

Clase 02 - Introducción Variables Aleatorias

Análisis de datos con R para Biociencias

Dra. María Angélica Rueda | Pontificia Universidad Católica de
Valparaíso

18 January 2022

PLAN DE LA CLASE

1.- Introducción

- ▶ Clasificación de variables aleatorias.
- ▶ Observar y predecir variables cuantitativas continuas y discretas.
- ▶ ¿Qué es un script?
- ▶ Formato correcto para importar datos a R.

2.- Práctica con R y Rstudio cloud

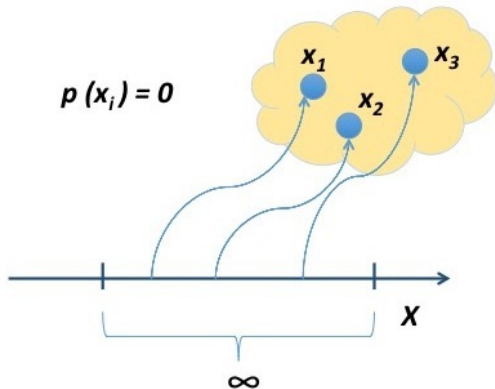
- ▶ Elaborar un script de R e importar datos desde excel.
- ▶ Observar y predecir variable aleatoria con distribución Normal.
- ▶ Observar y predecir variables aleatorias discretas con distribución Bernoulli o Binomial.

TIPOS DE VARIABLES ALEATORIAS



VARIABLE ALEATORIA CUANTITATIVA CONTINUA

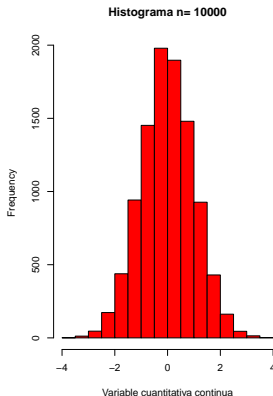
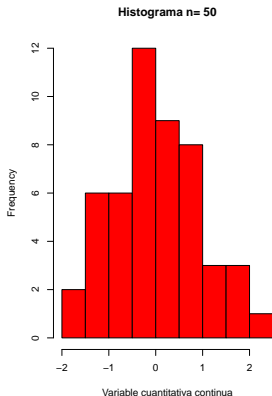
Definición: Puede tomar cualquier valor dentro de un intervalo (a,b) , (a,Inf) , $(-\text{Inf},b)$, $(-\text{Inf},\text{Inf})$ y la probabilidad que toma cualquier punto es 0, debido a que existe un número infinito de posibilidades.



OBSERVAR UNA VARIABLE CUANTITATIVA CONTINUA

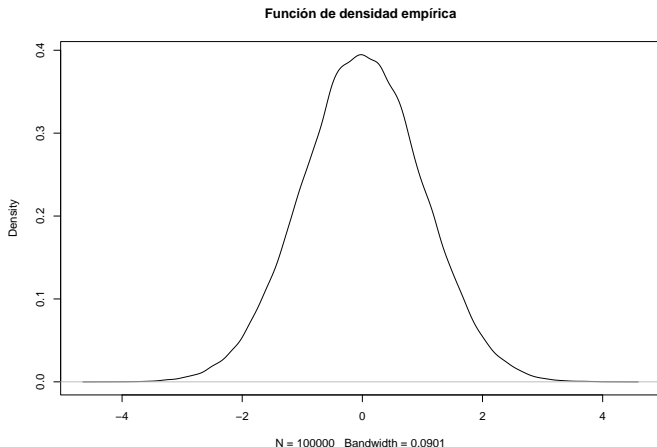
Al observar con un histograma **hist()** notamos que:

1. La frecuencia o probabilidad en un intervalo es distinta de cero.
2. Cuando aumenta el **n** muestral se perfila una distribución llamada **normal**.



