45:00

Estadística Aplicada y Procesamiento de Datos con R

Código en:

Clase 6. Estadística bivariada y Markdown

27 de septiembre, 2024

Andrés González Santa Cruz

andres.gonzalezs@mail.udp.cl

José Ruiz-Tagle Maturana

jose.ruiz-tagle@mail.udp.cl

udp Unidad de Postgrados

FACULTAD DE CIENCIAS SOCIALES E HISTORIA

Introducción

- Datos bivariados : "Son los valores de 2 variables diferentes que se obtienen del mismo elemento poblacional" (Johnson & Kuby, 2008, p. 146)
- Veremos algunos ejemplos aplicados de lo que quieren ver: estadística bivariada

▶ código

Chi-cuadrado

- Supongamos que una tasa alarmante de violaciones en USA se definió en 11 por 100,000 habitantes.
- ¿Existe una asociación entre una alta tasa de asesinatos y violaciones?

$$x^2 = \sum \frac{(O-E)^2}{E}$$

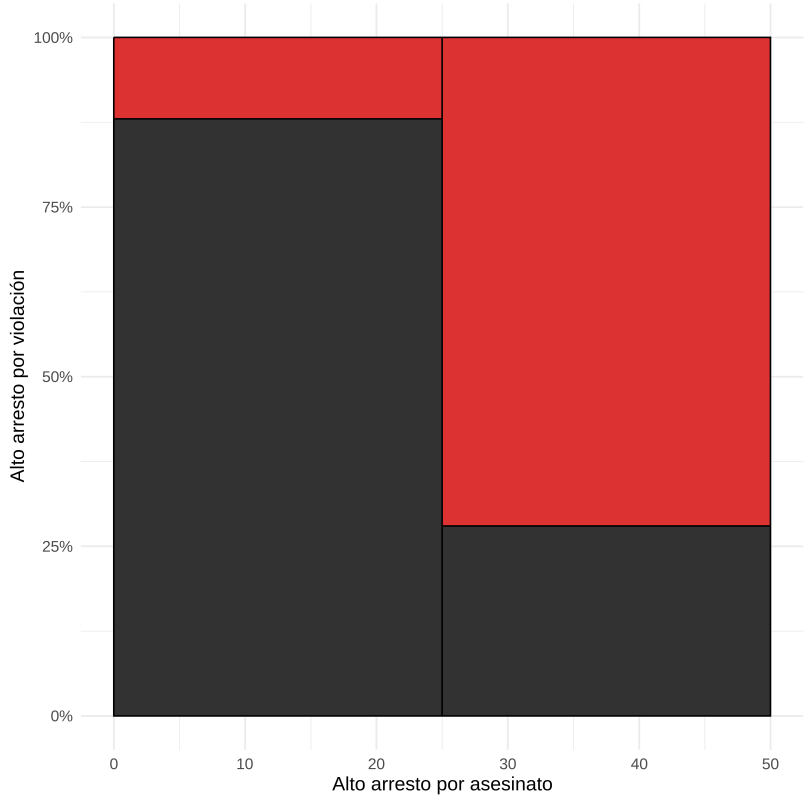
$$gl = (r-1)(e-1)$$

Sea:

- O= frecuencia observada de una casilla, E= Frecuencia esperada de una casilla
- gl= r= número de filas en la tabla cruzada / c= número de columnas en la tabla cruzada

▶ código

▶ código



▶ código

##

Odds ratio

▶ código

	Evento Y=1	Evento Y=0
x=1	a	c
x=0	b	d

- **Probabilidades**: La razón se incluye en el denominador

$$\frac{p}{1}$$

$$\frac{a}{a+c}$$

- **Chances/momios**:

$$\frac{p}{1-p}$$

$$\frac{a}{c}$$

$$\frac{\text{n de éxitos}}{\text{n de fracasos}}$$

- **Odds Ratio**: Las chances de producirse el resultado esperado (outcome) en expuestos a una VI, en relación a las chances de experimentar el mismo resultado en no expuestos a una VI.

