OPSO79-1-UCSH2021

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

29/10/2021

Estadística descriptiva II

Medidas de asociación entre variables

Introducción

Hace unas semanas vimos análisis descriptivos univariados.

Desde hoy comenzaremos con el análisis de dos (y más) variables, con propósito de investigar la asociación entre estas.

Una asociación existe entre dos o más variables si valores particulares de una son más probables que ocurran cuando ocurren ciertos valores en otra variable.

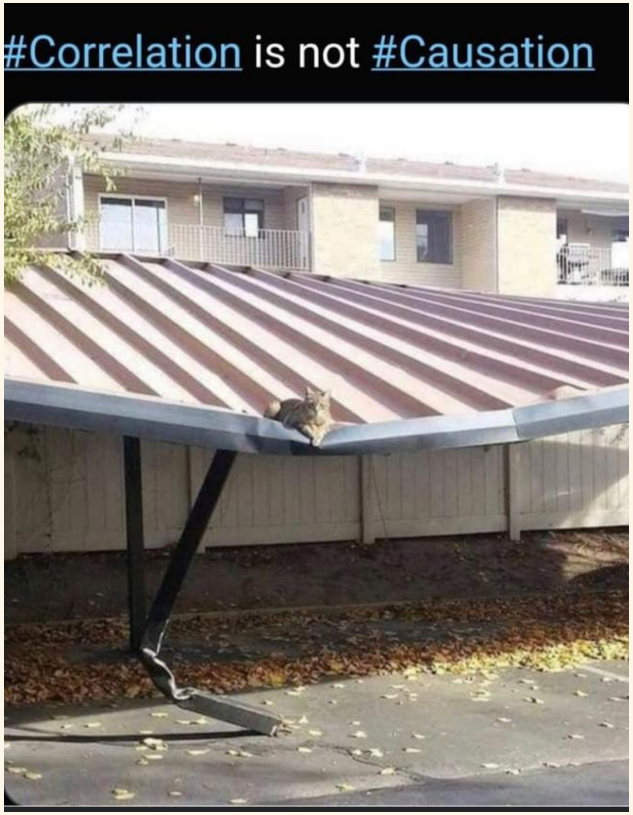
Asociación no es causalidad:

- Asociación necesaria pero no suficiente
- Causa antecede al efecto (explicación lógica)
- Control terceras variables (búsqueda de contrafactuales)

Hoy veremos métodos para estudiar si las asociaciones existen y para describir que tan fuertes son.

Se pueden distinguir según el tipo de variables (categóricas y numéricas)

Asociación no es causalidad



Asociaciones entre categóricas

Tablas de contingencia

Visualización de dos variables categóricas.

Las filas listan las categorías de una variable, mientras que las columnas las categorías de la otra variable.

Cada celda en la tabla es el número de observaciones en el set de datos que tiene la particular combinación de categorías de las dos variables.

¿Que posiciones políticas valoran más la democracia en América Latina?, ¿La izquierda, el centro, la derecha, los sin posición?

```
table(data$p47st_e,data$posicion_politica)
```

```
##
##      centro  der  izq ninguna
##   -2    178  212  197     212
##   1     207  253  229      52
##   2    1200  836  724     251
##   3    3069 1547 1632     720
##   4    2668 1500 1790    1159
```

Tablas de contingencia

Nivel de acuerdo con la frase *La democracia puede tener problemas pero es el mejor sistema de gobierno*, según posición política

```
library(sjmisc)
data <- data[data$p20st_a>=0,]  ## ojo, se quitan los valores perdidos
flat_table(data,p20st_a,posicion_politica)
```

```
##                       posicion_politica centro  der  izq ninguna
## p20st_a
## No preguntada                            0    0    0       0
## No aplicable                             0    0    0       0
## No sabe / No contesta                    0    0    0       0
## No sabe                                  0    0    0       0
## Muy de acuerdo                        1278  924  789     290
## De acuerdo                            3637 2154 1968    1025
## En desacuerdo                         1845  862 1271     616
## Muy en desacuerdo                      330  239  345     147
```

Dice poco...

Veamos con porcentajes...

Tablas de contingencia

Nivel de acuerdo con la frase *La democracia puede tener problemas pero es el mejor sistema de gobierno*, según posición política.