PREDECIR UNA VARIABLE CUANTITATIVA CONTINUA

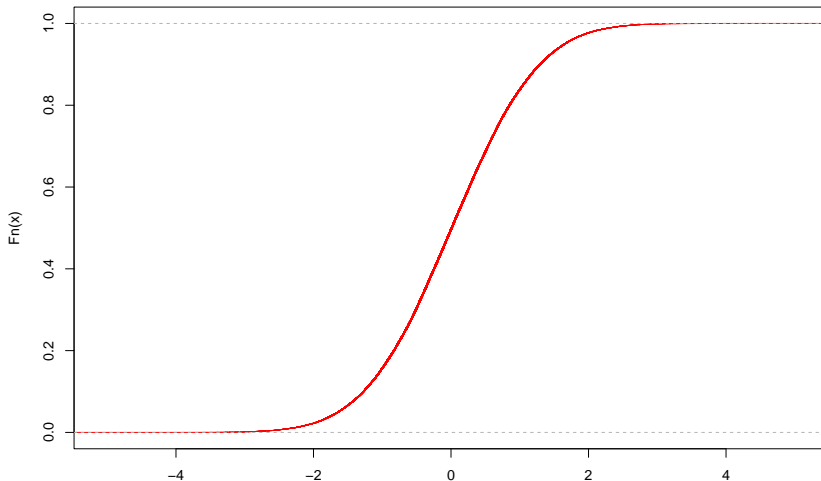
Podemos predecir la probabilidad de que la variable aleatoria tome un determinado valor usando la función de densidad empírica **density()**.



PREDECIR VARIABLES CONTINUAS 2

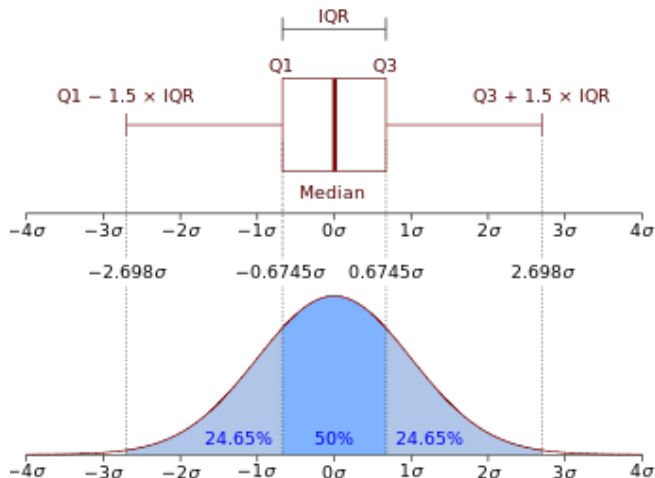
Podemos predecir la probabilidad de que la variable aleatoria tome un valor menor o igual a un determinado valor, usando la función de distribución empírica acumulada **ecdf()**.

Función de distribución empírica acumulada



OBSERVAR CON BOXPLOT

Las gráficas de cajas y bigotes (**boxplot()**) son muy adecuadas para observar variables aleatorias continuas.



VARIABLES ALEATORIAS DISCRETAS

Las variables aleatorias discretas son aquellas que presentan un número contable de valores; por ejemplo:

- ▶ **Número de parásitos** (1, 3, 5, 6, etc.).
- ▶ **Número de especies.**
- ▶ **Número de crías vivas o hijos o huevos.**
- ▶ **Número de semillas.**

IMPORTANCIA DE IDENTIFICAR Y ANALIZAR VARIABLES ALEATORIAS DISCRETAS

- ▶ Es importante identificar la naturaleza que tiene nuestra variable en estudio, y así evitar errores en los análisis estadísticos que llevemos a cabo.
- ▶ Usualmente cuando las variables en estudio son conteos, proporciones o binarias (éxito o fracaso, macho o hembra, sano o enfermo) deben ser consideradas como **variables aleatorias discretas**.
- ▶ Según sea la variable aleatoria discreta, ella tendrá una función de distribución de probabilidad asociada que no es normal. Por ejemplo: **Bernoulli, Binomial, Binomial Negativa, Poisson, entre otras**.
- ▶ En gran parte, la *distribución de variables aleatorias discretas* suelen ser **asimétricas a derecha o a izquierda**.

