

INFERENCIA ESTADÍSTICA

Maestría en estadística aplicada

Universidad de Nariño

Material preparado por:

Giovany Babativa, PhD

PRESENTACIÓN DEL CURSO

- **Sobre mí**
- **Objetivo**
- **Metodología**
- **Recursos y materiales**
- **Evaluación**



SOBRE MÍ



<http://jgbabativam.rbind.io/>



<https://scholar.google.es/citations?user=2NJRN8A8AAAJ&hl=es>



<https://github.com/jgbabativam>



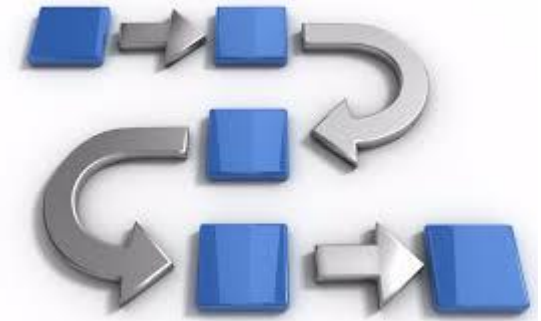
jgbabativam@unal.edu.co

OBJETIVO

Continuar con la formación global de los alumnos sobre las principales técnicas estadísticas orientadas a la experimentación e investigación científica, dotándolos de los conocimientos necesarios para utilizar de manera inteligente la metodología disponible para el análisis y la evaluación de la información proveniente de cualquiera de las ramas del saber.



METODOLOGÍA

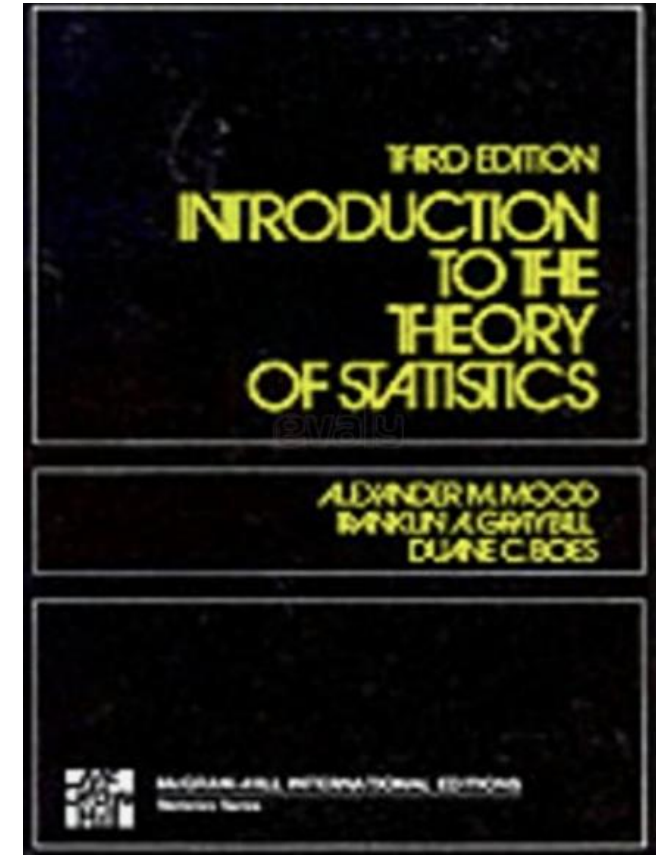
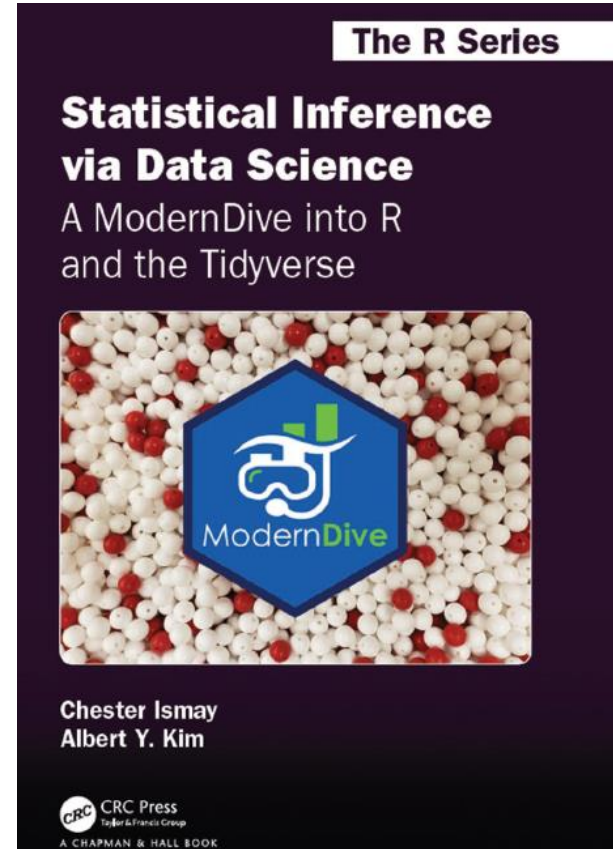
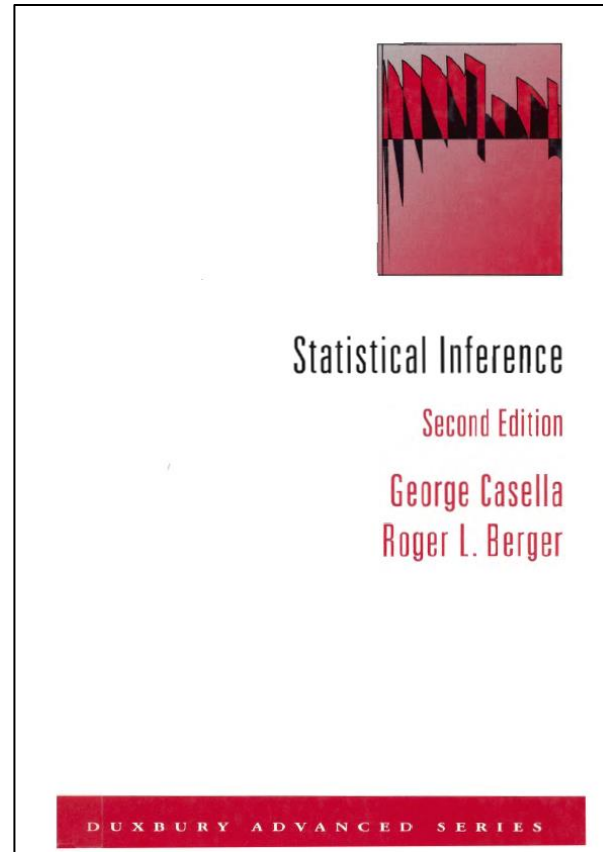
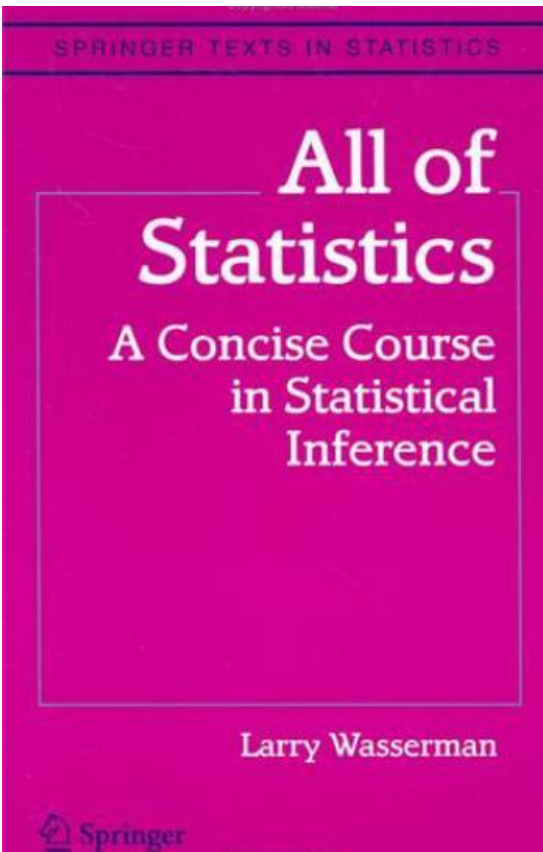


Clases magistrales teórico - prácticas. Se desarrollan los elementos teóricos, conceptuales y aplicados de cada tema. En cualquier momento se le puede pedir a cualquiera de los alumnos que comparta la pantalla y realice algún ejercicio del tema que se esté desarrollando



RECURSOS Y MATERIALES

- Los materiales y ejercicios se dejan en moodle cada semana.

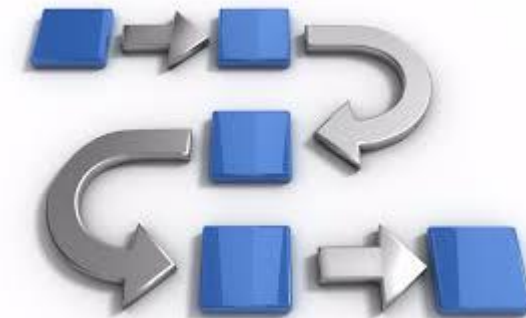


EVALUACIÓN

- 5 Exámenes con un peso del 20% cada uno.
- Se realiza un examen cada semana con el tema visto la semana anterior. Para el último examen se tendrá una franja que no afecte el desarrollo de la siguiente asignatura.