Porcentajes por filas (cada fila suma 100%)

```
flat_table(data,p20st_a,posicion_politica,margin = c("row"))
```

```
##                       posicion_politica centro   der   izq ninguna
## p20st_a
## No preguntada                           NaN   NaN   NaN     NaN
## No aplicable                            NaN   NaN   NaN     NaN
## No sabe / No contesta                   NaN   NaN   NaN     NaN
## No sabe                                 NaN   NaN   NaN     NaN
## Muy de acuerdo                        38.95 28.16 24.05    8.84
## De acuerdo                            41.40 24.52 22.40   11.67
## En desacuerdo                         40.16 18.76 27.67   13.41
## Muy en desacuerdo                     31.10 22.53 32.52   13.85
```

Tablas de contingencia

Nivel de acuerdo con la frase *La democracia puede tener problemas pero es el mejor sistema de gobierno*, según posición política.

Porcentajes por columnas (cada columna suma 100%)

```
flat_table(data,p20st_a,posicion_politica,margin = c("col"))
```

```
##                       posicion_politica centro   der   izq ninguna
## p20st_a
## No preguntada                             0.00  0.00  0.00    0.00
## No aplicable                              0.00  0.00  0.00    0.00
## No sabe / No contesta                     0.00  0.00  0.00    0.00
## No sabe                                   0.00  0.00  0.00    0.00
## Muy de acuerdo                           18.03 22.11 18.04   13.96
## De acuerdo                               51.30 51.54 45.00   49.33
## En desacuerdo                            26.02 20.63 29.06   29.64
## Muy en desacuerdo                         4.65  5.72  7.89    7.07
```

En los de derecha es donde hay más autopercibidos "demócratas" (73,6%), en la izquierda un un 63%

Tablas de contingencia

En Chile: nivel de acuerdo con la frase *La democracia puede tener problemas pero es el mejor sistema de gobierno*, según posición política.

Porcentajes por columnas (cada columna suma 100%)

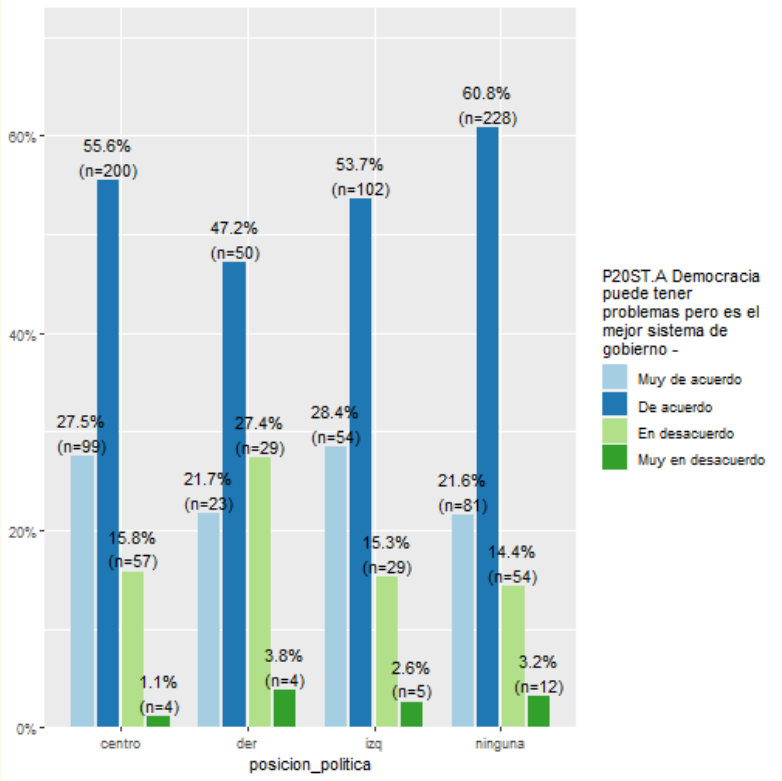
```
flat_table(data[data$idenpa==152,],
           p20st_a,posicion_politica,margin = c("col"))
```

```
##                       posicion_politica centro   der   izq ninguna
## p20st_a
## No preguntada                            0.00  0.00  0.00    0.00
## No aplicable                             0.00  0.00  0.00    0.00
## No sabe / No contesta                    0.00  0.00  0.00    0.00
## No sabe                                  0.00  0.00  0.00    0.00
## Muy de acuerdo                          27.50 21.70 28.42   21.60
## De acuerdo                              55.56 47.17 53.68   60.80
## En desacuerdo                           15.83 27.36 15.26   14.40
## Muy en desacuerdo                        1.11  3.77  2.63    3.20
```

