CLASE 02 - VARIABLES ALEATORIAS

DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R.

Dr. José Gallardo Matus | https://genomics.pucv.cl/

21 March 2022

PLAN DE LA CLASE

1. **Introducción**

- Diferencia entre variable, variable aleatoria, datos y factores.
- Clasificación de variables aleatorias.
- Observar y predecir variables cuantitativas continuas y discretas.
- 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.

CONCEPTOS Y DEFINICIONES

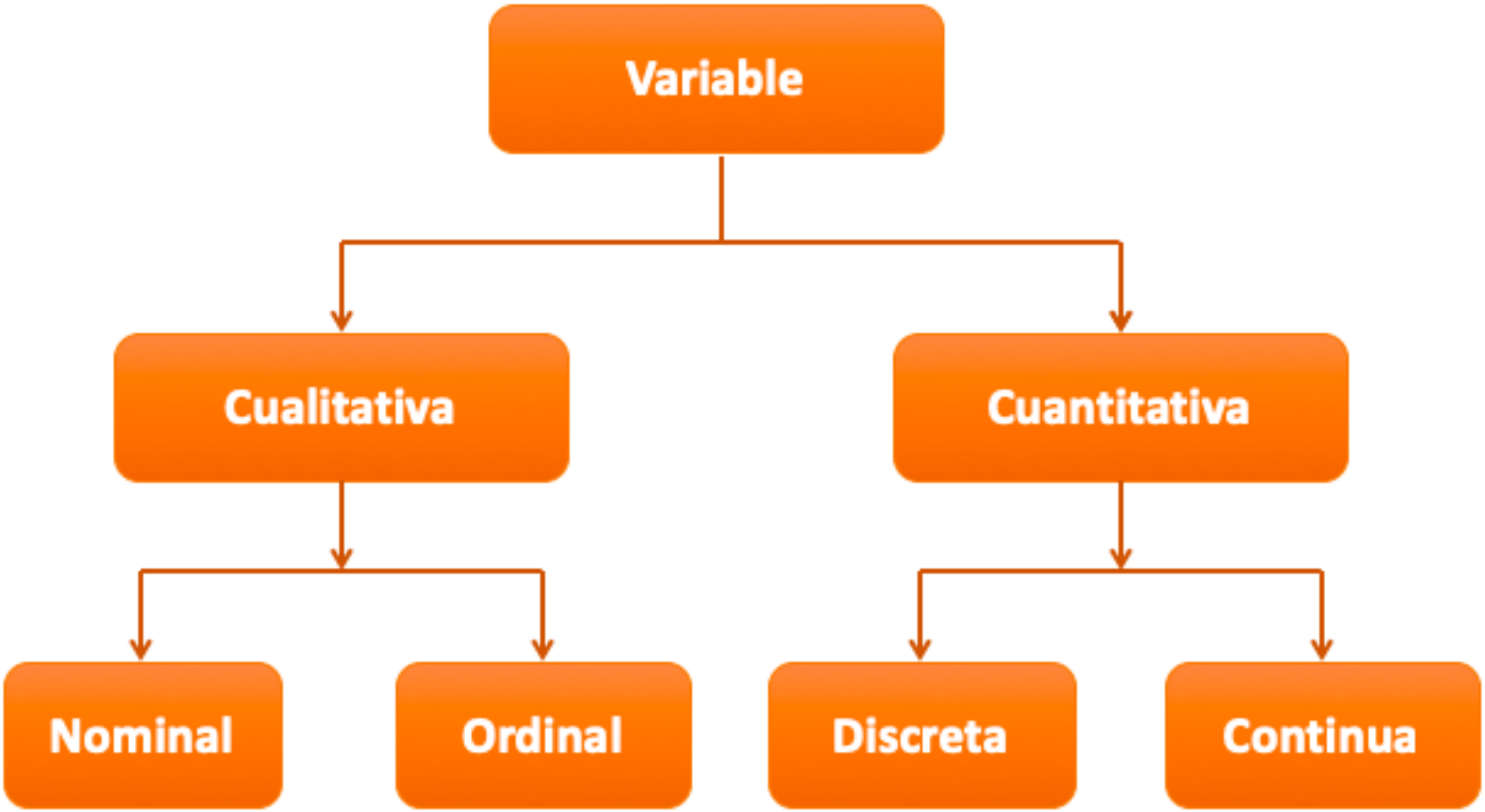
1. **Variable**: Características que se pueden medir u observar en un individuo o en un ambiente: peso, temperatura, Sexo, pH, Tipo de bacteria, abundancia, número de alelos, absorvancia.

2. **Variable aleatoria**: es un número que representa un resultado de un experimento aleatorio. Depende entonces de función matemática o distribución de probabilidad.

3. **Datos u observaciones**: Son los valores que puede tomar una variable aletoria. 25 gramos, 55 mm, 13ºC, 7 unidades de pH, 25 bacterias, 2 alelos, 32 ct, 1,5.

4. **Factor**: Usado para identificar tratamientos de un experimento o variables de clasificación. Se usan como *variables independientes o predictoras*, es decir tienen un efecto sobre una *variable respuesta o dependiente*. Ej. Sexo (niveles: macho o hembra) tiene un efecto sobre nivel de hormonas.