SESIONES

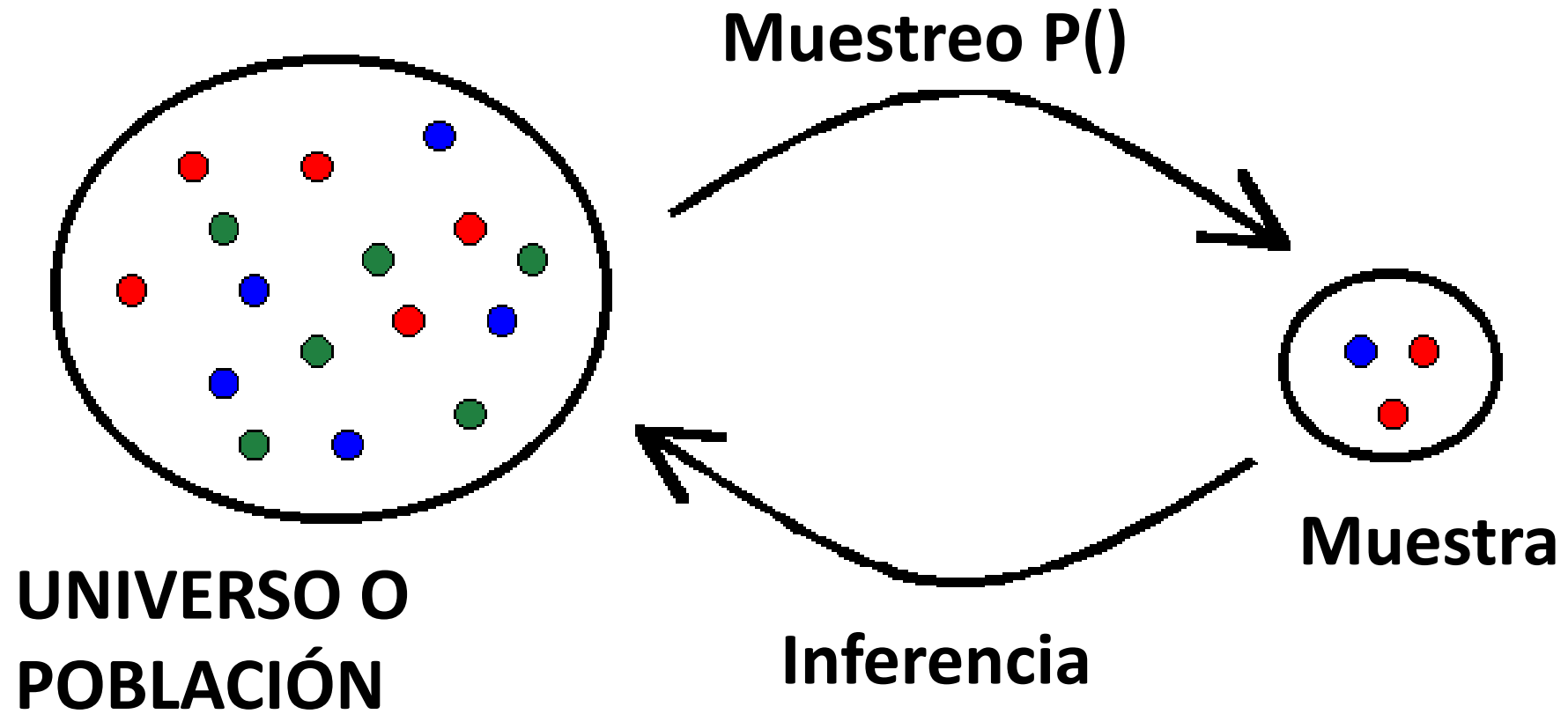


Viernes	Bloque 1	6:00 p.m.	7:00 p.m.	dudas	1		
	Bloque 2	7:00 p.m.	8:30 p.m.	examen	1,5		
	Bloque 3	8:30 p.m.	10:30 p.m.	clase	2		
Sabado	Bloque 4	7:00 a.m.	9:00 a.m.	clase	2	break	30 min.
	Bloque 5	9:00 a.m.	11:00 a.m.	clase	2	break	15 min
	Bloque 6	11:00 a.m.	1:00 p.m.	clase	2	break	15 min
					10,5		

Inicio: 23 de junio 2023

Fin: 22 de julio 2023*

INTRODUCCIÓN



REPASO

Variable aleatoria: $X : \Omega \rightarrow \mathbb{R}$

Función de distribución: $F_X(x) = \mathbb{P}(X \leq x)$.

Valor esperado: $\mathbb{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$

Varianza: $\sigma^2 = \mathbb{E}(X - \mu)^2$

DISTRIBUCIONES CONTINUAS

$X \sim \text{Uniform}(a, b)$, if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}, \quad x \in \mathbb{R}$$

$X \sim \text{Exp}(\beta)$, if

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0$$

$X \sim \text{Gamma}(\alpha, \beta)$, if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

DISTRIBUCIONES CONTINUAS

THE χ^2 DISTRIBUTION. X has a χ^2 distribution with p degrees of freedom — written $X \sim \chi_p^2$ — if

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad x > 0.$$

If Z_1, \dots, Z_p are independent standard Normal random variables then $\sum_{i=1}^p Z_i^2 \sim \chi_p^2$.

DISTRIBUCIONES CONTINUAS

t AND CAUCHY DISTRIBUTION. X has a t distribution with ν degrees of freedom — written $X \sim t_\nu$ — if

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$$

The t distribution is similar to a Normal but it has thicker tails. In fact, the Normal corresponds to a t with $\nu = \infty$. The Cauchy distribution is a special case of the t distribution corresponding to $\nu = 1$. The density is

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

PROPIEDADES DE LA DISTR. NORMAL

(i) If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

(ii) If $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

(iii) If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$ are independent, then

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

It follows from (i) that if $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned}\mathbb{P}(a < X < b) &= \mathbb{P}\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).\end{aligned}$$

PAQUETE ESTADÍSTICO



...pero ahora es del tipo...



DIFERENCIA ENTRE R y R-STUDIO





ENTORNO DE R-STUDIO

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

1. Script GDSC.R x Untitled5* x Untitled6* x 2. Funnel.R* x Calibración de universos.R* x Untitled7* x >>

Source on Save Run Source

marcas Next Prev All Replace Replace All

☐ In selection ☐ Match case ☐ Whole word ☐ Regex ☒ Wrap

```
182 dev.off()
183 |
184
185 f7 = Datos %>%
186     dplyr::filter(Ciudad == "TOTAL") %>%
187     ggplot(aes(reorder(var, Indicador), fill=Marca)) +
188     geom_bar(aes(y = Indicador), stat = "identity", width = 1)
189     geom_bar(aes(y = - Indicador), stat = "identity", width = 1)
190     geom_text(aes(y=0, label= paste(round(Indicador), '%')),
191               color='black') +
192
```

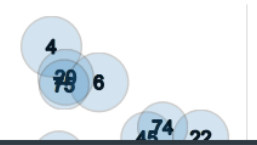
183:1 (Top Level) R Script

Files Plots Packages Help Viewer

Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust re $\lambda = 1$

Intertopic Distance Map (via multidimensional scaling)



PC2

0 200

new state name clue

Console Terminal x

G:/Mi unidad/JUGISAKA/Proyectos/2019/8. Agosto/5. Muestra Maestra/Expansión/

```
[946] 44 44 44 44 44 44 44
[953] 44 44 44 44 44 44 44
[960] 44 44 44 44 44 44 44
[967] 44 44 43 43 43 43 43
[974] 43 43 43 43 43 43 43
[981] 43 43 43 43 43 43 43
[988] 43 43 43 43 43 43 43
[995] 43 43 42 42 42 42 42
[ reached getOption("max.print") -- omitted 339
3 entries ]
> |
```

Environment History Connections

Import Dataset

Global Environment

Data

- Jeopardy Large list (5 elements, 19.3 Mb)

Values

- json Large character (723.9 Kb)

R y R-MARKDOWN





EJERCICIO

1. Suppose that $X \sim N(3, 5)$. Find $\mathbb{P}(X > 1)$.
2. Now find $q = \Phi^{-1}(0.2)$. This means we have to find q such that $\mathbb{P}(X < q) = 0.2$.



EJERCICIO

Un agente de seguros vende pólizas a cinco personas de la misma edad y que disfrutan de buena salud. Según las tablas actuales, la probabilidad de que una persona en estas condiciones viva 30 años o más es de $2/3$. Calcule la probabilidad de que transcurridos 30 años, vivan:

1. Las 5 personas
2. Al menos 3 personas, $P(X \geq 3) = 1 - P(X < 3)$

3. Use R para calcular:

Si $p = 0.4$ y $n = 50$, calcule $P(X = 15)$.

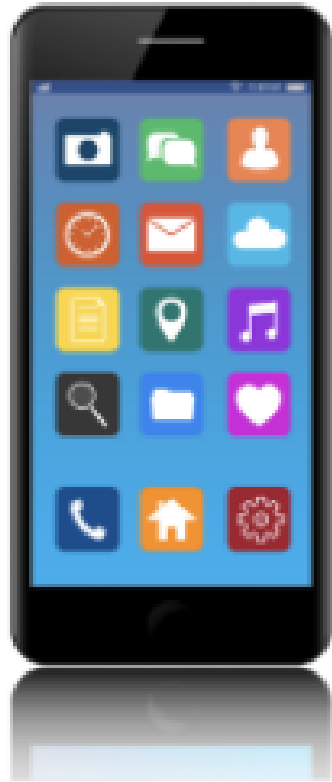
Si $p = 0.4$ y $n = 50$, calcule $P(X \leq 15)$.