La izquierda, el centro y los sin posición son más democráticos que la derecha

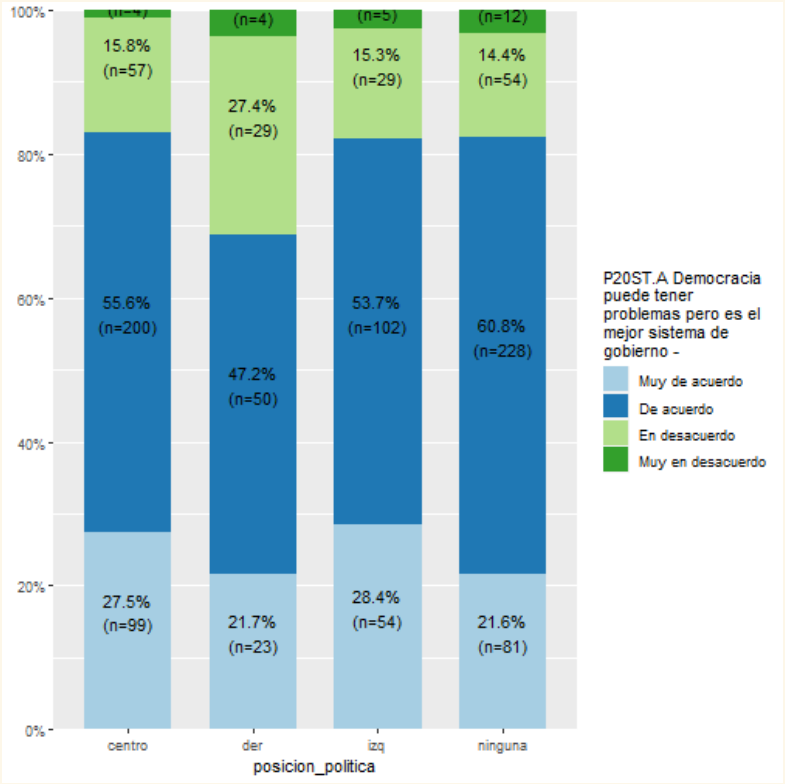
Gráficos de barras bivariados

```
library(sjPlot)
chile <- data[data$idenpa==152,]
plot_xtab(chile$posicion_politica,chile$p20st_a,margin = c("row"))
```



Gráficos de barras apilados

```
plot_xtab(chile$posicion_politica,chile$p20st_a,margin = c("row"),bar.po
```



Asociaciones entre categóricas

En conclusión, **existe asociación** entre la posición política las personas y su valoración a la democracia.

¿Como se vería si no existiese asociación? Algo así:

	centro	der	izq	ninguna
Muy de acuerdo	20%	20%	20%	20%
De acuerdo	50%	50%	50%	50%
En desacuerdo	15%	15%	15%	15%
Muy en desacuerdo	15%	15%	15%	15%

Para todas las posiciones políticas, los porcentajes de cada categoría de valoración de la democracia son iguales.

La fuerza de las asociaciones

Ratio entre las proporciones condicionales

¿Que tanta diferencia hay entre derecha e izquierda en el estar muy de acuerdo con la democracia?

Cuando no hay asociación:

```
0.20/0.20
```

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

Cuando hay asociación (entre izq y derecha)

```
0.28/0.21
```

```
## [1] 1.333333
```

La proporción de personas de izquierda muy de acuerdo con la democracia es 1,3 veces la de personas de derecha democráticas.

Asociaciones entre numérica y categórica

Asoc. entre categóricas y numéricas

La misma lógica que la descripción univariada, pero para cada grupo de interés.