$$OR = \frac{a \times d}{b \times c} = \frac{\frac{a}{c}}{\frac{b}{d}}$$

$$P(E) = \frac{O(E)}{O(E)+1} = \frac{\frac{a}{c}}{\frac{a}{c}+1}$$

- **Razón de prevalencia (PR)**: Razón de proporciones

$$PR = \frac{P_A(E)}{P_B(E)} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

Sea $P_A(E)$ la probabilidad de que se constate/presente la enfermedad en la condición A

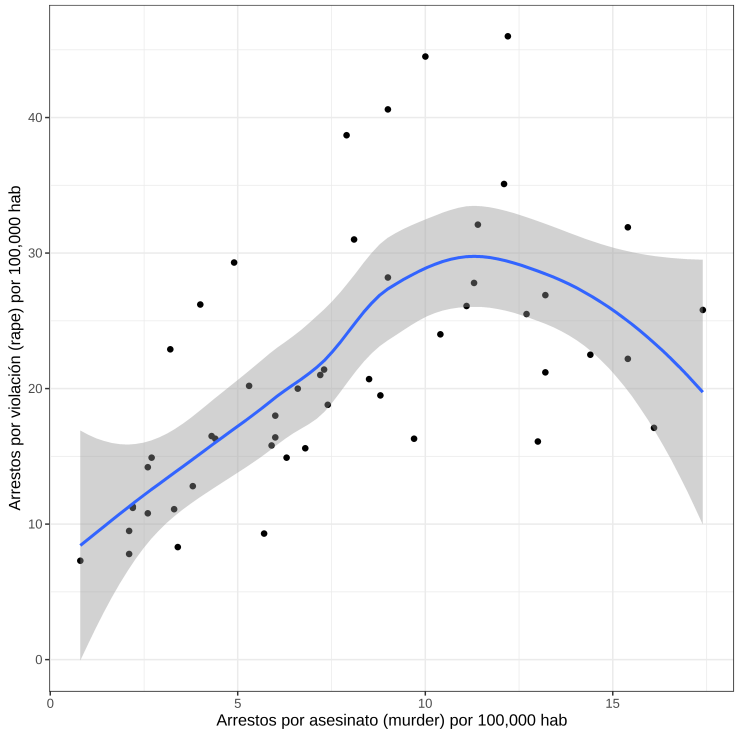
Correlación

45:00

- Correlación lineal : Mide la fuerza de una relación lineal entre dos variables

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

▶ código



Crímenes violentos por Estado en Estados Unidos en 1973

▶ código

```
##
##   Pearson's product-moment correlation
##
## data:  USArrests$Murder and USArrests$Rape
## t = 4.7267, df = 48, p-value = 2.031e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3383006 0.7277619
## sample estimates:
##      cor
## 0.5635788
##
##
##   Spearman's rank correlation rho
##
## data:  USArrests$Murder and USArrests$Rape
## S = 6675.9, p-value = 5.8e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.6794265
##
##
##   Spearman's rank correlation rho
##
## data:  USArrests$Murder and USArrests$Rape
## S = 6675.9, p-value = 5.8e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.6794265
##
##
##   Kendall's rank correlation tau
##
## data:  USArrests$Murder and USArrests$Murder
## z = 10.193, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
## tau
##   1
```

▶ código

correlación no implica causalidad (si no vea aquí)

Introducción a Rmarkdown y Quarto

$Fuente: Prof.\ Matias\ Placencio - Castro\ (placenci@bc.edu)$

**Torpedo sobre Rmarkdown*

- Markdown: Lenguaje de marcado ligero para formatear texto.
- Creador: John Gruber, 2004
- Uno de los lenguajes de formateo más populares.
- Rmarkdown: Extensión de Markdown para integrar código R y otros lenguajes.
- Usos:
  - Comunicar resultados para toma de decisiones.
  - Colaboración en investigaciones, incluyendo pasos seguidos.
  - Ambiente de trabajo moderno que captura procesos y flujo de trabajo
- Quarto: Extensión avanzada de Markdown.
- Integración fácil con Python y otros lenguajes.
- Mejora en el procesamiento de Markdown en "trabajos".

