos correspondiente.

Fecha : 08 agosto de 2021

Hora : 23:59 hora local

The background image shows a vast mountain range under a dramatic sky filled with white and grey clouds. In the foreground on the right, a person wearing an orange shirt and backpack stands on a rocky outcrop with their arms raised in triumph. The middle ground features a deep valley with green forests and winding roads.

Lo podemos lograr...

Daniel Enrique González Gómez

Imagen tomada de : <https://pixabay.com/es/images/search/paisaje/>