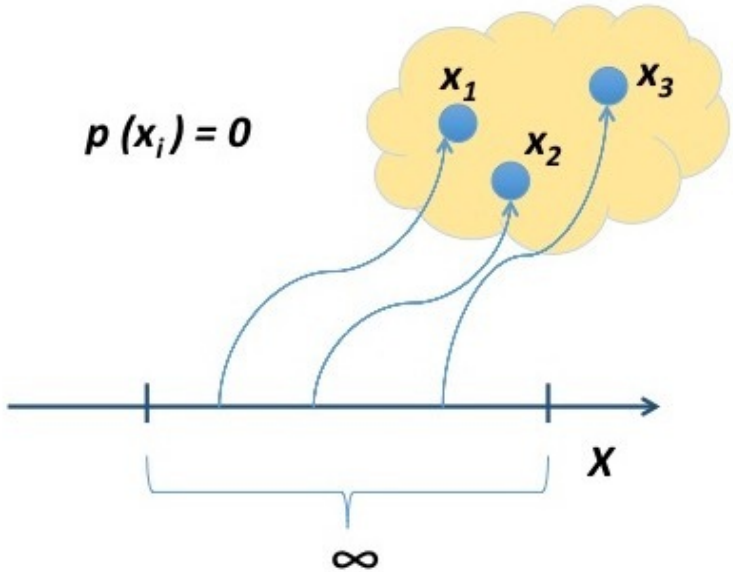
CLASIFICACIÓN DE VARIABLES



VARIABLE ALEATORIA CUANTITATIVA CONTINUA

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

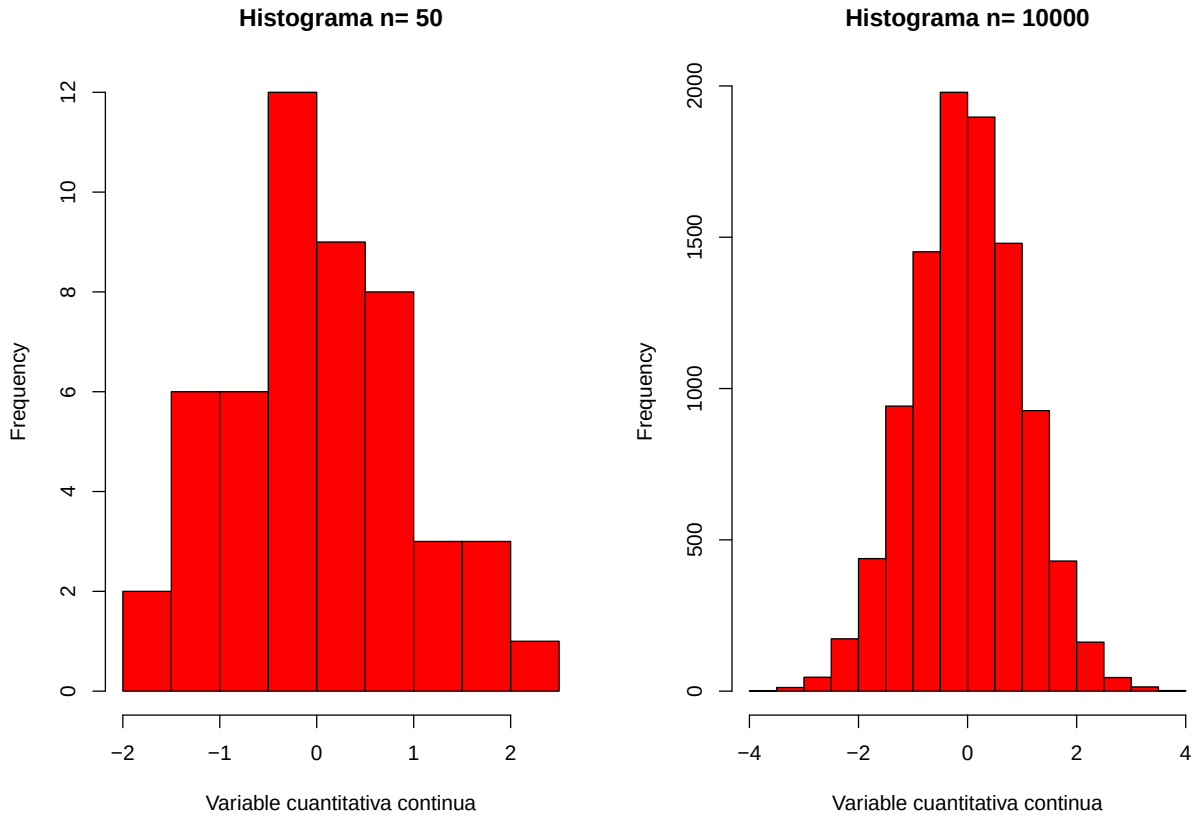
- Expresión relativa de un gen.
- Cantidad de un anticuerpo u hormonas producidas por un individuo.