Ventajas y desventajas

- Todo en un mismo lugar
- Automatiza
- Facilita colaboración (claridad en pasos, rastreables, etc.)

- Difícil lectura (mucho código)
- No es tan rapido manejarlo
- Difícil aprender al principio

Apuntes

- Apuntes Quarto
- Apuntes Rmarkdown

Estructura (1): YAML

- Debe ir al principio de cada documento y entre "---" tanto al principio como al final de esta sección
- Respete la indentación (ej., espacios como " " para definir jerarquías de los argumentos del YAML)
- Metadatos: Hablan de la estructura del documento (ej., nombre, título). Cuando usted saca una foto, ésta tiene una fecha, el nombre de la cámara, información sobre la profundidad de la imagen, la resolución etc. Esta es similar.

```
---
title: "Untitled" # sección de título
author: "ags" # autor
date: '2023-08-15' # fecha
output: html_document # formato de salida del documento
---
```

- Quarto

```
---
title: "Untitled" # sección de título
format: html # formato de salida del documento
editor: visual # interfaz gráfica que permite editar documentos sin tener que saber markdown
---
```

- PDF

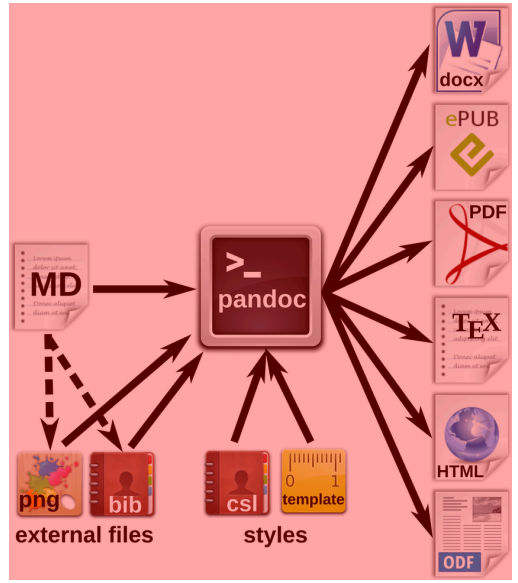
```
---
title: "Untitled"
author: "ags"
date: '2023-08-15'
output: pdf_document # para exportar a pdf
---
```

- PDF (xelatex)
- Si el anterior metadato para exportar a PDF no le funciona, puede instalar 'tinytex'

```
---
title: "Untitled"
author: "ags"
date: '2023-08-15'
output: # tenga en cuenta el indentado, la separación jerárquica de cada línea
  pdf_document:
    latex_engine: xelatex
---
```

Estructura (2): Chunks

Argumento	Ejemplo (por defecto)	Función
eval	eval=TRUE	El código corre y los resultados se incluyen en la salida.
include	include=TRUE	¿Se incluye el código y el resultado en la salida?
echo	echo=TRUE	¿Se despliega el código junto con los resultados?
warning	warning=TRUE	¿Se despliegan los mensajes de advertencia?
error	error=FALSE	¿Se despliegan los errores?, ¿sigue la compilación si hay errores?
message	message=TRUE	¿Se despliegan los mensajes?
tidy	tidy=FALSE	¿Se formatea el código para que parezca “limpio”?
results	results="markup"	"Cómo se ven los resultados? ""hide"" = sin resultados ""asis"" = resultados sin formato ""hold"" = se compilan los resultados al final del chunk (usar si hay muchos comandos)"
cache	cache=FALSE	¿Se guarda en el cache para compilaciones futuras?
comment	comment="##"	¿Cuál es el signo en que los caracteres no se evalúan?
fig.width, fig.height	fig.width=7	¿Cuál es el ancho y largo (en pies) de la figura?
fig.cap	fig.cap=""	Título de la figura
fig.align	fig.align="left"	Ubicación de la imagen: (izquierda) "left" (derecha) "right" (centro) "center"