Si $p = 0.4$ y $n = 50$, calcule $P(10 \leq X \leq 15)$.

Si $p = 0.4$ y $n = 50$, calcule q tal que $P(X \leq q) = 0.561$.

¿CÓMO SE TRABAJA EN R?

R: Nuevo teléfono



Paquetes: Aplicaciones que se pueden descargar



¿CÓMO SE TRABAJA EN R?

```
install.packages("packagename")
```

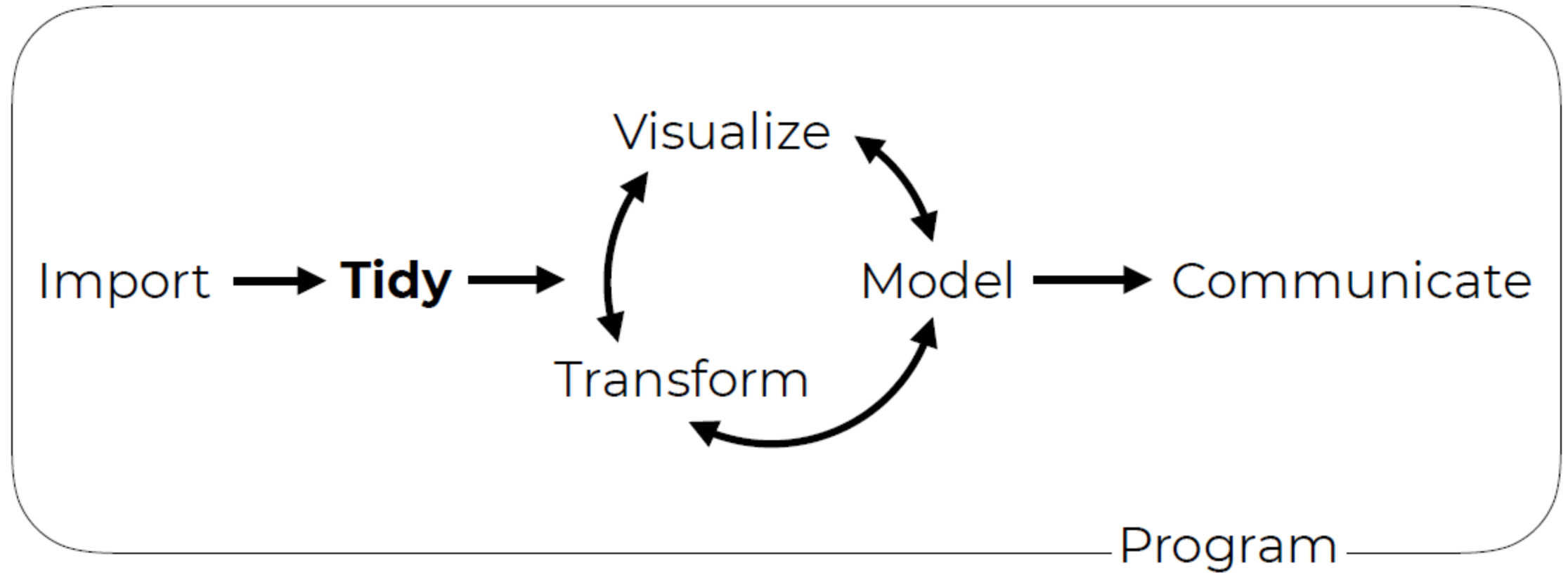
```
library(packagename)
```

BiplotML
tidyverse
readxl
skimr

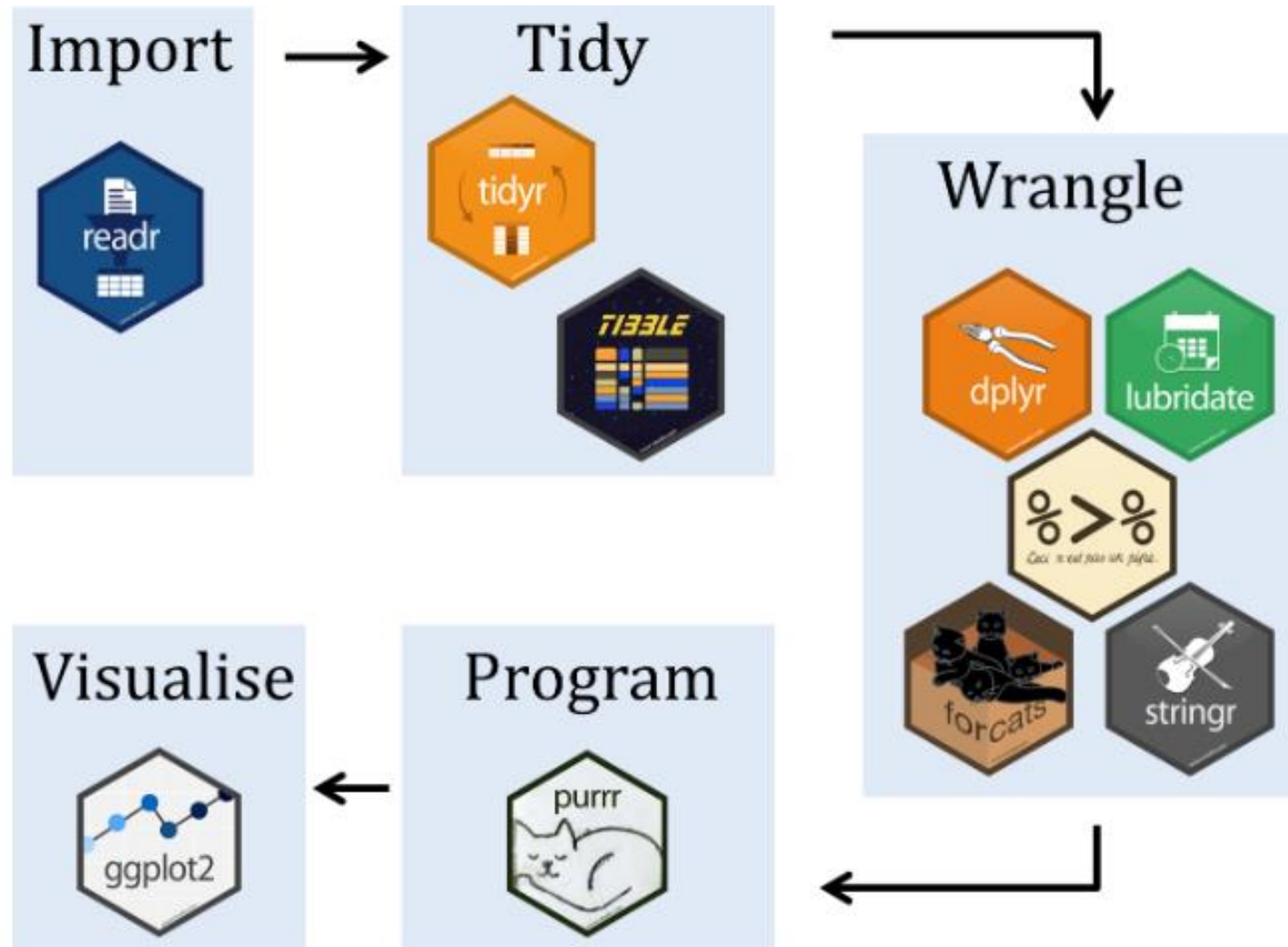
citation("package")



CICLO DE ANALÍTICA DE DATOS



EL MUNDO TIDYVERSE



EL MUNDO TIDYVERSE



```
library(tidyverse)
```



```
library(readr)  
library(dplyr)  
library(tidyr)  
library(ggplot2)  
library(purrr)  
library(tibble)  
library(stringr)  
library(forcats)
```

EXPLORACIÓN DE DATOS



Advertencia: Asegúrese de haber instalado los paquetes.

```
library(tidyverse)
library(nycflights13)
library(knitr)
```

```
x %>%
  f() %>%
  g() %>%
  h()
```

Cargue el conjunto de datos `flights`, revise la ayuda usando `?flights` y explore

- Use la función `View()` del entorno `R`.
- Use la función `glimpse()` del entorno `tidyverse`.
- Use la función `kable()` del paquete `knitr` para imprimir los primeros 5 registros del marco de datos.
- Use la función `skim()` del paquete `skimr` sobre la consola. Cargue el conjunto de datos y el paquete `nycflights13`.