OBSERVAR VARIABLE CONTINUA

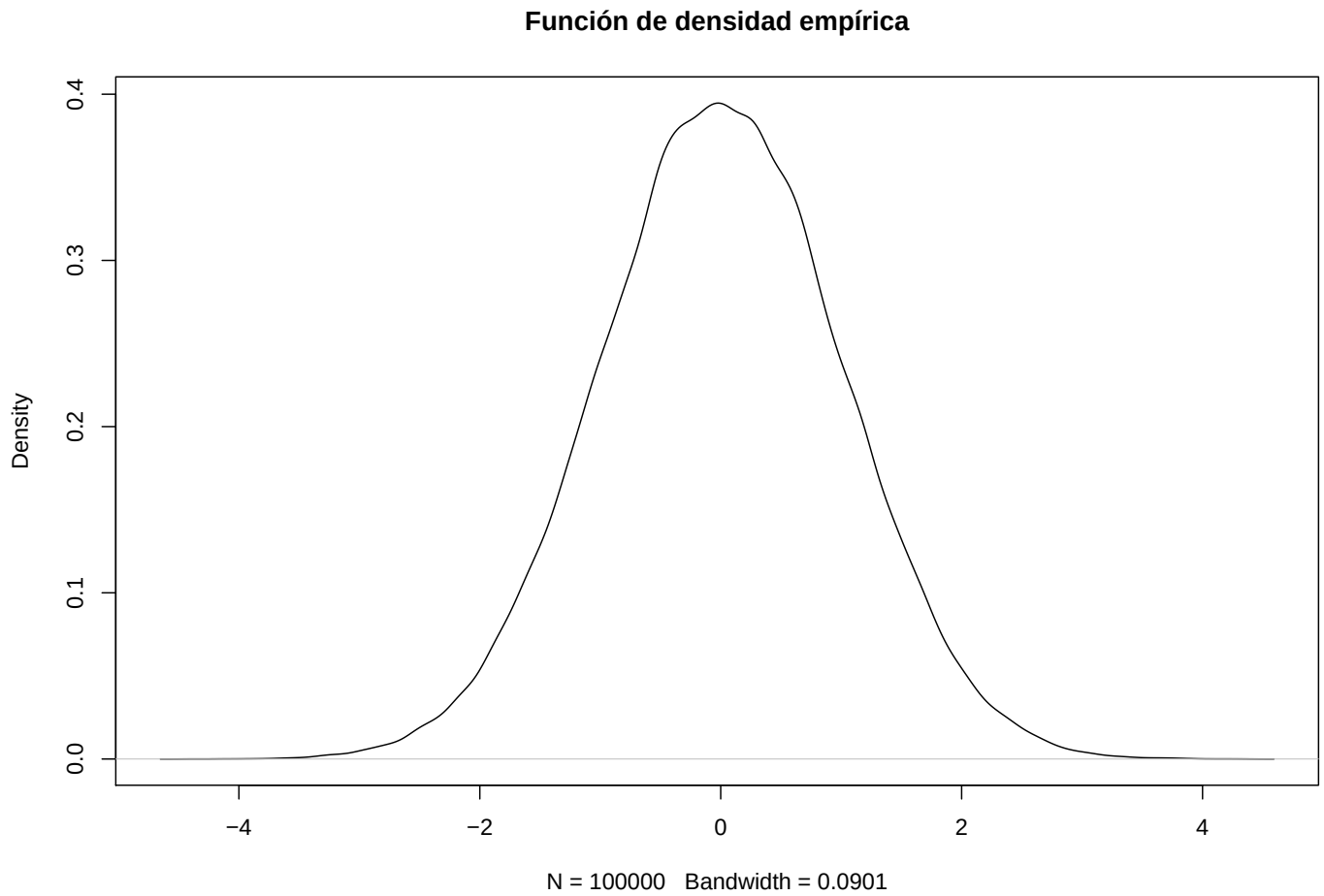
Al observar con un histograma notamos que:

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



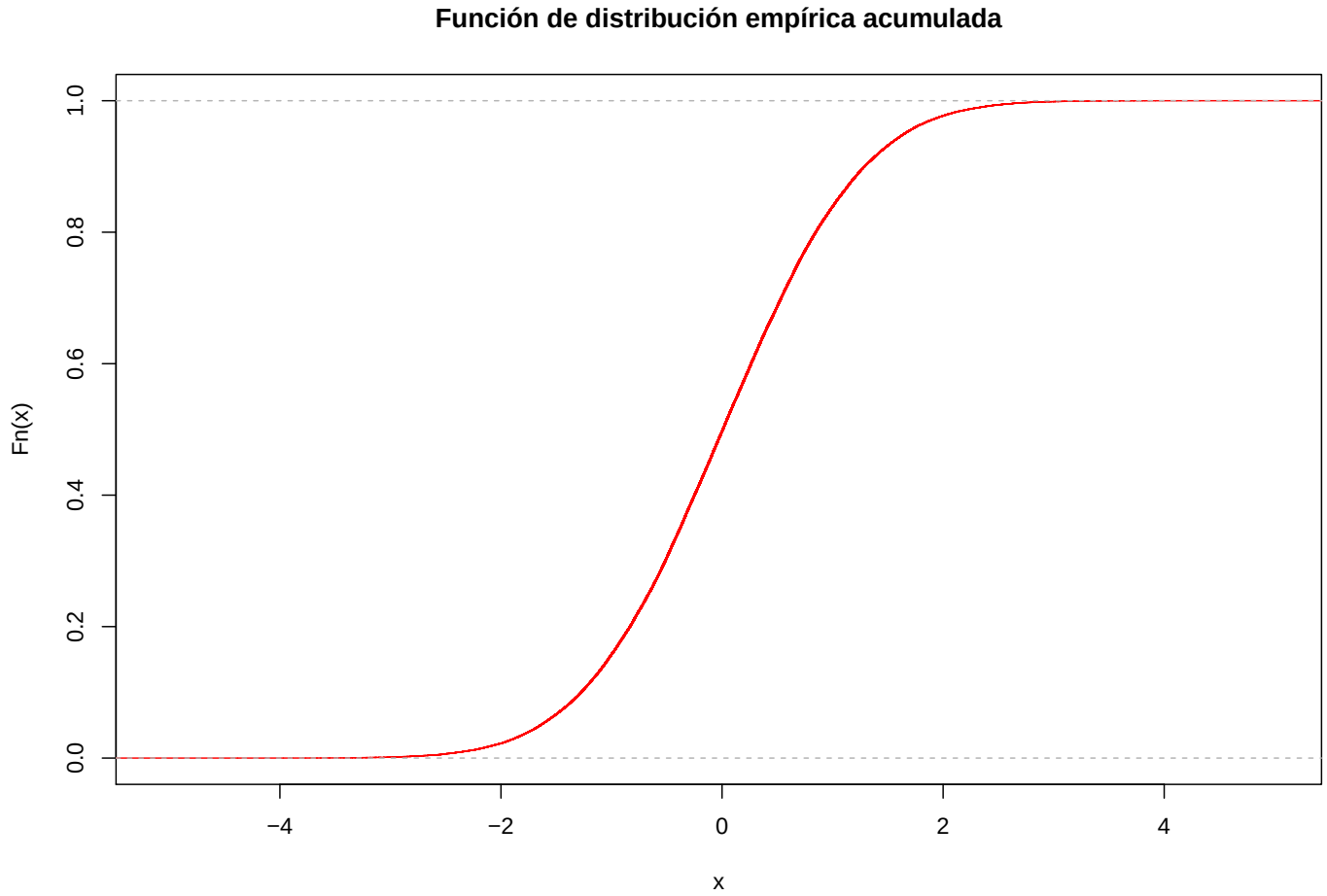
PREDECIR VARIABLE CONTINUA (V.C.)