EJEMPLO VARIABLE ALEATORIA BINARIA - DISTRIBUCIÓN BERNOULLI

Se realiza una prueba aleatoria de COVID-19 en los pasajeros de un avión (160 pasajeros en total) determinando que 8 de ellos son positivos. Sea $X=1$ si la persona tiene PCR+ y $X=0$ en el caso de que el PRC-. ¿Cuál es la distribución de X ? 8/160 = éxito, 152/160 = fracaso.

	Fracaso	Éxito
x	0	1
$f(x)=P(X=x)$	$1-p$ 0.95	p 0.05

$$f(x) = P(X = x) = \begin{cases} 1 - p & ; si \quad x = 0 \\ p & ; si \quad x = 1 \end{cases}$$

EJEMPLO VARIABLE ALEATORIA DISCRETA - DISTRIBUCIÓN BINOMIAL

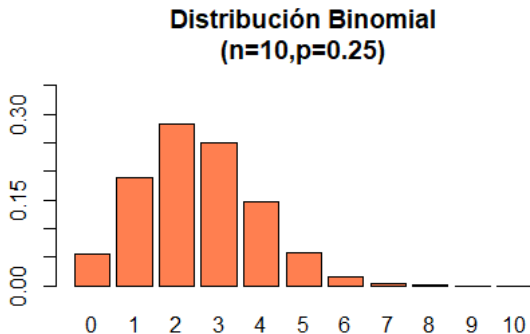


Figure 1: Número de parásitos por animal/planta

¿QUÉ ES UN SCRIPT?

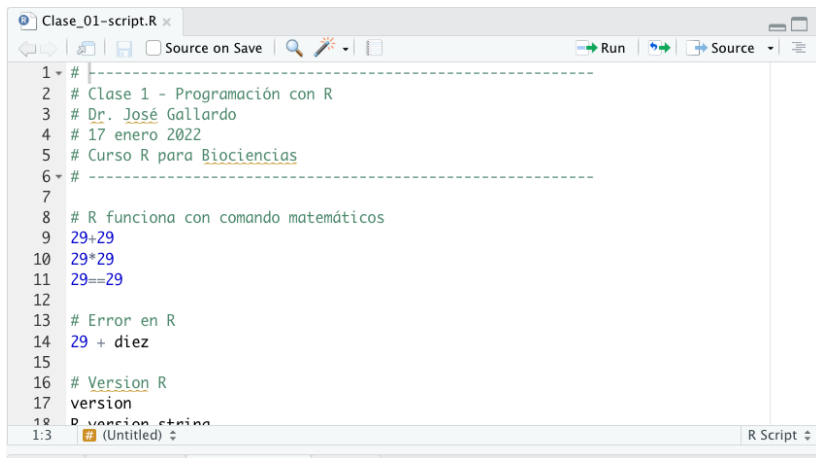
Los ***scripts*** son documentos de texto con una secuencia de comandos que permiten ejecutar programas.

Estos archivos son iguales a cualquier documentos de texto, pero **R puede leer y ejecutar** el código que contienen.

Los códigos de **R** están contenidos en librerías o packages.

Algunos ***script*** que usaremos en este curso tienen extensión de archivo **.R**, por ejemplo `mi_script.R` o **.Rmd** (R+markdown) por ejemplo `reporte.Rmd`.

EJEMPLO DE SCRIPT R



```
1 # -----  
2 # Clase 1 - Programación con R  
3 # Dr. José Gallardo  
4 # 17 enero 2022  
5 # Curso R para Biociencias  
6 # -----  
7  
8 # R funciona con comando matemáticos  
9 29+29  
10 29*29  
11 29==29  
12  
13 # Error en R  
14 29 + diez  
15  
16 # Version R  
17 version  
18 R.version.string  
1:3 # (Untitled) R Script
```