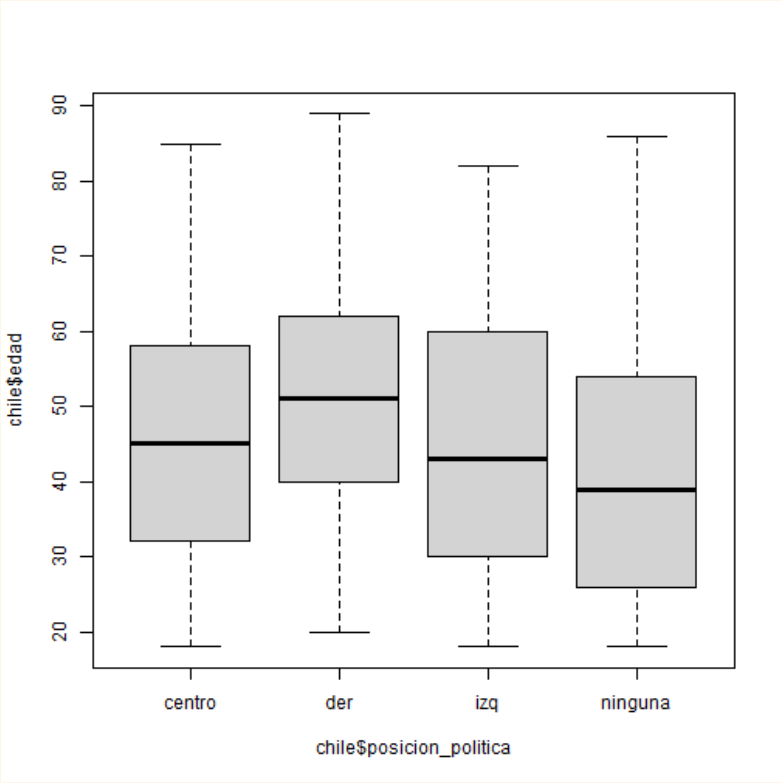
Relación entre edad y posición política

```
chile %>%
  group_by(posicion_politica) %>%
  summarise(mean=mean(edad),median=median(edad),sd=sd(edad),
            q25=quantile(edad,probs=0.25),q75=quantile(edad,probs=0.75))
```

posicion_politica	mean	median	sd	q25	q75
centro	45.57778	45	15.84534	32	58
der	50.28302	51	16.66176	40	62
izq	45.37368	43	17.80085	30	60
ninguna	41.10667	39	16.83053	26	54
NA	43.34247	42	17.48191	28	59

Gráfico de cajas con categórica

```
boxplot(chile$edad ~ chile$posicion_politica)
```



Asociaciones entre numéricas

Asociaciones entre numéricas

Lo más común es buscar una tendencia con un gráfico de puntos

Aunque no busquemos causalidad, por convención en el eje Y va la variable respuesta y en el X la variable explicativa.

¿Que tan asociada está la edad de una personas y sus ingresos?

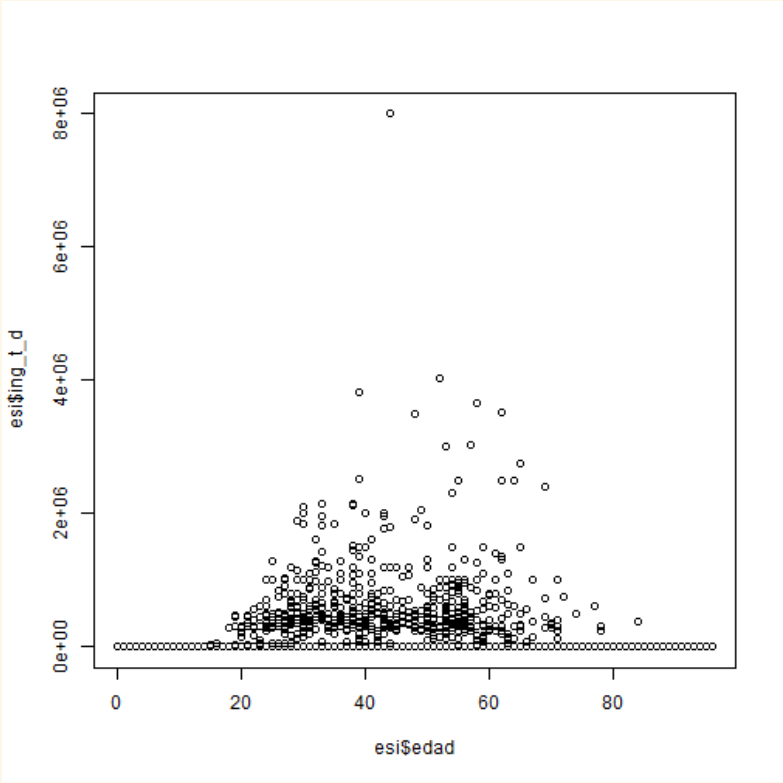
```
esi <- sample_n(haven::read_sav("data/esi-2019---personas_s.sav"),2000)
```