REGLAS DE LOS DATOS

country	year	cases	population
Afghanistan	1999	3775	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174603898
China	1999	212258	127291272
China	2000	216766	128042583

variables

country	year	cases	population
Afghanistan	1999	3775	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174603898
China	1999	212258	127291272
China	2000	216766	128042583

observations

country	year	cases	population
Afghanistan	99	75	987071
Afghanistan	00	66	595360
Brazil	99	737	006362
Brazil	00	488	603898
China	99	2258	27291272
China	00	6766	42583

values

ESTRUCTURA DE LOS DATOS

```
head(yourdata)
```

```
str(yourdata)
```

```
length(yourdata)
```

```
glimpse(yourdata)
```

```
names(yourdata)
```

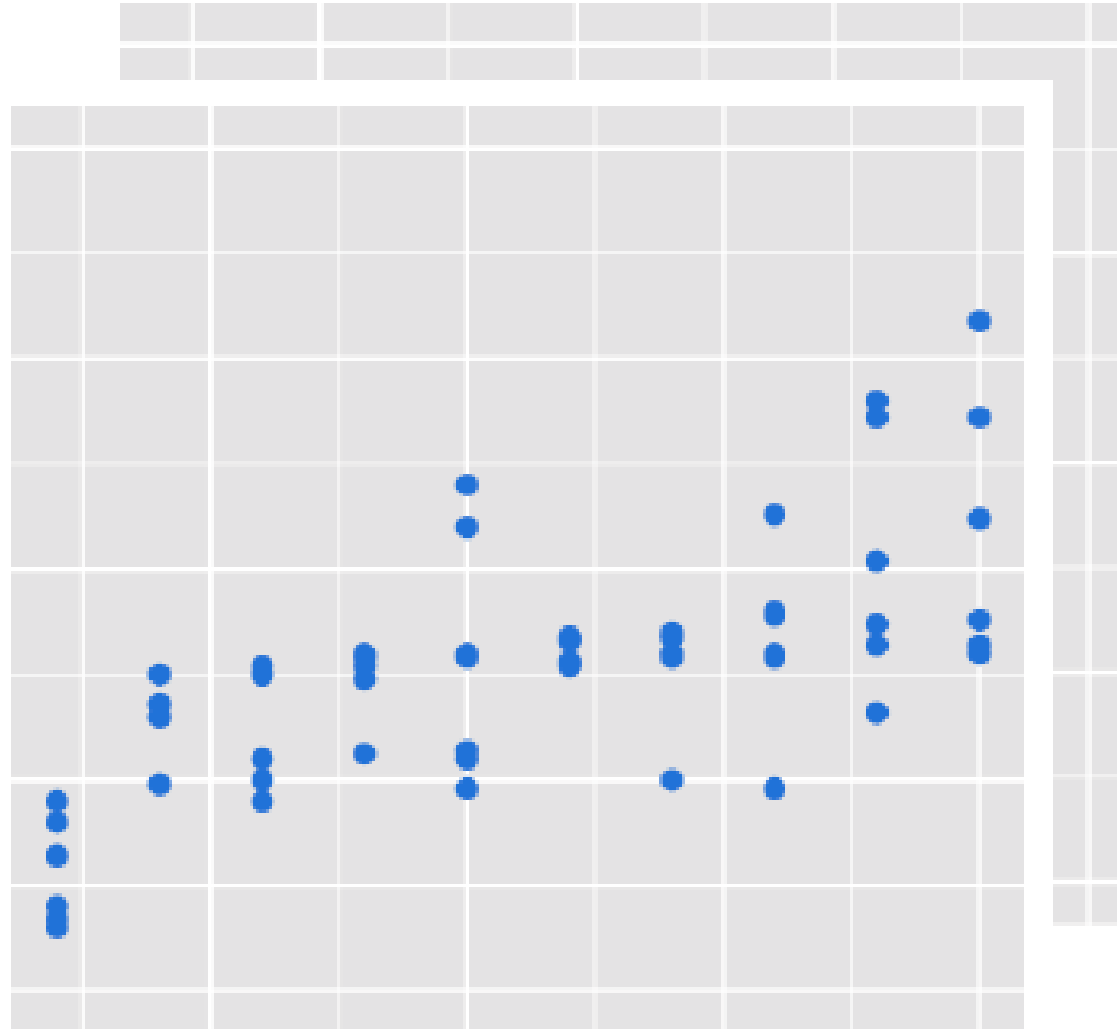

LA GRAMÁTICA DE LAS GRÁFICAS CON ggplot2



CAPAS -- LAYERS



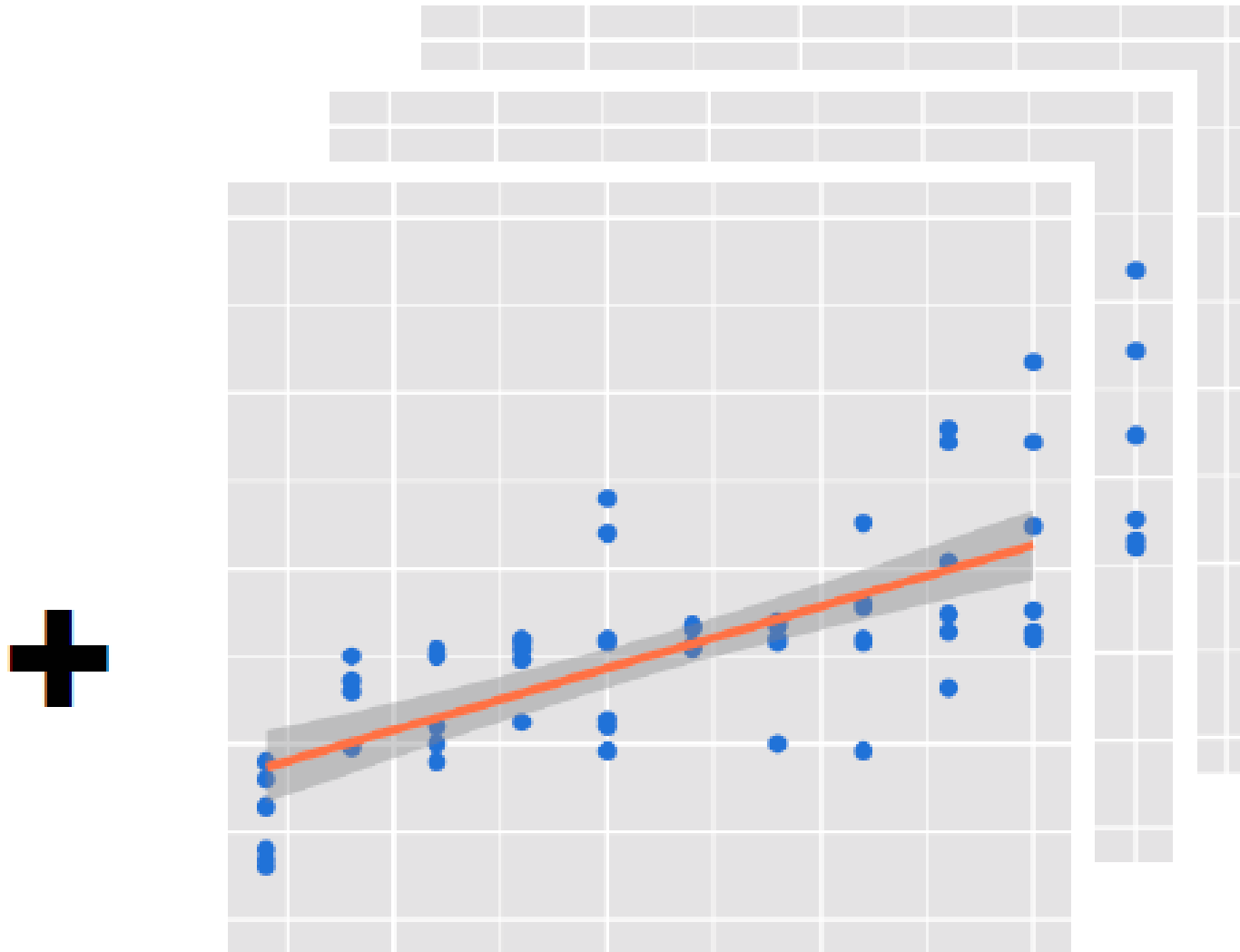
CAPAS -- LAYERS



geometric object



CAPAS -- LAYERS



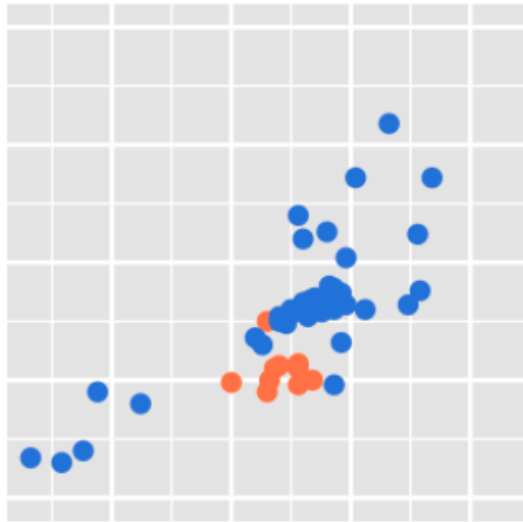
geometric object

Statistical transformation

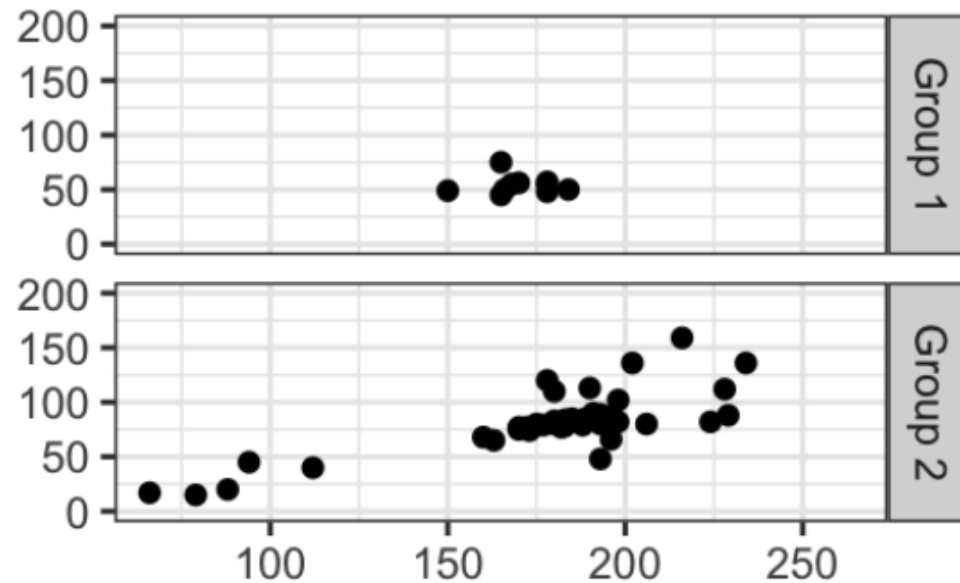
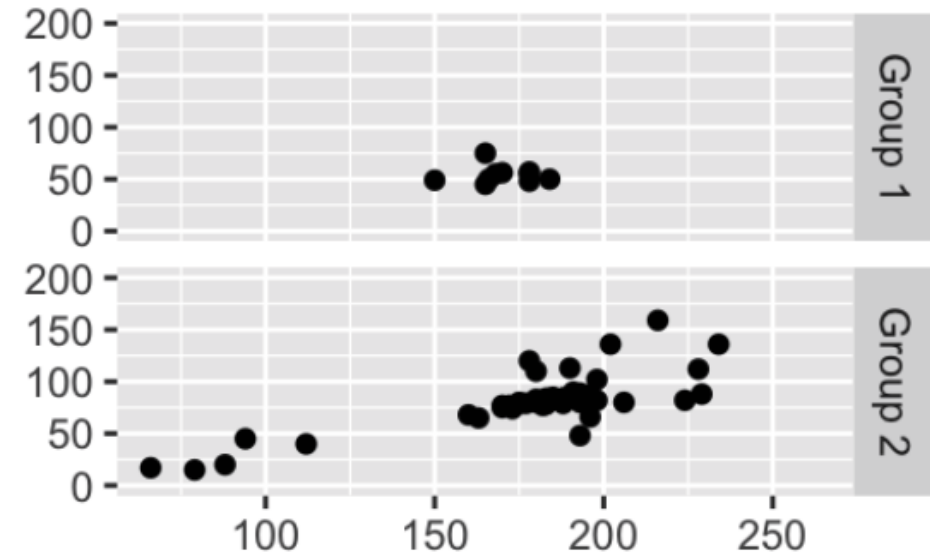
groups + facets + themes



CAPAS -- LAYERS



Groups



groups + facets + themes



LA GRAMÁTICA DE LAS GRÁFICAS CON ggplot2

Se requiere mínimo 3 cosas:

1. Datos
2. Aesthetics: variables
3. geometría

```
ggplot(data, aes(x = __, y = __)) +  
  geom_point()
```



LA GRAMÁTICA DE LAS GRÁFICAS CON ggplot2



```
ggplot(data, aes(x = __, y = __)) +  
  geom_point()
```

```
data %>%  
  filter(interesting_variable > z) %>%  
  ggplot(aes(x = __, y = __, colour = condition))  
  geom_point() +  
  facet_wrap(~ group)  
  
ggsave("your_first_ggplot.png")
```



DATA WRANGLING

filter() - Extraer casos



arrange() - Ordenar casos

group_by() - Agrupar casos



select() - Seleccionar variables

mutate() - Crear nuevas variables

summarize() - Resumir variables o crear casos



country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	666	2035360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	210258	1272015272
