OPSO79-1-UCSH2021

Corrección prueba 2, Estadística descriptiva II y transformación avazanzada de data frames

29/10/2021

Revisión prueba 2

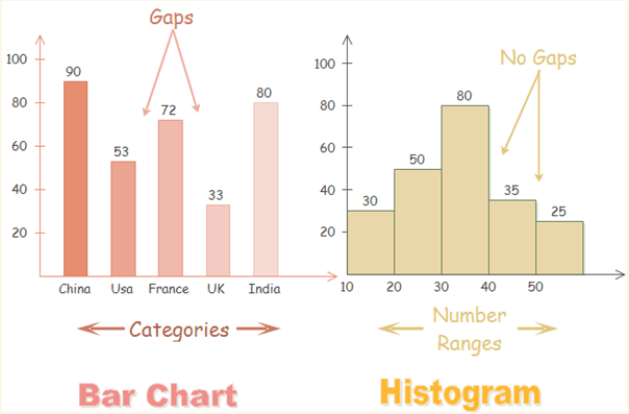
Estadística descriptiva y tranformación de data frames

Módulo I

Pregunta 1.1

Señale las principales diferencias entre un gráfico de barras y un histograma.

Gráfico de barras	Histograma
Variables categóricas	Variables numéricas
Espacio entre las barras	Barras una al lado de la otra
Anchura de las columnas no es relevante	Mientras más anchas las columnas mayor el intervalo representado



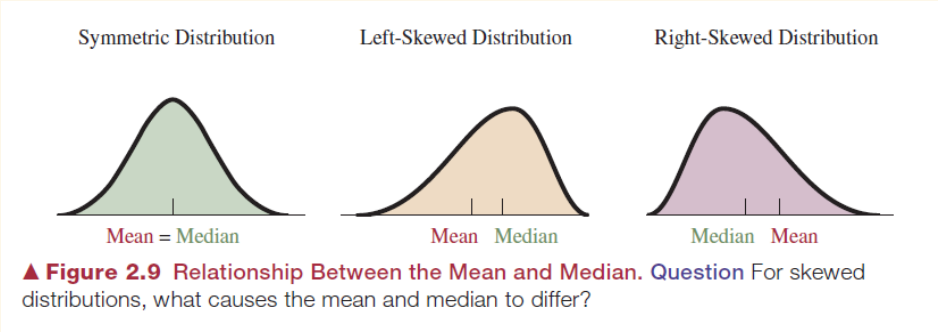
Módulo I

Pregunta 1.2

¿Qué es una distribución sesgada y mediante cuáles estadísticos se puede determinar?

Distribución en la que la media y mediana difieren de manera considerable por el efecto de valores atípicos (la media es tensionada hacia uno u otro lado).

Visualmente se identifica cuando una de las colas de la distribución de los datos es más larga que la otra.



▲ **Figure 2.9** Relationship Between the Mean and Median. **Question** For skewed distributions, what causes the mean and median to differ?

Módulo I

Pregunta 1.2

También con el estadístico skew (sesgo). Entre -0.5 y 0.5 datos bastante simétricos.

Datos simétricos

```
aleatoria <- rnorm(10000,mean = 5,sd=1)
library(moments)
summary(aleatoria)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.505   4.321   4.979   4.987   5.650   8.763
```

```
skewness(aleatoria)
```

```
## [1] 0.01713998
```

Módulo I

Pregunta 1.2

Datos Asimétricos

```
mean(x)
```

```
## [1] 4.278
```

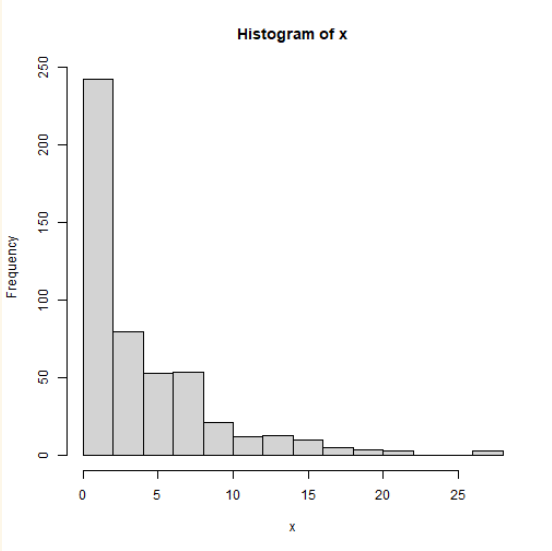
```
median(x)
```

```
## [1] 3
```

```
skewness(x)
```

```
## [1] 1.820958
```