Se espera que a mayor edad mayores ingesos (hasta cierta edad).

Sin duda que los ingresos serían la dependiente (por lógica)

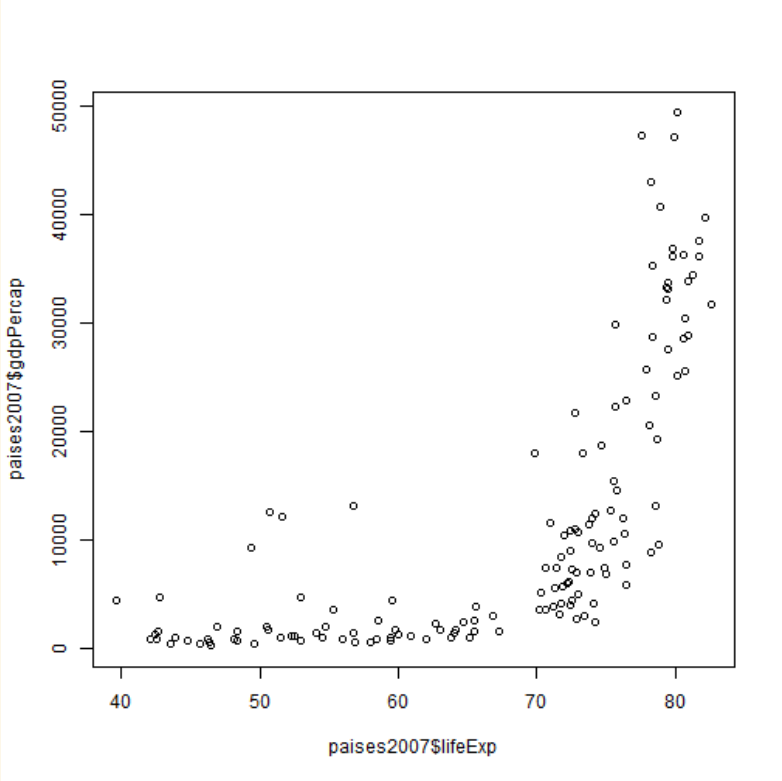
Asociaciones entre numéricas

```
plot(esi$edad,esi$ing_t_d)
```



Asociaciones entre numéricas

```
library(gapminder)
paises2007 <- gapminder %>% filter(year==2007)
plot(paises2007$lifeExp,paises2007$gdpPercap)
```



Asociaciones entre numéricas

La asociación será **positiva** si a medida que aumentan los valores en X aumentan los valores en Y.

La asociación será **negativa** si a medida que aumentan los valores en X disminuyen los valores en Y.