China	2000	213036	128043583

variables

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	666	2035360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	210258	1272015272
China	2000	213036	128043583

observations

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	666	2035360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	210258	1272015272
China	2000	213036	128043583

values

MUESTREO

- MAS
- PPT
- Bernoulli
- Estratificado
- Multietápico



DEFINICIÓN IID

Definition 5.1.1 The random variables X_1, \dots, X_n are called a *random sample of size n from the population $f(x)$* if X_1, \dots, X_n are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, X_1, \dots, X_n are called *independent and identically distributed random variables with pdf or pmf $f(x)$* . This is commonly abbreviated to iid random variables.

A partir de lo anterior la función conjunta (joint pdf):

$$f(x_1, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n) = \prod_{i=1}^n f(x_i).$$

Si la fdp está indexada por espacio de parámetros, entonces:

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta),$$

EJEMPLO: Función conjunta IID exponencial

Example (Sample pdf–exponential) Let X_1, \dots, X_n be a random sample from an exponential(β) population. Specifically, X_1, \dots, X_n might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \dots, x_n | \beta) = \prod_{i=1}^n f(x_i | \beta) = \prod_{i=1}^n \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1 + \dots + x_n)/\beta}.$$

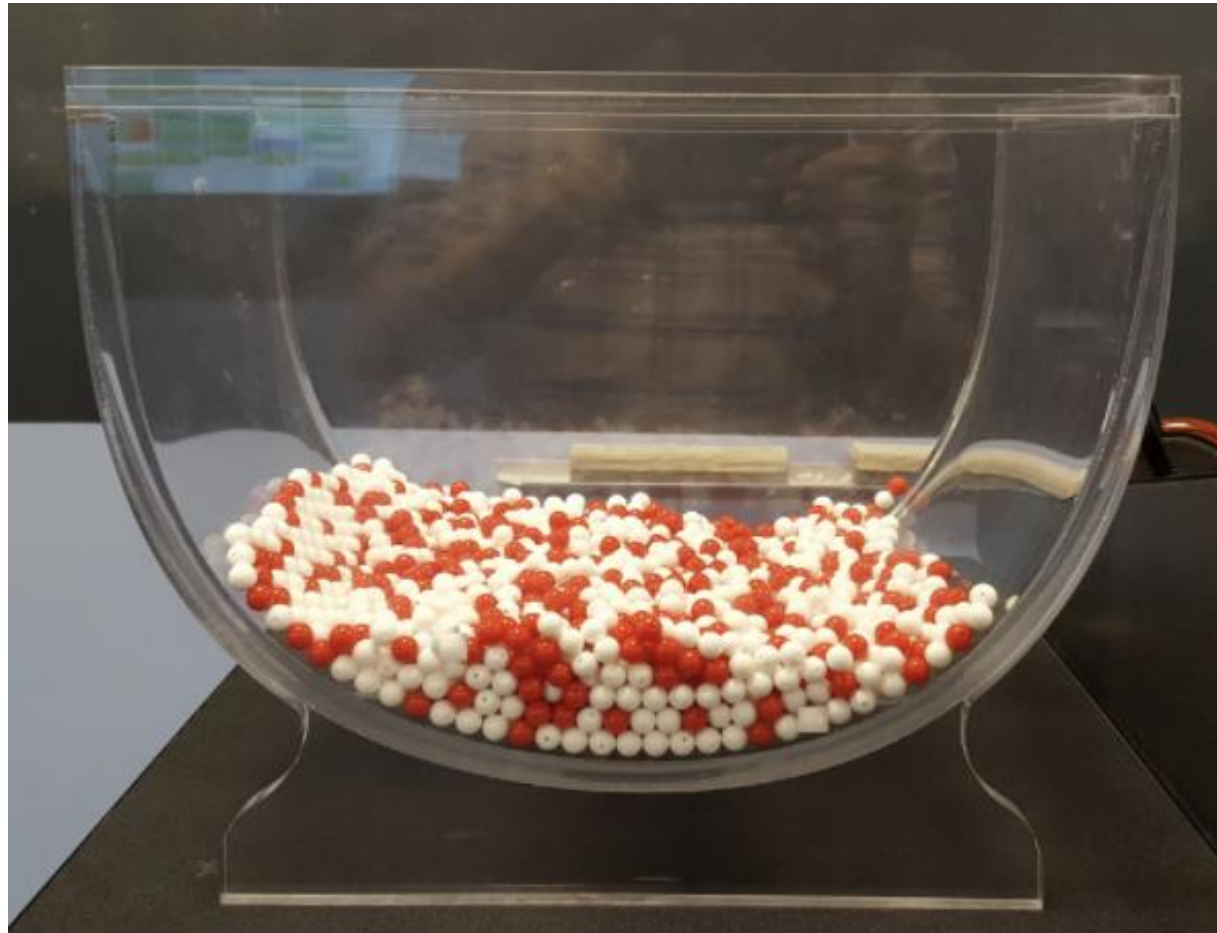
EJERCICIO.

Pruebe que:

- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \frac{\sigma^2}{n}$
- $E(S^2) = \sigma^2$

UNIVERSO, PARÁMETROS Y ESTIMADORES

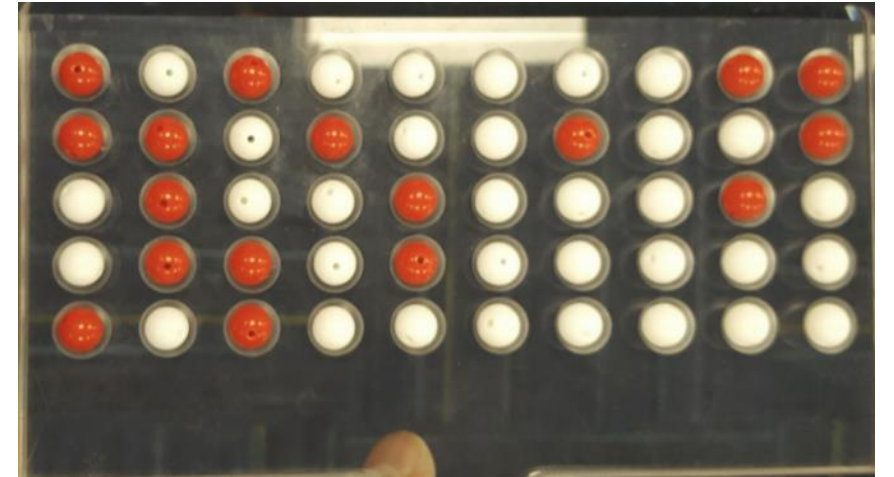
Se tienen bolas rojas y bolas blancas en una taza, todas del mismo tamaño y del mismo peso. Además, se mezclaron de antemano, para evitar patrones en la distribución espacial de las bolas rojas y blancas. ¿Qué proporción de las bolas de este cuenco son rojas?