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



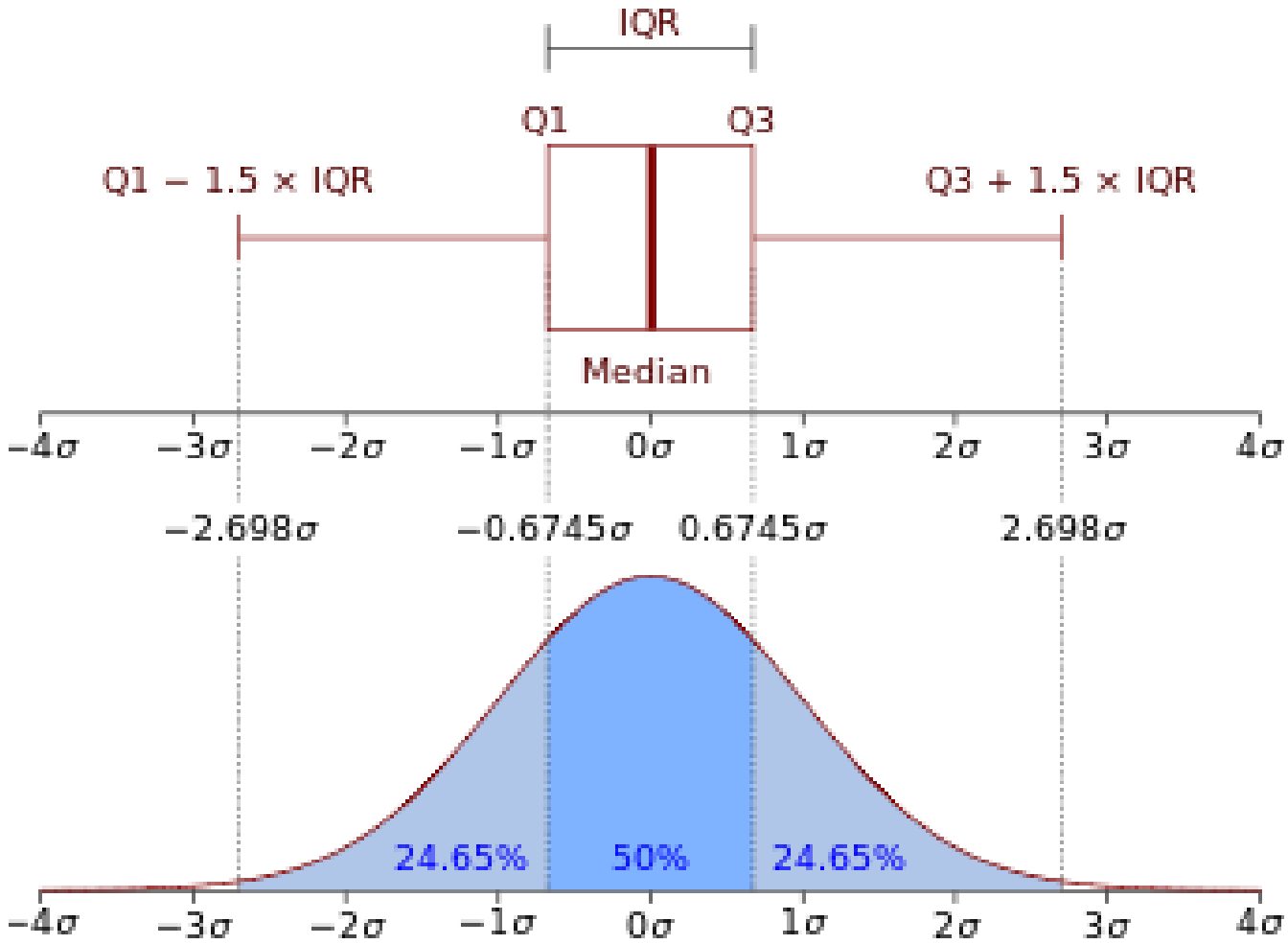
PREDECIR V.C.: DISTRIBUCIÓN ACUMULADA

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()**.