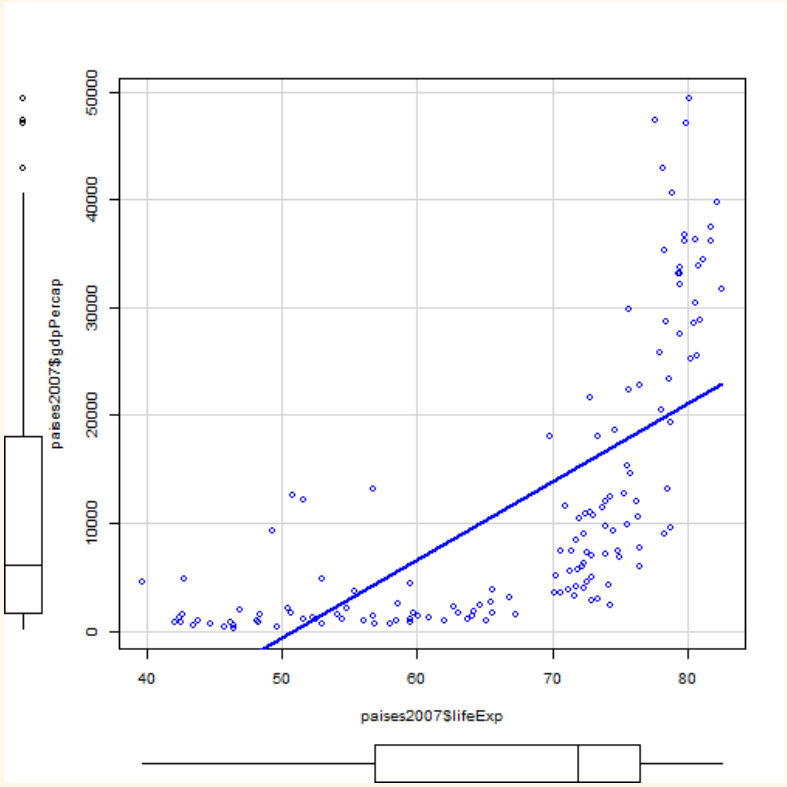
El paso siguiente es resumir la fuerza de la asociación.

Cuando la relación entre los datos sigue una tendencia de línea recta, se dice que los datos tienen una relación lineal.

A veces los toman una forma muy cercana a la recta (r = 1 o r = -1), otras veces ni se le parecen

La correlación mide la fuerza de esta asociación lineal.