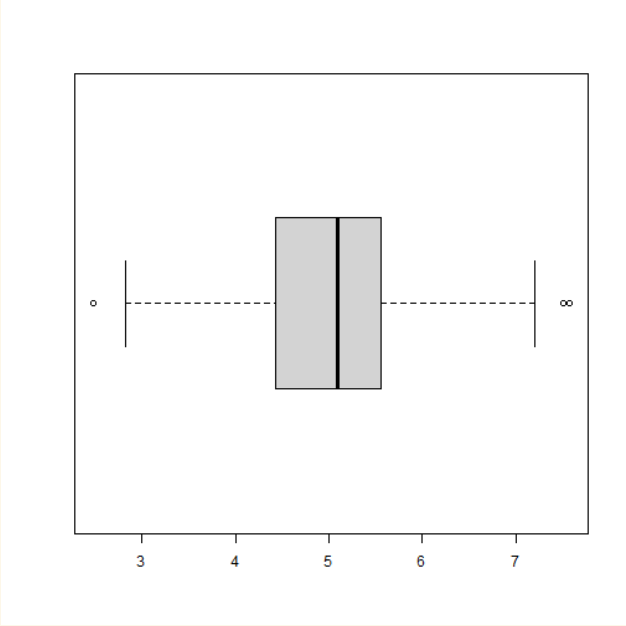
```
hist(x)
```



Módulo I

Pregunta 1.3

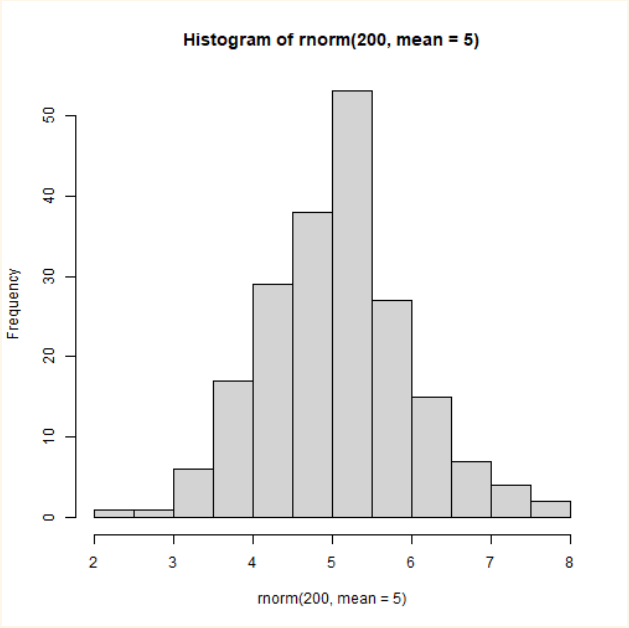
¿Cuáles son los estadísticos y otros elementos que un gráfico de cajas nos permite conocer de una variable?



Módulo I

Pregunta 1.3

¿Cuáles son los estadísticos y otros elementos que un gráfico de cajas nos permite conocer de una variable?



Módulo II

Trabaje con la Encuesta Latinobarómetro del 2020:
https://www.latinobarometro.org/latContents.jsp

Cargar la data y paquetes

```
library(dplyr)
data <- readRDS("data/Latinobarometro_2020_Esp_Rds_v1_0.rds")
```

Módulo II

Pregunta 2.1

¿Cuántas personas responden la encuesta en cada país de América Latina?, ¿Cuál es el país con mayor número de casos? Use `dplyr`.

```
data %>%
  group_by(idenpa) %>%
  summarise(n=n()) %>%
  arrange(-n) %>%
  slice(1)
```

```
## # A tibble: 1 x 2
##   idenpa          n
##   <hvn_lbll> <int>
## 1 76          1204
```

Brasil es el país con más casos (1204 casos)

Módulo II

Pregunta 2.2

¿Cuáles son las religiones que tienen mayor presencia en América Latina?, ¿Como es la distribución de las religiones?, ¿Cuáles son las principales diferencias de Chile respecto de América Latina en General?

América Latina

```
data %>% group_by(s10) %>%
  summarise(n=n()) %>% arrange(-n)
```

```
## # A tibble: 3 x 2
##   s10            n
##   <hvn_lbll> <int>
## 1  1         11262
## 2  2          3776
## 3 97          3309
```

```
sjmisc::frq(data[,"s10"])
```

Módulo II

Chile

```
data %>%
  filter(idenpa==152) %>%
  group_by(s10) %>%
  summarise(n=n()) %>% arrange(-n)
```

```
## # A tibble: 3 x 2
##   s10            n
##   <hvn_lbll> <int>
## 1  1           584
## 2 97           423
## 3  2           115
```

```
sjmisc::frq(data[data$idenpa==152,"s10"])
```

En Chile la católica también es la principal, con 48,7% de preferencias (7 puntos porcentuales bajo lo que alcanza en América Latina).

La segunda religión con mayor importancia es "ninguna", con un 35,25%.

Módulo II

Pregunta 2.3

¿Cómo es la distribución de la variable edad?, ¿Existen algún *outlier*?

```
summary(data$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      16      26      39      41      54     100
```

```
boxplot(data$edad, horizontal = TRUE)
```

Módulo II

Pregunta 2.3

Dos valores atípicos hacia arriba. ¿Cuáles son?

Son los valores mayores a 96, dado que $Q3 + 1,5 * (Q3 - Q1) = 96$

```
data %>% filter(edad>54+(1.5*(54-26))) %>%
  select(1,2,edad)
```

```
## # A tibble: 2 x 3
##   numinves   idenpa     edad
##   <hvn_lbll> <hvn_lbll> <hvn_lbll>
## 1 2020       188        100
## 2 2020       214         98
```

Módulo II

Pregunta 2.4

¿Que país presenta mayor variabilidad en la variable edad?