¿Desempleo?

MUESTRA

Seleccione aleatoriamente $n=50$ bolas



17 de las bolas son rojas, $0.34 = 34\%$.

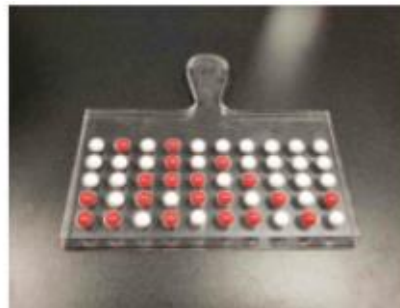
VARIACIÓN MUESTRAL Y TAMAÑO MUESTRAL

- Genere un universe de 2500 bolas con 625 rojas y el resto blancas.
- Seleccione 1000 muestras de tamaño 25, estime la proporción de bolas rojas en cada muestra, represente la distribución muestral del estimador.
- Seleccione 1000 muestras de tamaño 50, estime la proporción de bolas rojas en cada muestra, represente la distribución muestral del estimador.
- Seleccione 1000 muestras de tamaño 100, estime la proporción de bolas rojas en cada muestra, represente la distribución muestral del estimador.

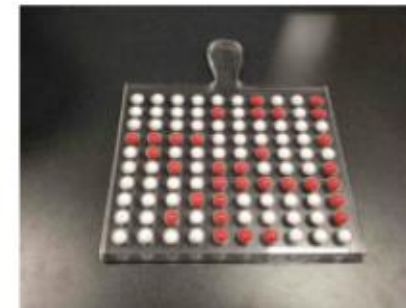
$n = 25$



$n = 50$



$n = 100$



ELEMENTOS RELEVANTES

- Si una muestra se genera al azar y de forma probabilística, entonces la estimación puntual resultante es una "buena aproximación" del verdadero parámetro desconocido en la población.
- En la práctica solo tiene una de las estimaciones y esta no necesariamente coincide con el verdadero valor, de modo que se debe considerar la variación muestral.
- Algunas estimaciones están muy lejanas del verdadero valor.
- Cuando el tamaño muestral aumenta, disminuye la variación muestral y por lo tanto las estimaciones tienden a estar más cerca del valor verdadero (precisión).

CONVERGENCE

1. The law of large numbers says that the sample average $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ converges in probability to the expectation $\mu = \mathbb{E}(X_i)$. This means that \bar{X}_n is close to μ with high probability.
2. The central limit theorem says that $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to a Normal distribution. This means that the sample average has approximately a Normal distribution for large n .

CONVERGENCE

Theorem (The Central Limit Theorem (CLT)). *Let X_1, \dots, X_n be IID with mean μ and variance σ^2 . Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then*

$$Z_n \equiv \frac{\bar{X}_n - \mu}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0, 1)$. In other words,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

TODO ESTO SIGNIFICA LO MISMO

$$Z_n \approx N(0, 1)$$

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X}_n - \mu \approx N\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2)$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx N(0, 1).$$

EJEMPLO:

Example. Suppose that the number of errors per computer program has a Poisson distribution with mean 5. We get 125 programs. Let X_1, \dots, X_{125} be the number of errors in the programs. We want to approximate $\mathbb{P}(\bar{X}_n < 5.5)$. Let $\mu = \mathbb{E}(X_1) = \lambda = 5$ and $\sigma^2 = \mathbb{V}(X_1) = \lambda = 5$. Then,

$$\begin{aligned}\mathbb{P}(\bar{X}_n < 5.5) &= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma}\right) \\ &\approx \mathbb{P}(Z < 2.5) = .9938. \quad \blacksquare\end{aligned}$$

EJEMPLO

El viaje en un autobús especial para ir de un campus de una universidad al campus de otra en una ciudad toma, en promedio, 28 minutos, con una desviación estándar de 5 minutos. En cierta semana un autobús hizo el viaje 40 veces. ¿Cuál es la probabilidad de que el tiempo promedio del viaje sea mayor a 30 minutos?

```
> 1-pnorm(30, mean = 28, sd = 5/sqrt(40))  
[1] 0.005706018
```

DIFERENCIA DE MEDIAS

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

Se distribuye normal estándar

EJEMPLO

Tiempo de secado de pinturas. Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura, el A y el B . Con la pintura tipo A se pintan 18 especímenes y se registra el tiempo (en horas) que cada uno tarda en secar. Lo mismo se hace con la pintura tipo B . Se sabe que la desviación estándar de población de ambas es 1.0.

Si se supone que los especímenes pintados se secan en el mismo tiempo medio con los dos tipos de pintura, calcule $P(\bar{X}_A - \bar{X}_B > 1.0)$, donde \bar{X}_A y \bar{X}_B son los tiempos promedio de secado para muestras de tamaño $n_A = n_B = 18$.

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0$$

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

```
> 1-pnorm(3)
[1] 0.001349898
```

CONDICIONES DE CTL

The central limit theorem tells us that $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0,1)$. However, we rarely know σ . Later, we will see that we can estimate σ^2 from X_1, \dots, X_n by

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

This raises the following question: if we replace σ with S_n , is the central limit theorem still true? The answer is yes.

THE DELTA METHOD

Theorem (The Delta Method). *Suppose that*

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0, 1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0, 1).$$

In other words,

$$Y_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{implies that} \quad g(Y_n) \approx N\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right).$$

EJEMPLO

Example. Let X_1, \dots, X_n be IID with finite mean μ and finite variance σ^2 . By the central limit theorem, $\sqrt{n}(\bar{X}_n - \mu)/\sigma \rightsquigarrow N(0, 1)$. Let $W_n = e^{\bar{X}_n}$. Thus, $W_n = g(\bar{X}_n)$ where $g(s) = e^s$. Since $g'(s) = e^s$, the delta method implies that $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$. ■

DISTRIBUCIÓN MUESTRAL DE S^2

Si S^2 es la varianza de una muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2 , entonces el estadístico

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $\nu = n - 1$ grados de libertad.

DISTRIBUCIÓN MUESTRAL DE S^2

Un fabricante de baterías para automóvil garantiza que su producto durará, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de 1.9, 2.4, 3.0, 3.5 y 4.2 años, ¿el fabricante continuará convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que las duraciones de las baterías siguen una distribución normal.

```
ggplot() +  
  geom_function(aes(color = "chi-cuadrado"),  
               fun = dchisq, args = list(df = 4)) +  
  xlim(0,20) +  
  geom_vline(xintercept = qchisq(0.025, df = 4), linetype = 2, color = "blue") +  
  geom_vline(xintercept = qchisq(0.975, df = 4), linetype = 2, color = "blue")
```

<http://jgbabativam.rbind.io/post/2020-07-14-graficas-de-funciones-con-ggplot2/>

DISTRIBUCIÓN MUESTRAL DE S^2

```
> x <- c(1.9, 2.4, 3.0, 3.5, 4.2)
> xb <- mean(x)
> s <- sd(x)
> n <- length(x)
> x2 <- (n-1)*s^2/1^2
```

Agregue el valor de X2 al gráfico y concluya, coloque un color específico para esta línea.

BERRY-ESSÈEN THEOREM

Theorem. *Assume the same conditions as the CLT. Then,*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \rightsquigarrow N(0, 1).$$

Para n pequeño se aproxima por una distribución t-student con $n-1$ gl.

DISTRIBUCIÓN t (VARIANZA DESCONOCIDA)

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}}$$

Para n mayor que 30, se aproxima por una distribución t-student con n-1 gl.

Use el argumento **fun=dt** en la geometría **geom_fuction** y use varias capas para mostrar el resultado con `v = c(4, 10, 20, 50, 100, 200)` y agregue una distribución normal.

DISTRIBUCIÓN F

$$F = \frac{U/v_1}{V/v_2},$$

Es el cociente entre dos v.a. con distribución chi-cuadrado

Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianzas σ_1^2 y σ_2^2 , respectivamente, entonces,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

EJERCICIO

Un ingeniero químico afirma que el rendimiento medio de la población de un cierto proceso de lotes es 500 gramos por mililitro de materia prima. Para verificar dicha afirmación muestrea 25 lotes cada mes. Si el valor t calculado cae entre $-t_{0.05}$ y $t_{0.05}$, queda satisfecho con su afirmación. ¿Qué conclusión debería sacar de una muestra que tiene una media $\bar{x} = 518$ gramos por mililitro y una desviación estándar muestral $s = 40$ gramos? Suponga que la distribución de rendimientos es aproximadamente normal.