OBSERVAR CON BOXPLOT

Las gráficas de cajas y bigotes son muy adecuadas para observar variables aleatorias continuas.



VARIABLES ALEATORIAS DISCRETAS

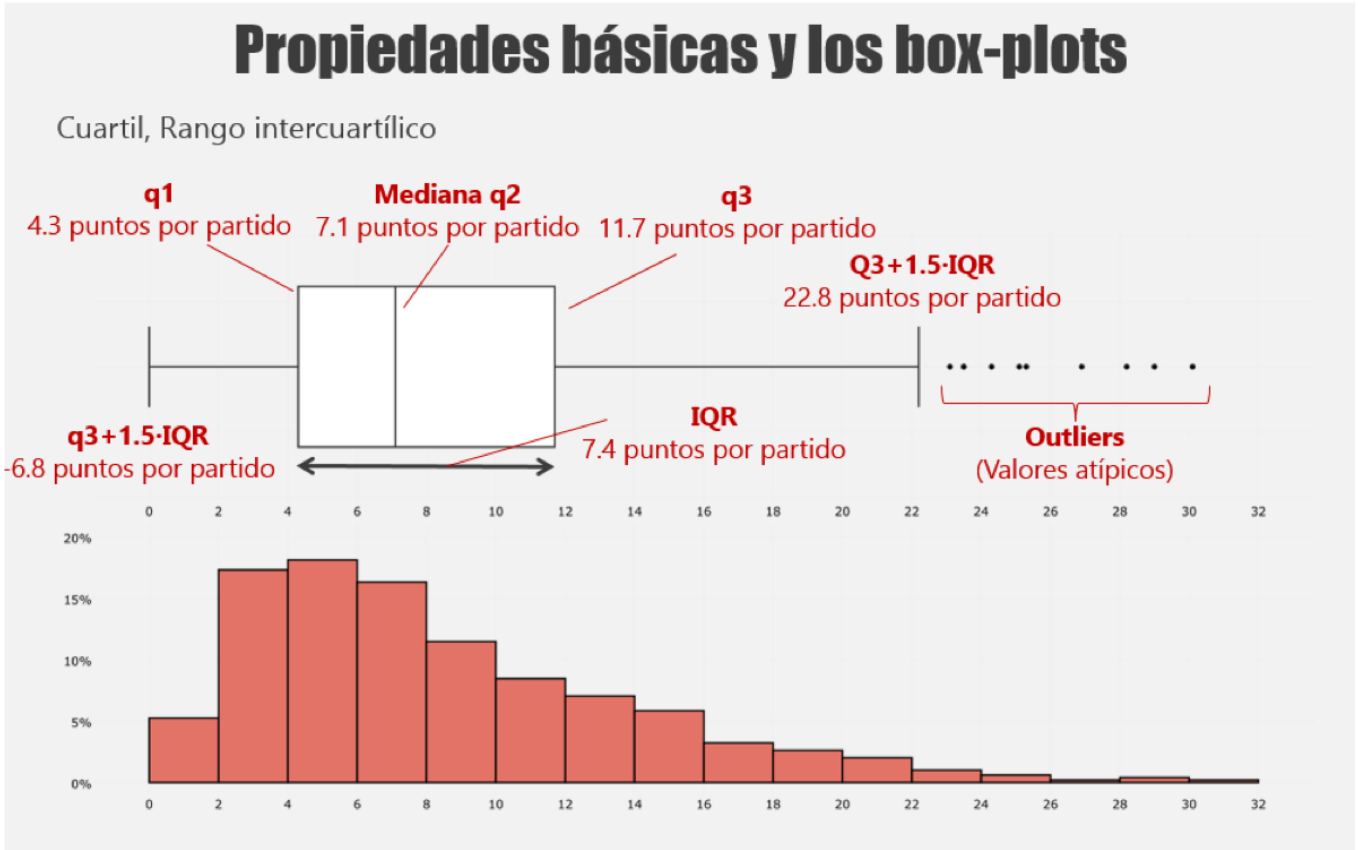
Las variables aleatorias discretas son aquellas que presentan un número contable de valores; por ejemplo:

- **Número de mutaciones** (1, 3, 5, 6, etc.).
- **Número de bacterias**.
- **Número de nucleótidos similares entre dos secuencias**.
- **Número de semillas de una fruta**.

IDENTIFICA CORRECTAMENTE TU VARIABLE

- 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**.