FORMATO CORRECTO PARA IMPORTAR A R

	A	B	C	D	E	F
1	sample_id	Weight	sex			
2	1	17,2	female			
3	2	18,8	female			
4	3	27,8	male			
5	4	20,4	male			
6	5	20,6	male			
7	6	28,6	male			
8	7	22,3	male			
9	8	13,7	female			
10	9	16,6	female			
11	10	17,8	female			
12	11	26,1	female			
13	12	21,8	male			
14	13	22	male			
15	14	20,6	male			
16	15	17,2	female			
17	16	28,9	male			
18	17	22,5	male			
19	18	10,2	female			
20	19	23,5	male			
21	20	17,6	female			
22	21	14,7	female			
23	22	18,9	female			
24	23	14,9	female			
25	24	16,4	female			
26	25	16,9	female			
27	26	11,6	female			

Nombres de
variables

Observaciones
o datos

Figure 2: Formato correcto de archivo excel para que sea importado a R

ERRORES EN FORMATO EXCEL

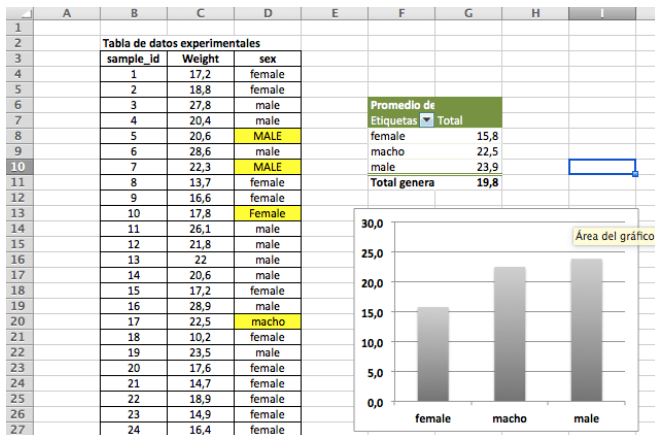


Figure 3: Errores comunes antes de importar a excel

Importante: No colocar símbolos matemáticos por ejemplo (%,\$,+) como nombres de las **(variables)**.

ERRORES EN FORMATO EXCEL 2

sample_id	Weight	sex		sample_id	Weight	sex	Observaciones
1	17,2	female		1	17,2	female	
2	18,8	female		2	18,8	female	
3	27,8	male		3	27,8	male	
4	20,4	male		4	20,4	male	
5	20,6	male		5	20,6	male	
6	28,6	male		6	28,6	male	
7	sin registro	male		7		male	
8	13,7	female		8	13,7	female	
9	16,6	female		9	16,6	female	
10	17,8	female		10	17,8	female	
11	26,1	male		11	26,1	male	
12	21,8	male		12	21,8	male	
13	22	Indeterminado		13	22	NA	Sexo Indeterminado
14	20,6	male		14	20,6	male	
15	17,2	female		15	17,2	female	
16	28,9	male		16	28,9	male	
17	22,5, cola deforme	male		17	22,5	male	cola deforme
18	10,2	female		18	10,2	female	
19	23,5	male		19	23,5	male	

Figure 4: Errores comunes antes de importar a excel

Importante: No colocar comentarios en las celdas de datos. Dejar celdas vacias o usar el simbolo *NA* es preferido cuando hay datos faltantes.

COMO IMPORTAR DATOS EXCEL A R

Antes de importar un archivo en formato excel (**.xlsx** o **.xls**) debe instalar y tener habilitada la librería **readxl**

```
library(readxl)
dat <- read_excel("Data.xlsx", sheet = 1)
dat <- read_excel("Data.xlsx", sheet = "Poblacion 1")
```

PRÁCTICA ANÁLISIS DE DATOS

- 1.- Guía de trabajo Rmarkdown disponible en drive.
- 2.- La tarea se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- ▶ Identificamos y clasificamos variables aleatorias.
- ▶ Observamos la distribución de una variable cuantitativa continua usando histograma y boxplot.
- ▶ Predecimos el comportamiento de una variable cuantitativa continua con distribución normal usando funciones de densidad y de distribución acumulada.
- ▶ Estudiamos sobre variables aleatorias discretas y algunas distribuciones de probabilidad asociadas (Bernoulli y Binomial).