

Ayudantia3

Instalar librerías

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

Cargar librerías

```
## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.3      v purrr   0.3.4
## v tibble  3.1.1      v dplyr  1.0.5
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##   %+%, alpha

##
## Attaching package: 'FinCal'

## The following objects are masked from 'package:psych':
##
##   geometric.mean, harmonic.mean

## Registered S3 methods overwritten by 'rmutil':
##   method      from
##   plot.residuals psych
##   print.response httr
```

Cargar base de datos

A continuación se presenta la base de datos CASEN 2017, esta ha sido simplificada para el uso en ayudantía. La original tiene muchas más variables, para mayor información visitar <http://observatorio.ministeriodesarrollsocial.gob.cl/encuesta-casen-2017>

```
Casen_2017_simplificada <- read_dta('Casen2017.dta', encoding = 'ISO-8859-1')
```

Estadígrafos

Su funcionalidad es obtener características que describan los datos analizados. Estos son indicadores o medidas de resumen estadístico.

Estadígrafos de posición o de medidas de posición

Analizaremos la columna edad

```
summary(Casen_2017_simplificada$edad)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   19.00   36.00   37.78   56.00   117.00
```

```
glimpse(Casen_2017_simplificada$edad)
```

```
## num [1:216439] 56 21 24 28 26 26 1 62 36 59 ...
## - attr(*, "label")= chr "Edad"
## - attr(*, "format.stata")= chr "%10.0g"
```

Moda Es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor. No tiene por qué ser única, en caso de que no hay valor que se repita se dice que es amodal.

```
mfv(Casen_2017_simplificada$edad) #Indica el o los valores con más frecuencia
```

```
## [1] 50
```

Media La media aritmética se obtiene al sumar todos los datos que tenemos y dividir el resultado entre el número total de esos datos

```
mean(Casen_2017_simplificada$edad, na.rm = TRUE)
```

```
## [1] 37.78026
```

Mediana Representa el valor de la variable de posición central en un conjunto de datos ordenados.

```
median(Casen_2017_simplificada$edad, na.rm = TRUE)
```

```
## [1] 36
```

Cuantiles Medidas de posición no central que nos permiten reconocer otros puntos característicos de la distribución. Dividen a la distribución en un cierto número de partes de manera que en cada una de ellas hay el mismo número de valores de la variable.

Cuartiles Dividen a la distribución en cuatro partes iguales (tres divisiones). C1, C2, C3, correspondientes a 25%, 50%, 75%.

```
quantile(Casen_2017_simplificada$edad, probs = c(0, 0.25, 0.5, 0.75, 1))
```

```
##  0%  25%  50%  75% 100%
##   0   19   36   56  117
```

Deciles Dividen a la distribución en 10 partes iguales (nueve divisiones). D1, D2, D3, D4, ..., D10, correspondiente a 10%, 20%, 30%, 40%, ..., 90%.

```
quantile(Casen_2017_simplificada$edad, probs = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))

##    0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##     0   8  15  22  28  36  45  52  60  69 117
```

Percentiles Dividen a la distribución en 100 partes iguales (99 divisiones). P1, P2, P3, P4, ..., P100, correspondientes a 1%, 2%, 3%, 4%, ..., 99%.

```
quantile(Casen_2017_simplificada$edad, probs = c(.44, .77, .85))

## 44% 77% 85%
##  31  57  64
```

Estadígrafos de dispersión o medidas de dispersión

Para poder determinar esta variación en un grupo de datos respecto a una variable determinada se recurre a medidas de desviación o variación cuyo objetivo principal es medir el grado de dispersión o concentración de los valores o datos, alrededor de las medidas de tendencia central.

Rango intercuartil El rango intercuartilico de una variable de observación es la diferencia de sus cuartiles superior e inferior. Es una medida de cuán alejada está la porción media de los datos en valor. Se define como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), es decir: $RQ = Q3 - Q1$.

```
quantile(Casen_2017_simplificada$edad, probs = c(0, 0.25, 0.5, 0.75, 1))

##    0%  25%  50%  75% 100%
##     0  19  36  56 117

print('El rango intercuartil es:')

## [1] "El rango intercuartil es:"
IQR(Casen_2017_simplificada$edad, na.rm = T)

## [1] 37
```

Varianza La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones.

```
var(Casen_2017_simplificada$edad, na.rm = T)

## [1] 526.737
```

Desviación estándar La desviación estándar es un índice numérico de la dispersión de un conjunto de datos (o población). Mientras mayor es la desviación estándar, mayor es la dispersión de la población. La desviación estándar es un promedio de las desviaciones individuales de cada observación con respecto a la media de una distribución. En otras palabras, la desviación estándar es la raíz cuadrada de la varianza.

```
sd(Casen_2017_simplificada$edad, na.rm = T)

## [1] 22.95075
```

Asimetría Es la medida que indica la simetría de la distribución de una variable respecto a la media aritmética, sin necesidad de hacer la representación gráfica. Por regla general, la asimetría negativa indica que la media de los valores de los datos es menor que la mediana y que la distribución de los datos es sesgada a la izquierda. La asimetría positiva indicaría que la media de los valores de los datos es mayor que la mediana y que la distribución de los datos es sesgada a la derecha.

```
skew(Casen_2017_simplificada$edad, na.rm = T)
```

```
## [1] 0.2286902
```

Curtosis Una curtosis superior a 0 significa que la distribución es leptocúrtica y, por tanto, tiene un pico alto con colas delgadas. Por el contrario, una curtosis inferior a 0 significa que la distribución es platicúrtica y, por tanto, tiene un pico bajo y colas gruesas.

```
kurtosi(Casen_2017_simplificada$edad, na.rm = T)
```

```
## [1] -0.9278745
```

Coefficiente de variación Su fórmula expresa la desviación estándar como porcentaje de la media aritmética, mostrando una interpretación relativa del grado de variabilidad, independiente de la escala de la variable, a diferencia de la desviación típica o estándar. Por otro lado presenta problemas ya que a diferencia de la desviación típica este coeficiente es fuertemente sensible ante cambios de origen en la variable. Por ello es importante que todos los valores sean positivos y su media dé, por tanto, un valor positivo. A mayor valor del coeficiente de variación mayor heterogeneidad de los valores de la variable; y a menor C.V., mayor homogeneidad en los valores de la variable.

```
coefficient.variation(sd(Casen_2017_simplificada$edad), mean(Casen_2017_simplificada$edad))
```

```
## [1] 0.60748
```

```
(sd(Casen_2017_simplificada$edad)/mean(Casen_2017_simplificada$edad))
```

```
## [1] 0.60748
```