HISTOGRAMA Y BOXPLOT DE VARIABLE DISCRETA



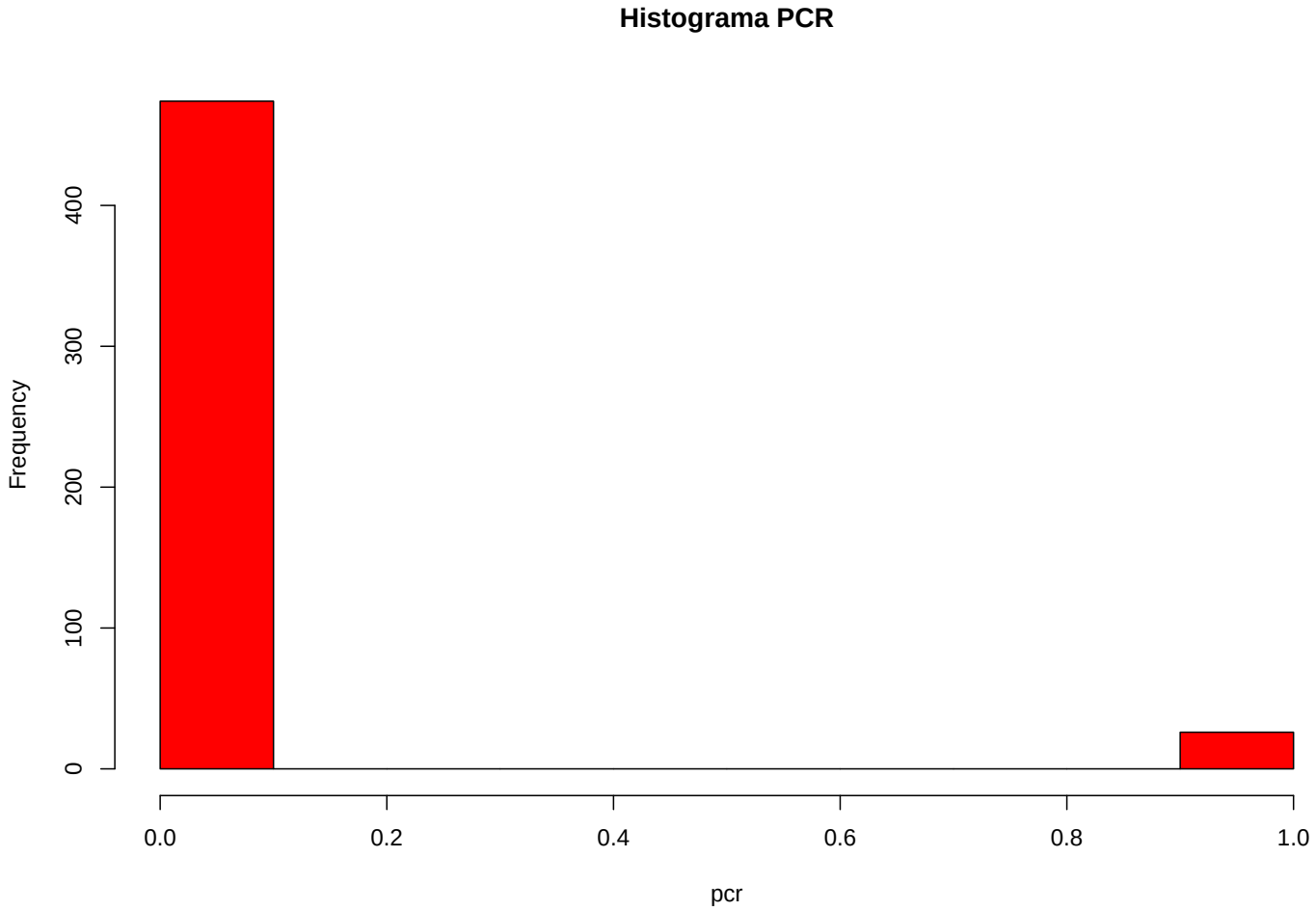
VARIABLE DISCRETA: DISTRIBUCIÓN BERNOULLI

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

VARIABLE DISCRETA: DISTRIBUCIÓN BERNOULLI

Representación en un histograma de la frecuencia de recuperados y fallecidos.