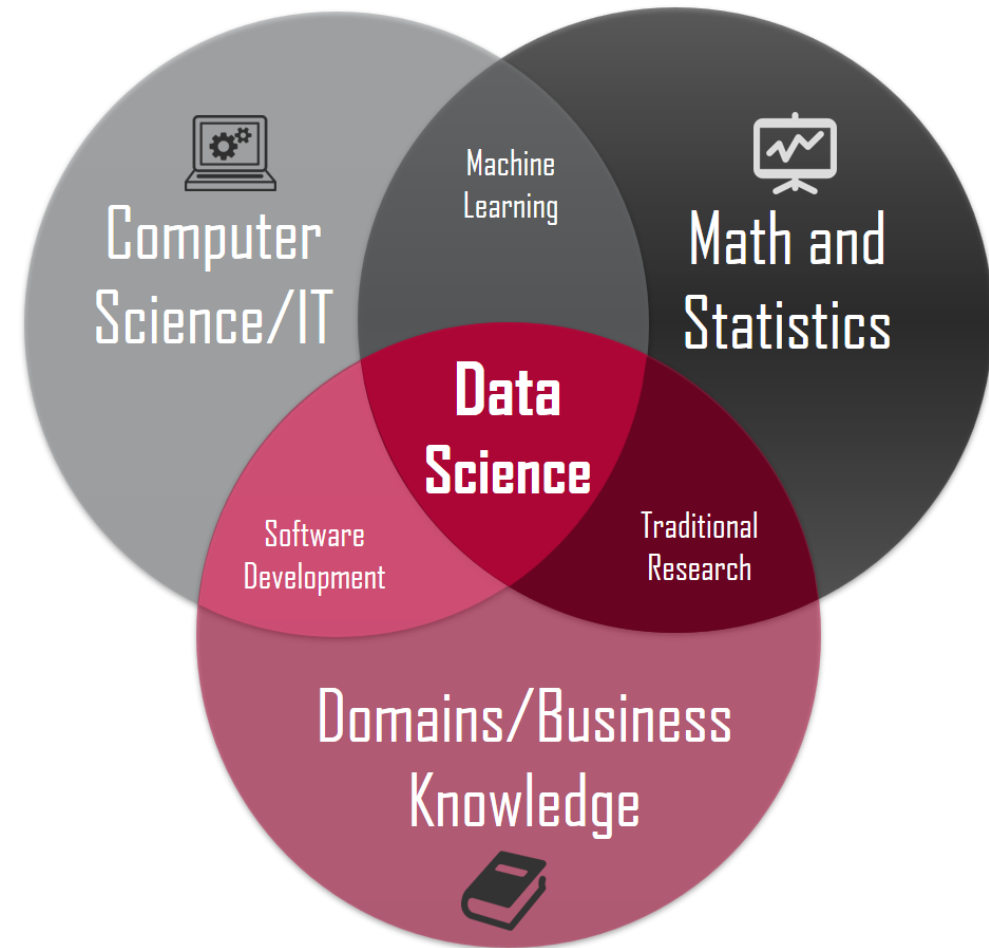
```
> qt(0.05, df = 24)
[1] -1.710882
> qt(0.95, df = 24)
[1] 1.710882
```

```
> xb <- 518
> s <- 40
> n <- 25
> t <- (xb - 500)/(s/sqrt(n))
> t
[1] 2.25
```

ESTADÍSTICA INFERENCIAL

STATISTICAL LEARNING

DATA SCIENCE



INTRODUCCIÓN

Statistical inference, or “learning” as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is:

Given a sample $X_1, \dots, X_n \sim F$, how do we infer F ?

In some cases, we may want to infer only some feature of F such as its mean.

SOPORTE TEÓRICO

A statistical model \mathfrak{F} is a set of distributions (or densities or regression functions). A **parametric model** is a set \mathfrak{F} that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}, \quad \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

This is a two-parameter model. We have written the density as $f(x; \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, a parametric model takes the form

$$\mathfrak{F} = \left\{ f(x; \theta) : \theta \in \Theta \right\}$$

SOPORTE TEÓRICO

Métodos paramétricos: Se denominan así porque suponen que los datos provienen de alguna distribución F indexada por un parámetro. El supuesto que se impone sobre la distribución de X , en pruebas de hipótesis sobre medias, es que ésta es de familia normal. Cuando se aplican estos métodos sin que se cumplan los supuestos requeridos las pruebas tienden a no conservar el error de tipo I, luego las conclusiones extraídas de la misma pueden ser erradas.

Métodos no paramétricos: Se denominan así porque no se impone ningún supuesto sobre la distribución de X (métodos libres de la distribución) o los supuestos que se imponen son solo condiciones de regularidad (continuidad, simetría), pero estos datos aún se encuentran dados por una familia de distribuciones F indexada por infinitos parámetros.

ESTIMACIÓN PUNTUAL

Point estimation refers to providing a single “best guess” of some quantity of interest. The quantity of interest could be a parameter in a parametric model, a CDF F , a probability density function f , a regression function r , or a prediction for a future value Y of some random variable.

By convention, we denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$. Remember that θ is a fixed, unknown quantity. The estimate $\hat{\theta}$ depends on the data so $\hat{\theta}$ is a random variable.

More formally, let X_1, \dots, X_n be n IID data points from some distribution F . A point estimator $\hat{\theta}_n$ of a parameter θ is some function of X_1, \dots, X_n :

$$\hat{\theta}_n = g(X_1, \dots, X_n).$$

SESGO – ESTIMADOR INSESGADO

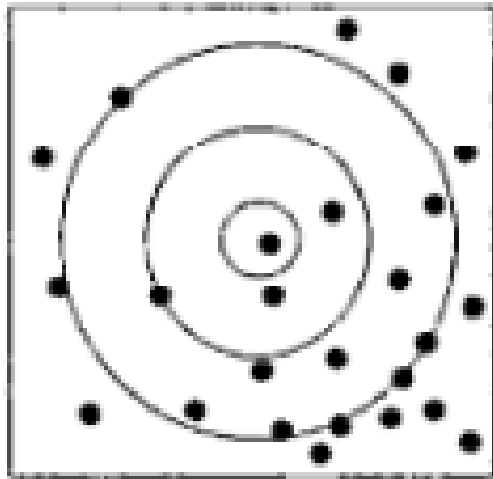
$$\text{bias}(\hat{\theta}_n) = \mathbb{E}_{\theta}(\hat{\theta}_n) - \theta.$$

$\hat{\theta}_n$ is **unbiased** if $\mathbb{E}(\hat{\theta}_n) = \theta$.

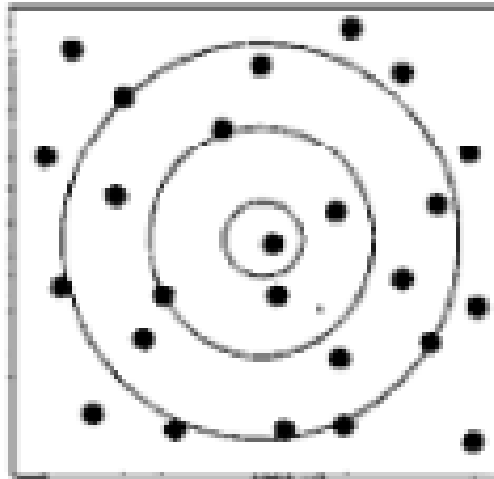
Se dice que un estadístico $\hat{\Theta}$ es un **estimador insesgado** del parámetro θ si