Fuente: https://ourcodingclub.github.io/tutorials/rmarkdown/

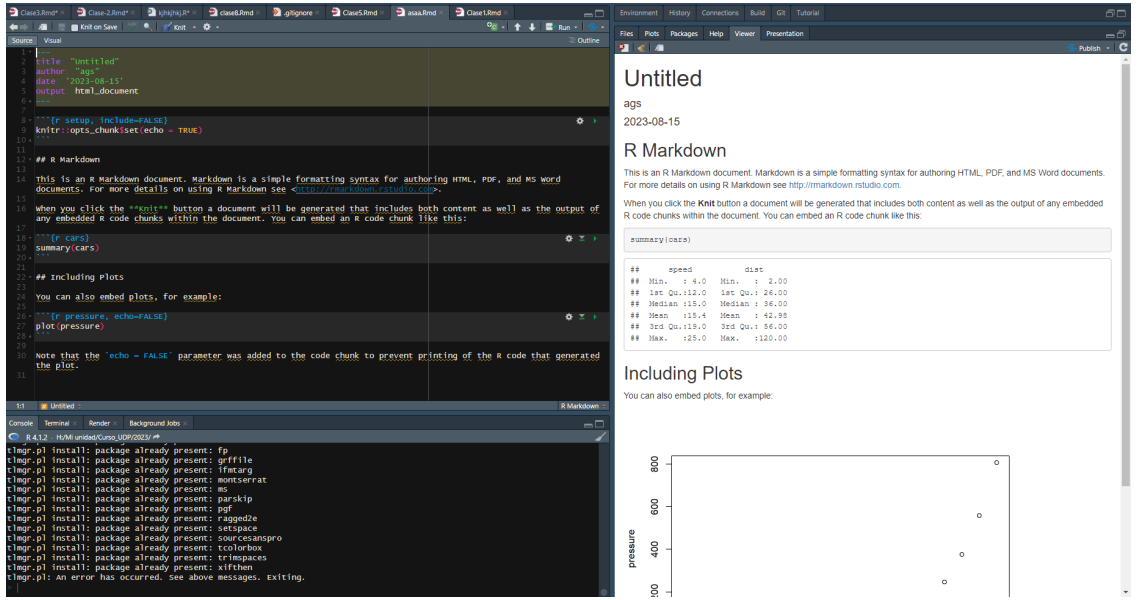
Nota: En quarto, los argumentos de cada chunk se presentan en formato #| debajo de cada chunk

Ejemplos

45:00

```
```{r, "rmarkdown-output", echo = F, eval=TRUE, out.width = '50%', fig.align = 'center', fig.cap="Ejemplo markdown"}
url <- './_figs/rmarkdown.PNG'
knitr::include_graphics(url)
```
```

Salida:



Ejemplo markdown

- La media de 5 números es 18. Si uno es excluido, la media es 16. ¿Cuál es el número excluido?

```
$$
\frac{(a + b+ c+ d+ x)}5=18
$$

$$
\frac{(a + b+ c+ d)}4=16
$$

$$
```

10 01:29

Ejemplos (2)

Necesario para hacer la interfaz con Python:

if(!require(reticulate)){install.packages("reticulate")}

```
```{python, "ejercicio-python", echo = F}
import numpy as np
np.random.seed(42)
data = np.random.randn(100)
mean_data = np.mean(data)
mean_data
```
```{r,"ejercicio-r", echo = F}
set.seed(42)
r_data <- rnorm(100)
mean_r_data <- mean(r_data)
mean_r_data
```
```

Salida:

0

-0.10384651739409384

[1] 0.03251482

Ejercicio 1

Despliegue el siguiente ejercicio en un markdown en formato .html

- De algunas de las base de datos de permisos de circulación pagados y tramitados en la Municipalidad de Cochamó el 2016 (https://datos.gob.cl/dataset/permisoscirculacion2016cochamo),
    - Obtenga el porcentaje por columna, según corresponda al tipo de variable y nivel de medición.
    - Obtenga la media y la mediana, según corresponda al tipo de variable y nivel de medición.
    - EXTRA: Obtenga una tabla de 2 vías, según corresponda al tipo de variable y nivel de medición.