Desviación estándar sirve para conocer la variabilidad en un grupo, pero al tener cada grupo medias diferentes es necesario homogeneizar la medición de la variabilidad ocupando el **coeficiente de variación** (sd/media).

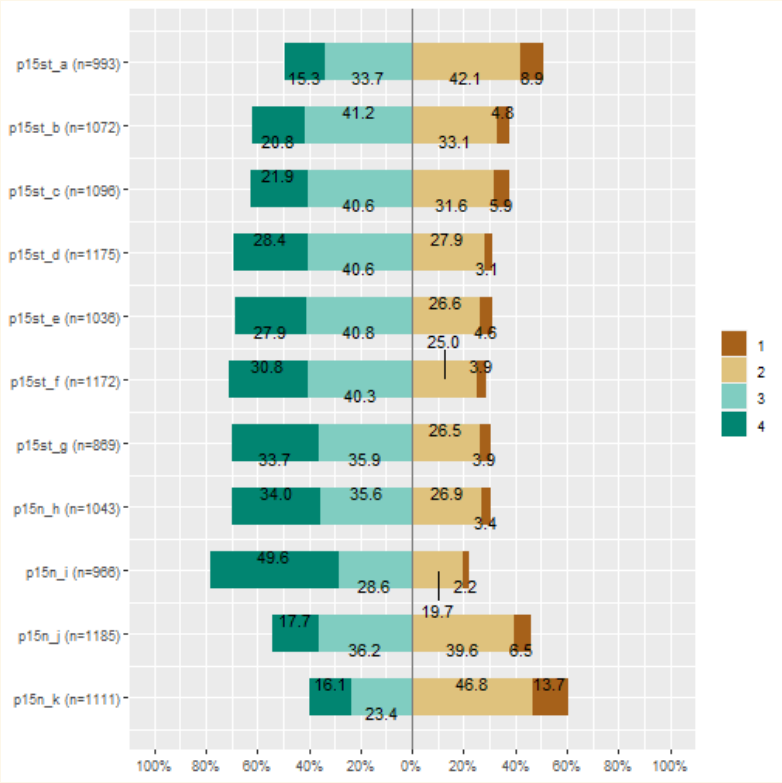
```
data %>%
  group_by(idenpa) %>%
  summarise(sd=sd(edad),
            cv = (sd(edad)/mean(edad)) ) %>%
  arrange(-cv)
```

```
## # A tibble: 2 x 3
##   idenpa        sd    cv
##   <hvn_lbll> <dbl> <dbl>
## 1 558         15.4 0.431
## 2 222         17.2 0.419
```

Nicaragua tiene el coeficiente de variación más alto (43,11%).

Módulo II

En Chile, ¿cuál de las instituciones listadas en P15 goza de mayor confianza?



Clínicas privadas, un 60,5% de los encuestados confía mucho o algo en ellas.

Módulo II

Pregunta 2.5

```
## Opción lenta (Esto para cada variable y comparar)
data %>%
  filter(idenpa==152) %>%
  mutate(p15n_k=if_else(p15n_k < 0, NA_real_, as.numeric(p15n_k))) %>%
  select(p15n_k) %>%
  table() %>%
  prop.table()
```

```
## .
##         1         2         3         4
## 0.1368137 0.4680468 0.2340234 0.1611161
```

Forma insuficiente (no se quitan valores perdidos):

```
prop.table(table(data[data$idenpa==152,]$p15n_k))
```

```
##
##         -2         -1          1          2          3          4
```

Módulo II (anexo)

Código detrás del gráfico de likert

```
data %>%
  filter(idenpa==152) %>%
  select(starts_with("p15")) %>%
  mutate(p15st_a=if_else(p15st_a < 0, NA_real_, as.numeric(p15st_a)),
         p15st_b=if_else(p15st_b < 0, NA_real_, as.numeric(p15st_b)),
         p15st_c=if_else(p15st_c < 0, NA_real_, as.numeric(p15st_c)),
         p15st_d=if_else(p15st_d < 0, NA_real_, as.numeric(p15st_d)),
         p15st_e=if_else(p15st_e < 0, NA_real_, as.numeric(p15st_e)),
         p15st_f=if_else(p15st_f < 0, NA_real_, as.numeric(p15st_f)),
         p15st_g=if_else(p15st_g < 0, NA_real_, as.numeric(p15st_g)),

         p15n_h=if_else(p15n_h < 0, NA_real_, as.numeric(p15n_h)),
         p15n_i=if_else(p15n_i < 0, NA_real_, as.numeric(p15n_i)),
         p15n_j=if_else(p15n_j < 0, NA_real_, as.numeric(p15n_j)),
         p15n_k=if_else(p15n_k < 0, NA_real_, as.numeric(p15n_k))) %>%
  plot_likert(catcount = 4)
```

Recursos web utilizados

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Bibliografía utilizada

Agresti, A. and C. Franklin (2018). *Statistics the Art and Science of Learning from Data*. Pearson Education Limited.