VARIABLE DISCRETA: DISTRIBUCIÓN BINOMIAL NEGATIVA

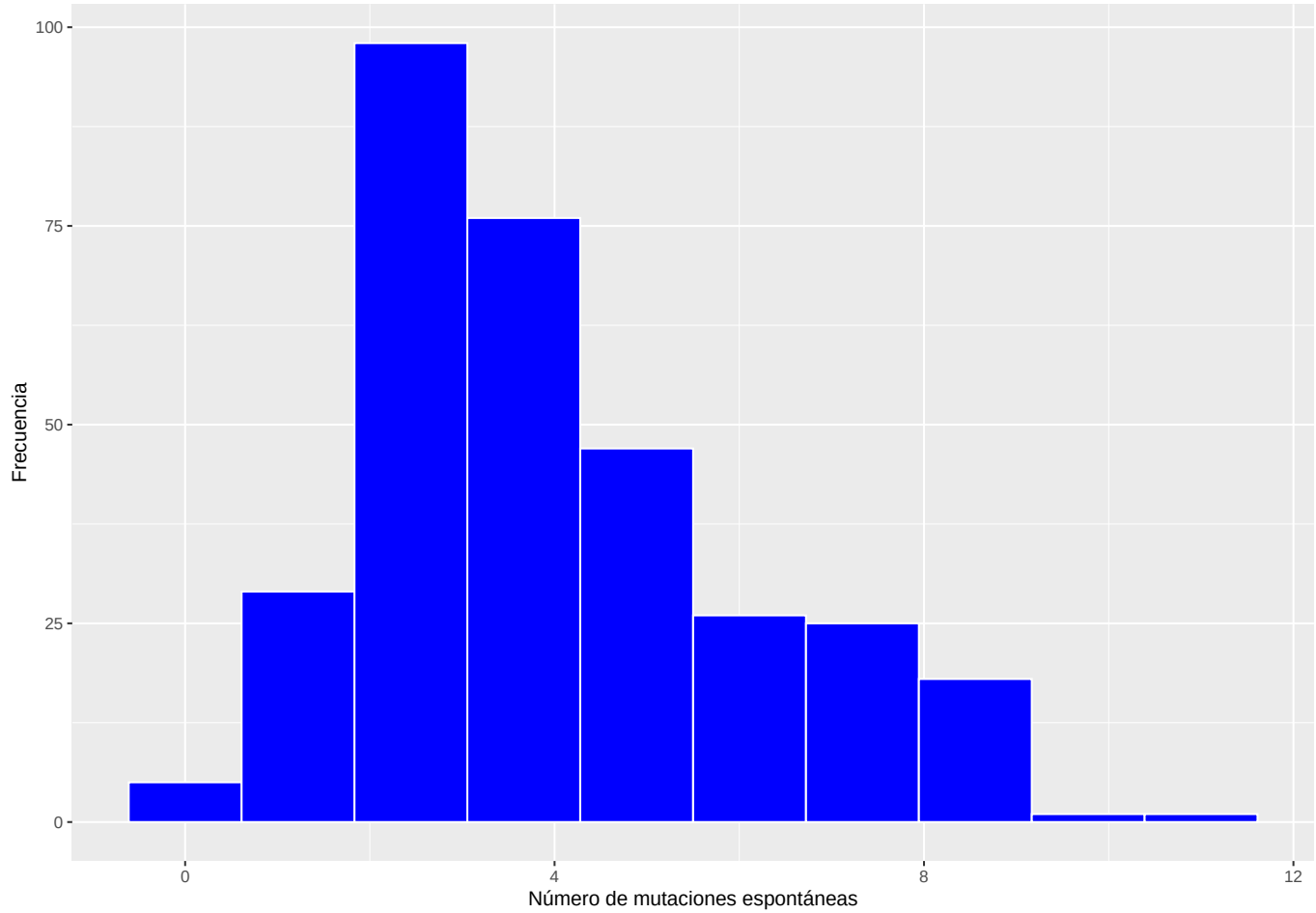
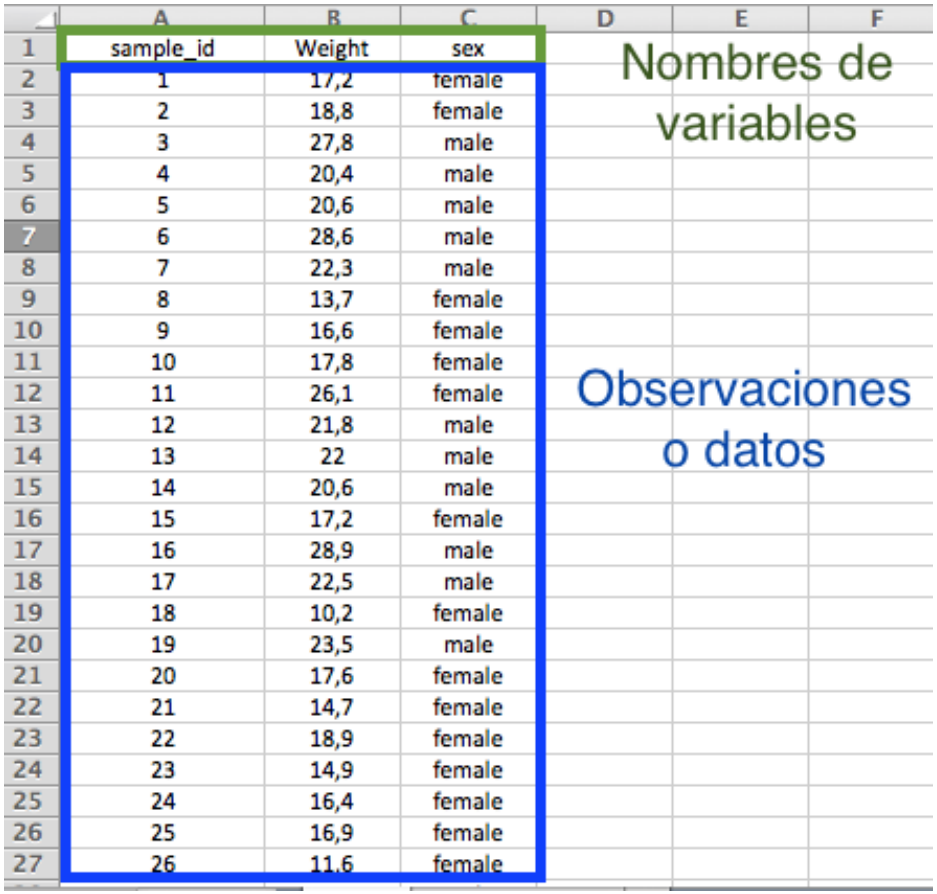


Figure 1: Número de mutaciones espontáneas en 326 líneas de levadura.

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			

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

ERRORES EN FORMATO EXCEL

Tabla de datos experimentales

sample_id	Weight	sex
1	17,2	female
2	18,8	female
3	27,8	male
4	20,4	male
5	20,6	MALE
6	28,6	male
7	22,3	MALE
8	13,7	female
9	16,6	female
10	17,8	Female
11	26,1	male
12	21,8	male
13	22	male
14	20,6	male
15	17,2	female
16	28,9	male
17	22,5	macho
18	10,2	female
19	23,5	male
20	17,6	female
21	14,7	female
22	18,9	female
23	14,9	female
24	16,4	female

Promedio de Etiquetas	Total
female	15,8
macho	22,5
male	23,9
Total genera	**19,8**