Ejercicio 2

De los datos sobre interrupción voluntaria del embarazo, genere un gráfico de lineas en que el eje x sea el AÑO y las líneas sean la frecuencia. Cada línea debe representar cada causal (rojo= Causal 1: Peligro para la vida de la mujer; azul=Causal 2: Inviabilidad fetal de carácter letal; morado= Causal 3: Embarazo por violación), utilizando tidyverse

```
#https://deis.minsal.cl/#tableros
#notese, que no escribimos con ñ por notación
data_df <- data.frame(
 ANIO = c(2018, 2018, 2018, 2019, 2019, 2019, 2020, 2020, 2020, 2021, 2021, 2021, 2022, 2022, 2022, 2023, 2023, 2023),
 Frecuencia = c(262, 346, 124, 267, 414, 137, 160, 348, 154, 250, 442, 130, 254, 368, 209, 103, 162, 142),
 CAUSAL = c("Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3",
        "Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3", "Causal 1", "Causal 2", "Causal 3")

)

library(tidyverse)

#Ejemplo de un gráfico con el total, sin división en causales
data_df %>%
  group_by(ANIO) %>%
  summarise(total=sum(Frecuencia, na.rm=T)) %>%
  ggplot(aes(ANIO, total, group=1))+ #en algunos casos podría solicitarle el último argumento
  geom_point()+ #líneas
 theme_minimal() #temática
```

Fuentes

- Aron, A. & Aron, E. (2002). Estadística Para Psicología. Brasil. Prentice-Hall.
- Ayçaguer, L. (1997). Cultura Estadística e Investigación Científica en el Campo de la Salud: Una Mirada Crítica. Díaz de Santos.
- Glen, S. (s.f). "Bivariate Analysis Definition & Example" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/bivariate-analysis/
- Johnson, R. & Kuby, P. (2008). Estadística Elemental: lo esencial México: Cengage Learning.
- Montero-Alonso, M.A. (2007, Mayo). Apuntes de Estadística II. Tema 4. Universidad de Granada. Obtenido el 19 de Octubre de 2021 desde: https://www.ugr.es/~eues/webgrupo/Docencia/MonteroAlonso/estadisticaII/tema4.pdf. ISBN: 9788469056639
- Ritchey, F. (2008). Estadística para las Ciencias Sociales. México: McGrawll Hill.
- Ritchie, H. & Roser, M. (2020). "Energy". Publicado en línea en OurWorldInData.org. Obtenido el 16 de Octubre de 2021 desde: 'https://ourworldindata.org/energy' [Recurso en-línea]
- s.a. Datos de países, Índices de Precios de Consumo (IPC). Datosmacro. Obtenido el 16 de Octubre de 2021 desde: https://datosmacro.expansion.com/ipc-paises
- Sandilands D. (2014) Bivariate Analysis. En: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_222
- Santibáñez, J. (2021). Conceptos básicos de la inferencia estadística. [Material de clases].Unidad 1. Introducción. Javier Santibáñez. IIMAS, UNAM jsantibanez@sigma.iimas.unam.mx. Semestre 2020-1. Obtenido el 18 de Octubre de 2021 desde: http://sigma.iimas.unam.mx/jsantibanez/Cursos/Inferencia/2020_1/notas/u1.pdf
- Starmer, J. "Hypothesis testing and the null hypothesis, clearly explained!!!". Youtube. Obtenido el 17 de Octubre de 2021 desde: https://youtu.be/0oc49DyA3hU