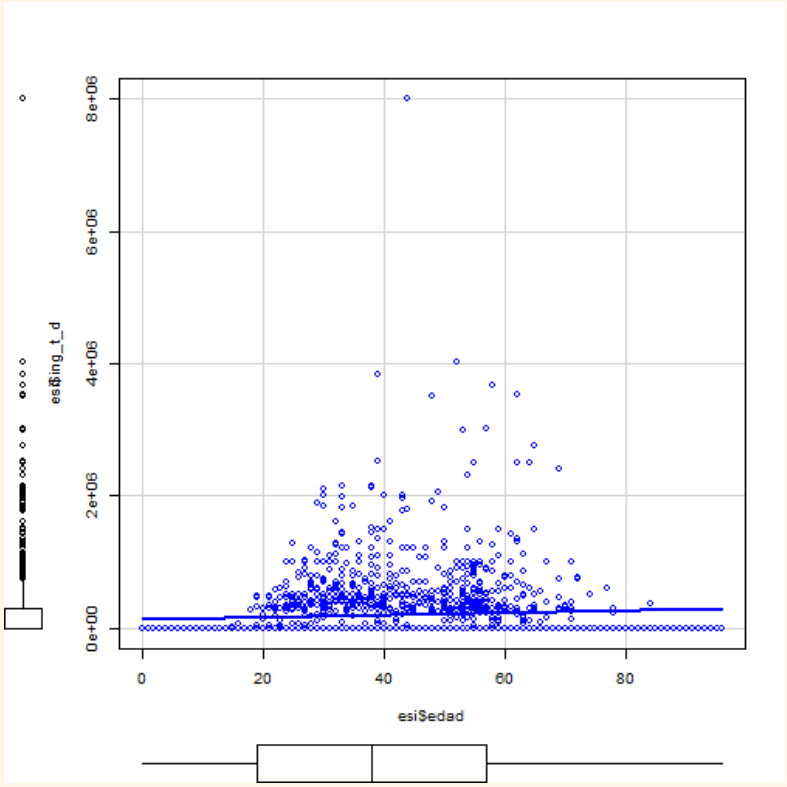
Asociaciones entre numéricas



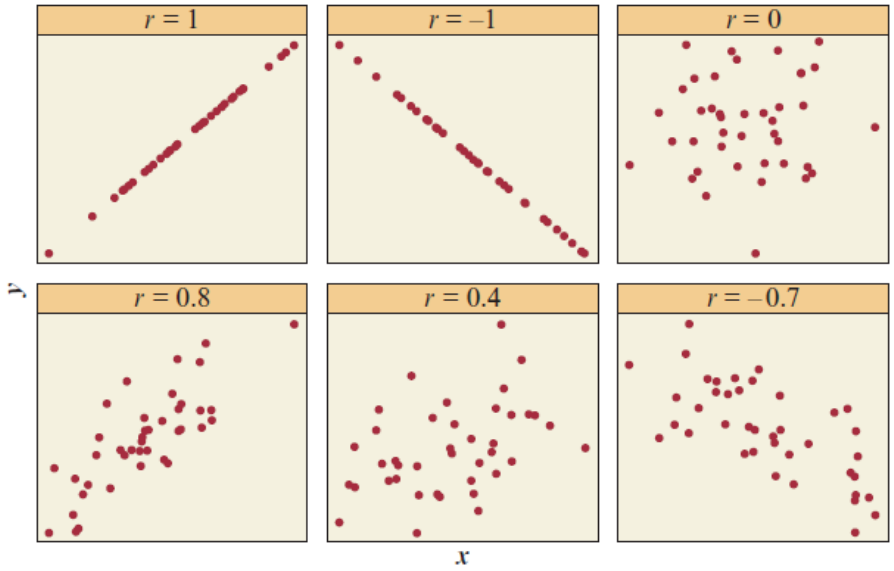
Hay mejores representaciones, pero una recta sirve.

Para edad e ingresos definitivamente no es lo más certero.

Asociaciones entre numéricas



Correlaciones



▲ **Figure 3.7** Some Scatterplots and Their Correlations. The correlation gets closer to ±1 when the data points fall closer to a straight line. **Question** Why are the cases in which the data points are closer to a straight line considered to represent stronger association?

Correlaciones

Se obtiene el valor Z de cada observación en X y en Y.

Ambos valores se multiplican y luego se suman. El resultado se divide por el tamaño de la muestra.

$$r = \frac{1}{n-1}\sum(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})$$

Valores Z para Brasil

```
brazil <- paises2007[15,c(1,4,6)]
brazil %>%
  mutate(z_lifeExp=(lifeExp-mean(paises2007$lifeExp))/sd(paises2007$life
   z_gdpPercap=(gdpPercap-mean(paises2007$gdpPercap))/sd(paises2007$gdpP
```

country	lifeExp	gdpPercap	z_lifeExp	z_gdpPercap
Brazil	72.39	9065.801	0.4458352	-0.203288

Correlaciones

En R ya escribieron la función:

```
cor(paises2007$lifeExp,paises2007$gdpPercap)
```

```
## [1] 0.6786624
```

A mano algo así:

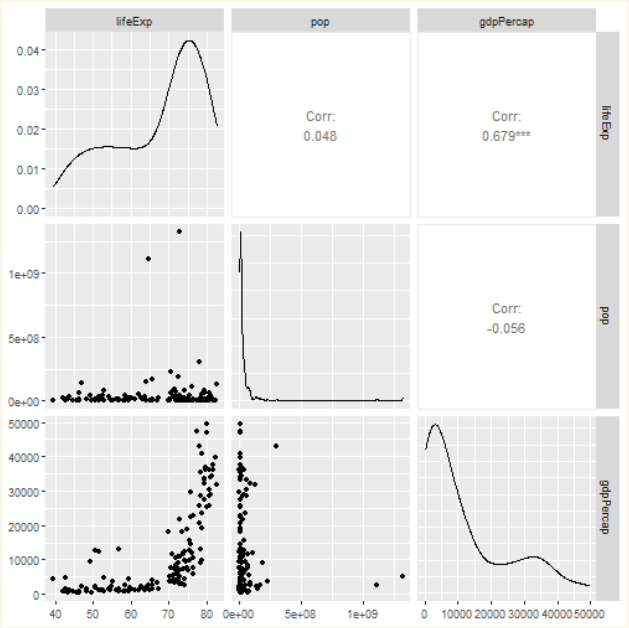
```
zeta <- paises2007 %>%
  select(country,lifeExp,gdpPercap) %>%
  mutate(
    meanlifeExp=mean(lifeExp),
    meangdpPercap=mean(gdpPercap),
    sdlifeExp=sd(lifeExp),
     sdgdpPercap=sd(gdpPercap),
    z_lifeExp=(lifeExp-mean(paises2007$lifeExp))/sd(paises2007$lifeExp),
   z_gdpPercap=(gdpPercap-mean(paises2007$gdpPercap))/sd(paises2007$gdpP

sum(zeta$z_gdpPercap*zeta$z_lifeExp)/(nrow(zeta)-1)
```

```
## [1] 0.6786624
```