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

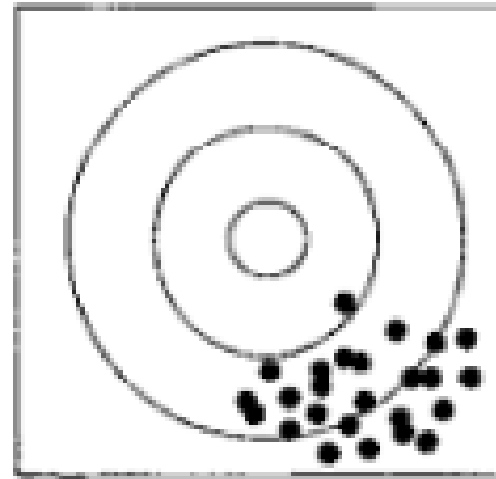
PROPIEDADES DE UN ESTIMADOR



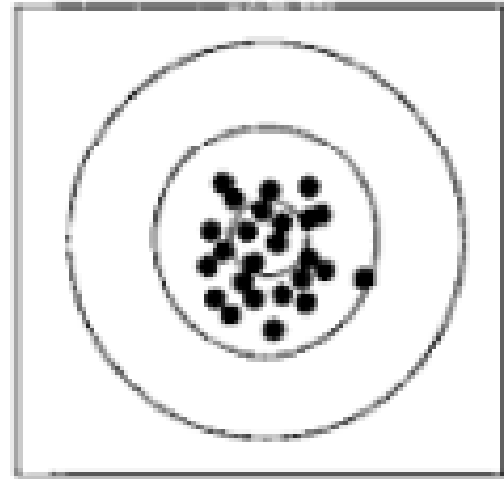
(a) Biased, imprecise



(b) Unbiased, imprecise



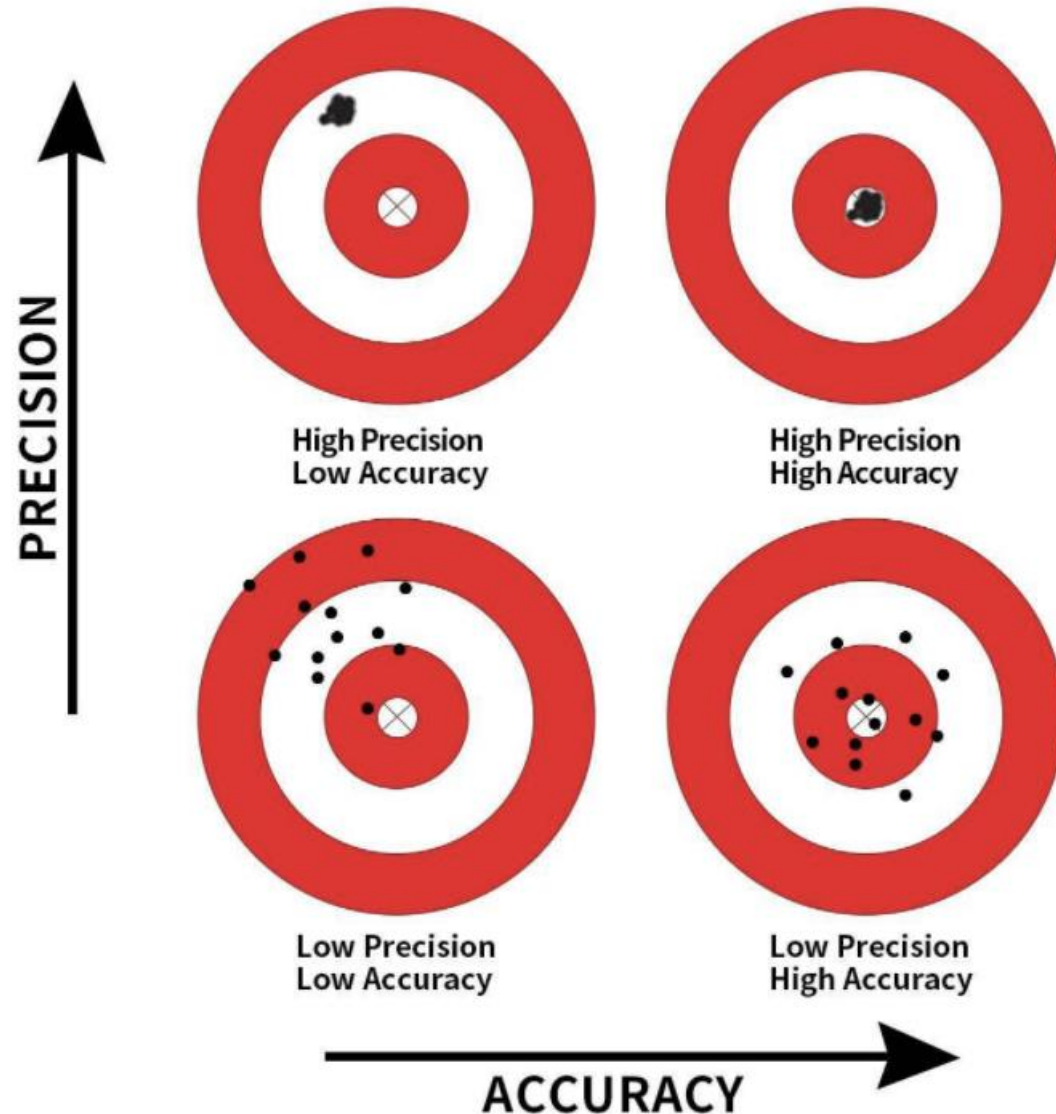
(c) Biased, precise



(d) Unbiased, precise

PROPIEDADES DE UN ESTIMADOR

Sesgo (Exactitud) vs Precisión



¿El tamaño de la muestra resuelve todos los problemas?

PROPIEDADES DE UN ESTIMADOR



CONSISTENCIA

Definition. *A point estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$.*

ERROR ESTÁNDAR

$$\text{se} = \text{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}.$$

ERROR CUADRÁTICO MEDIO

PARÁMETROS DE INTERÉS

Often, we are only interested in some function $T(\theta)$. For example, if $X \sim N(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ then $\mu = T(\theta)$ is called the **parameter of interest** and σ is called a **nuisance parameter**. The parameter of interest might be a complicated function of θ as in the following example.

MÉTODO DE LOS MOMENTOS

Definition. *The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ such that*

$$\alpha_1(\hat{\theta}_n) = \hat{\alpha}_1$$

$$\alpha_2(\hat{\theta}_n) = \hat{\alpha}_2$$

$$\vdots \quad \vdots \quad \vdots$$

$$\alpha_k(\hat{\theta}_n) = \hat{\alpha}_k.$$

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i^1, \quad \mu'_1 = EX^1,$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2, \quad \mu'_2 = EX^2,$$

$$\vdots$$

$$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad \mu'_k = EX^k.$$

EJEMPLO

Example. Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\hat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$. By equating these we get the estimator

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad \blacksquare$$

EJEMPLO

Example. Let $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations¹

$$\begin{aligned}\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}^2 + \hat{\mu}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2.\end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\begin{aligned}\hat{\mu} &= \overline{X}_n \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2. \quad \blacksquare\end{aligned}$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

MÉTODO DE LA MÁXIMA VEROSIMILITUD

Definition. *The likelihood function is defined by*

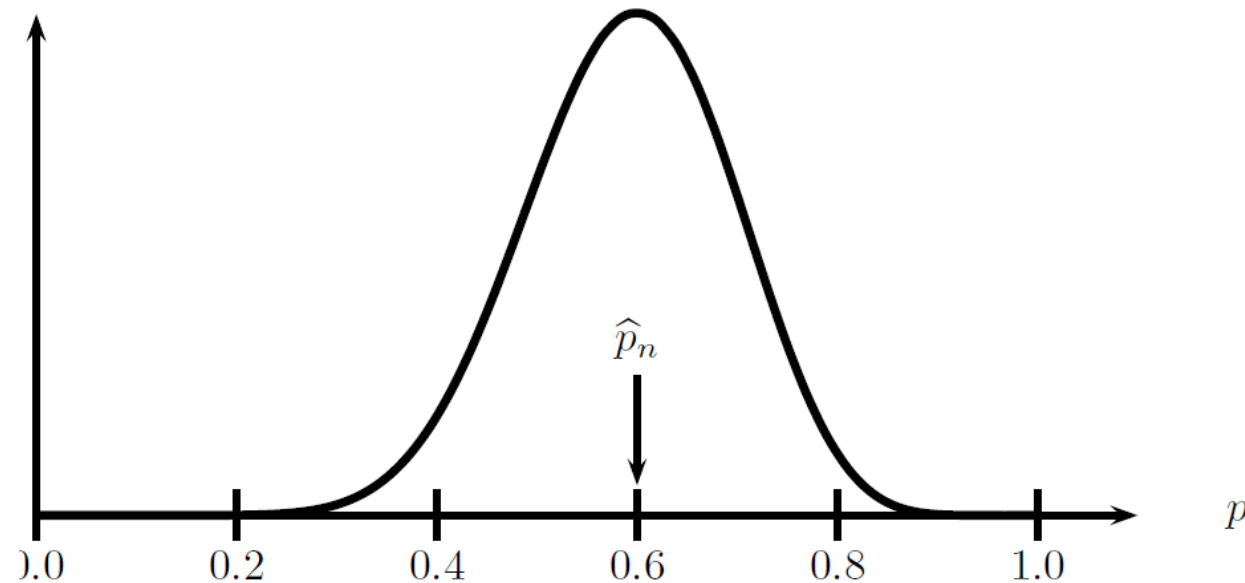
$$\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i; \theta).$$

The log-likelihood function is defined by $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we **treat it is a function of the parameter θ** . Thus, $\mathcal{L}_n : \Theta \rightarrow [0, \infty)$. The likelihood function is not a density function: in general, it is **not** true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

MÉTODO DE LA MÁXIMA VEROSIMILITUD

Definition. *The maximum likelihood estimator MLE, denoted by $\hat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$.*



EJEMPLO

Example. Suppose that $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p . Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(p) = S \log p + (n - S) \log(1 - p).$$

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\hat{p}_n = S/n$.

EJEMPLO

Example. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned}\mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\}\end{aligned}$$

EJEMPLO

$$= \sigma^{-n} \exp \left\{ -\frac{nS^2}{2\sigma^2} \right\} \exp \left\{ -\frac{n(\bar{X} - \mu)^2}{2\sigma^2} \right\}$$

where $\bar{X} = n^{-1} \sum_i X_i$ is the sample mean and $S^2 = n^{-1} \sum_i (X_i - \bar{X})^2$. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$ and then expanding the square. The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood. ■

ESTIMADORES COMUNES

Scenario	Population parameter	Notation	Point estimate	Symbol(s)
1	Population proportion	p	Sample proportion	\hat{p}
2	Population mean	μ	Sample mean	\bar{x} or $\hat{\mu}$
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proportions	$\hat{p}_1 - \hat{p}_2$
4	Difference in population means	$\mu_1 - \mu_2$	Difference in sample means	$\bar{x}_1 - \bar{x}_2$
5	Population regression slope	β_1	Fitted regression slope	b_1 or $\hat{\beta}_1$

A VECES NO ES POSIBLE USAR MÉTODOS EXACTOS

Open Access Article

Logistic Biplot by Conjugate Gradient Algorithms and Iterated SVD

by  Jose Giovany Babativa-Márquez ^{1,2,*}   and  José Luis Vicente-Villardón ¹  

¹ Department of Statistics, University of Salamanca, 37008 Salamanca, Spain

² Facultad de Ciencias de la Salud y del Deporte, Fundación Universitaria del Área Andina, Bogotá 1321, Colombia

* Author to whom correspondence should be addressed.

Academic Editor: Liangxiao Jiang

Mathematics **2021**, *9*(16), 2015; <https://doi.org/10.3390/math9162015>

Received: 24 June 2021 / Revised: 19 August 2021 / Accepted: 19 August 2021 / Published: 23 August 2021

(This article belongs to the Special Issue **Multivariate Statistics: Theory and Its Applications**)

Download PDF

Browse Figures

Citation Export

<https://www.mdpi.com/2227-7390/9/16/2015/htm>