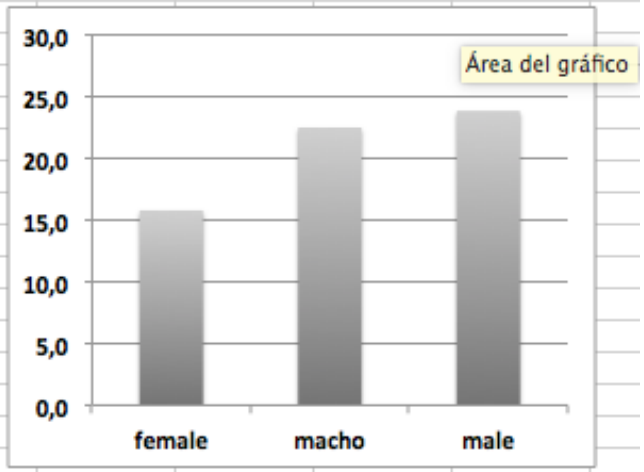


Figure 3: Errores comunes antes de importar a excel

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

ERRORES EN FORMATO EXCEL 2

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

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

Asuntos importantes:

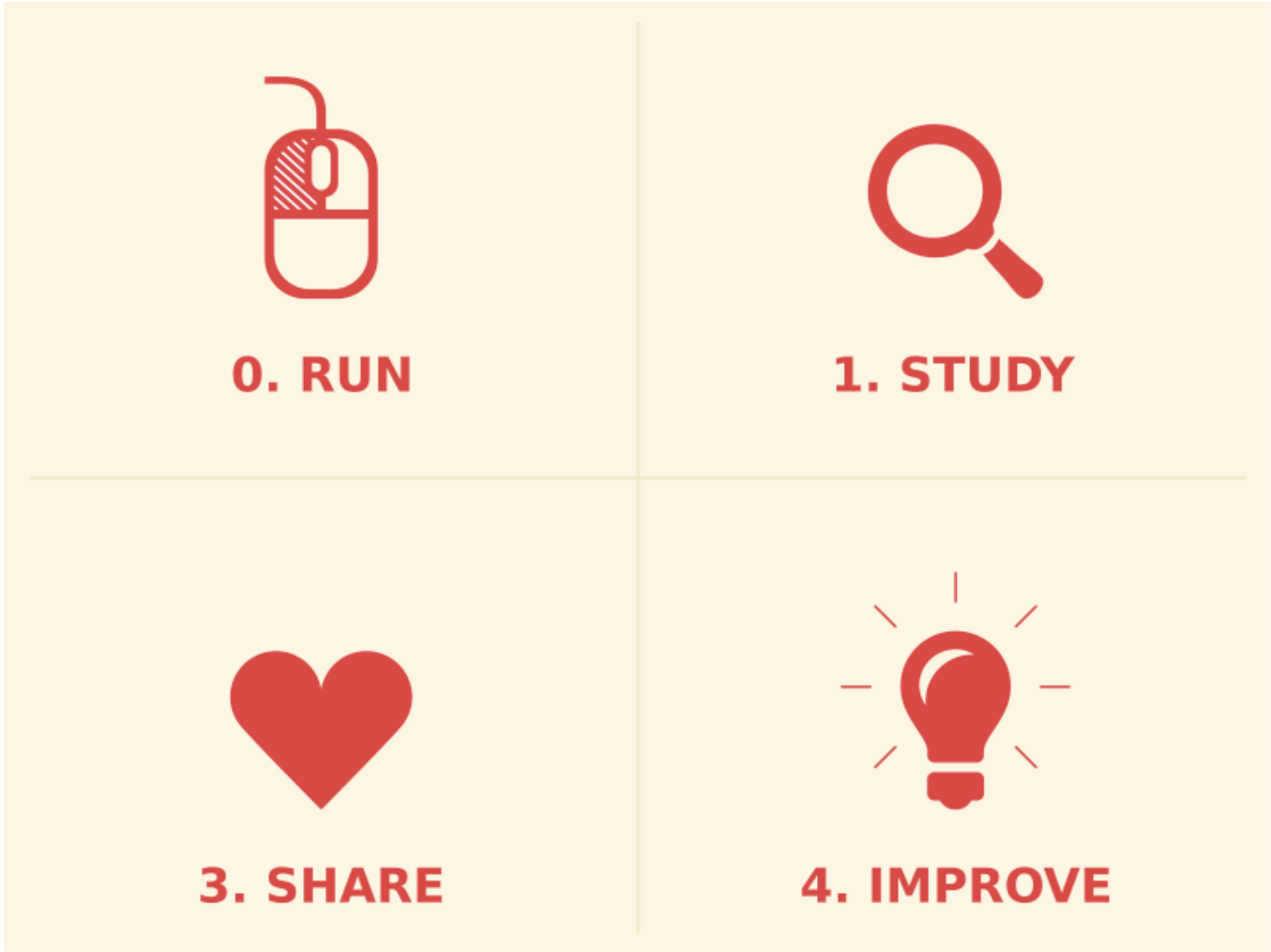
1. Prefiera archivos sin formato commo **txt, csv o tsv**. Si tiene un excel se recomienda transformarlo, particularmente cuando trabaje con miles de filas o columnas.

2. Ojo con separador de columnas, decimales y valores perdidos.

```
library(readr)
mouse <- read.csv("Data.csv", header = TRUE,
          sep = ";", dec = ",", na.strings=c(""))
```

PRÁCTICA VARIABLES ALETORIAS

Guía de trabajo programación con R en Rstudio.cloud.



RESUMEN DE LA CLASE

- Identificamos y clasificamos variables.
- 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.
- Reconocemos variables aleatorias discretas y algunas distribuciones de probabilidad asociadas (Bernoulli y Binomial).