Correlaciones

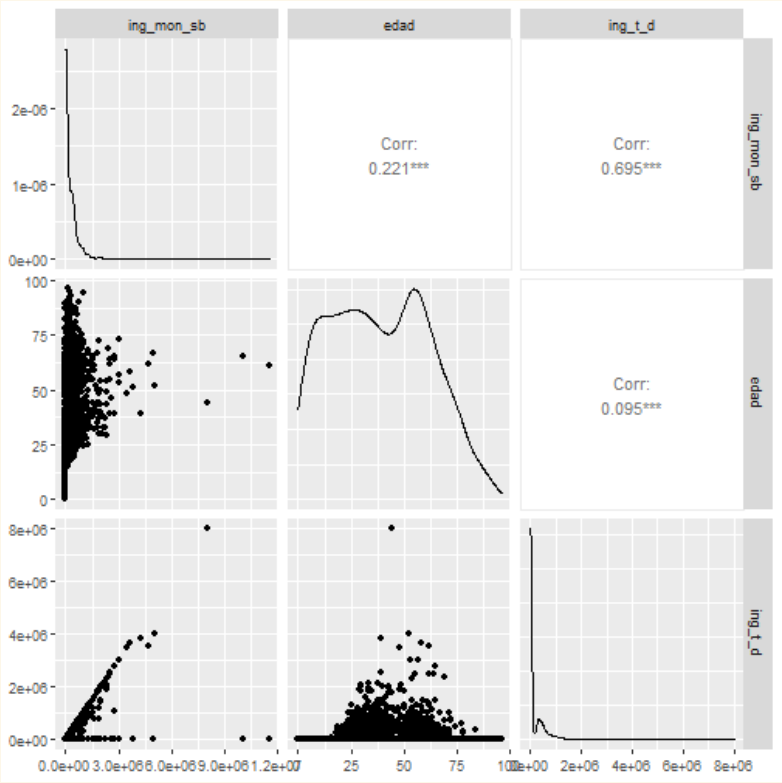
También hay muchos paquetes que permiten visualizar las correlaciones

```
library(GGally)
ggpairs(paises2007[,c("lifeExp","pop","gdpPercap")])
```



Correlaciones

```
ggpairs(esi[,c("ing_mon_sb","edad","ing_t_d")])
```



Correlaciones

La correlación siempre cae entre 1 y -1

Mientras más cercana sea al valor absoluto de 1 más fuerte será la asociación lineal

Correlación positiva indica asociación positiva. Correlación negativa indica asociación negativa.

El valor de la correlación se encuentra estandarizado (se calcula desde valores z). Por tanto no depende de la unidad de medida de las variables

La correlación no depende cuál variable es considerada como de respuesta o explicativa.

Correlación entre 0.3 y 0.5 es de fuerza media, aceptable en ciencias sociales.

Correlación de 0.6 o más excelente (de 1 imposible)

El valor de r^2 es la proporción de la variación de y que está explicada por la relación lineal entre x y y.

Conclusiones

Existen distintos estadísticos y visualizaciones para conocer la asociación entre variables (dependen del tipo de variable)

Por muy fuerte que sea una asociación, **correlación no implica causalidad**

Puede existir una relación entre x y y, aun cuando no haya una correlación lineal

Los análisis realizados son solamente representativos del conjunto de datos

Hasta ahora no hemos realizado ninguna inferencia hacia la población

En clase subsiguiente comenzaremos con muestras e inferencia, para así poder sacar conclusiones sobre la población chilena o latinoamericana desde los datos.

Para esto hay que considerar el cómo se seleccionó la muestra, el peso de cada caso (factor de expansión) y los niveles de error asociados a las estimaciones.

Recursos web utilizados

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

Para seguir aprendiendo

Boxplot por grupo

Bibliografía utilizada

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

Triola, M. F. (2009). *Estad\'istica. Décima Edición, Editorial Pearson Educación*. México, DF: Pearson.