

CLASE 02 - VARIABLES ALEATORIAS CONTINUAS

Diplomado en Análisis de Datos y Modelamiento Predictivo con
Aprendizaje Automático para la Acuicultura.

Dra. María Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

08 April 2023

PLAN DE LA CLASE

1. Introducción

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

CONCEPTOS Y DEFINICIONES

1. **Variable**

Características que se pueden medir u observar en un individuo o en un ambiente: peso, temperatura, sexo, crecimiento, madurez, flotabilidad, rendimiento, sobrevivencia, biomasa cosechada.

2. **Variable aleatoria**

Es un número que representa un resultado de un experimento aleatorio. Depende entonces de función matemática o distribución de probabilidad.

CONCEPTOS Y DEFINICIONES

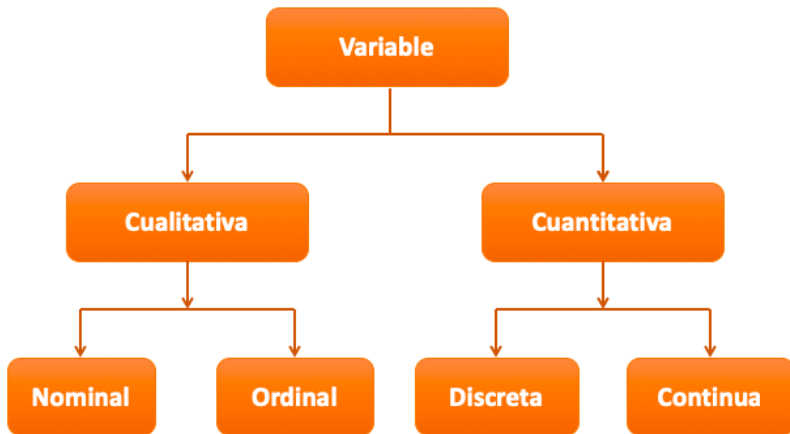
3. Datos u observaciones

Son los valores que puede tomar una variable aleatoria. 6078 gramos, 55 mm, células por mililitro, macho / hembra, 13°C, Maduro /No maduro, Kg wfe/N, vivo/muerto.

4. Factor

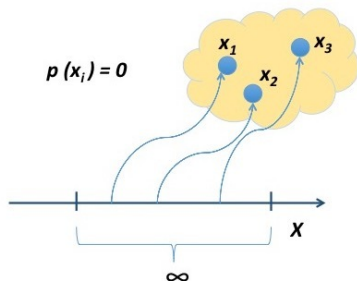
Usado para identificar tratamientos de un experimento o variables de clasificación. Se usan como *variables independientes o predictoras*, es decir, tienen un efecto sobre una *variable dependiente o respuesta*. Ej. Sexo (niveles: macho o hembra) tiene un efecto sobre nivel de hormonas.

CLASIFICACIÓN DE VARIABLES



VARIABLE ALEATORIA CUANTITATIVA CONTINUA

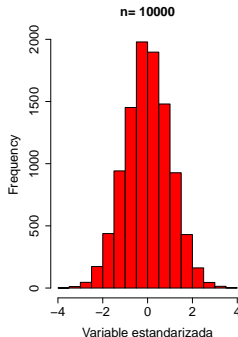
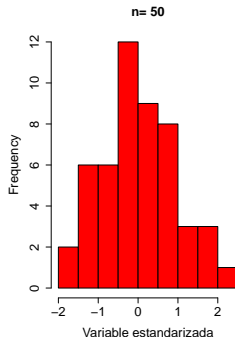
Definición: Puede tomar cualquier valor dentro de un intervalo (a,b) , (a,Inf) , $(-\text{Inf},b)$, $(-\text{Inf},\text{Inf})$ y la probabilidad que toma cualquier punto es 0, debido a que existe un número infinito de posibilidades.



OBSERVAR VARIABLE CONTINUA CON HISTOGRAMA

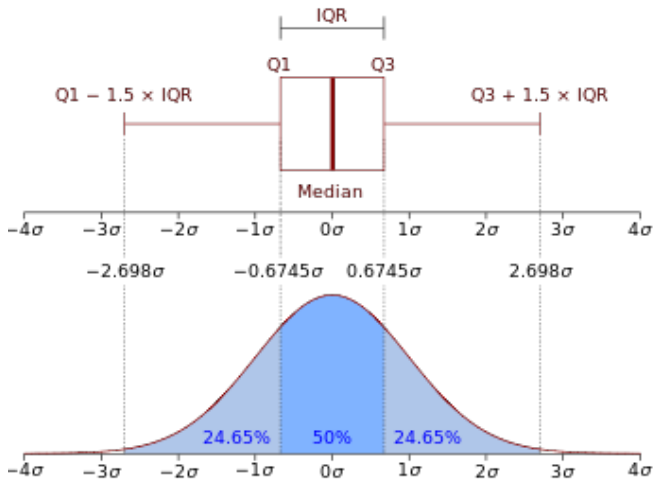
Al observar con un histograma **hist()** notamos que:

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



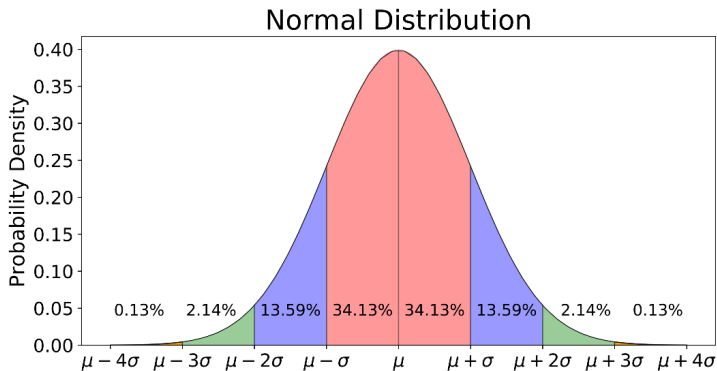
OBSERVAR CON BOXPLOT

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



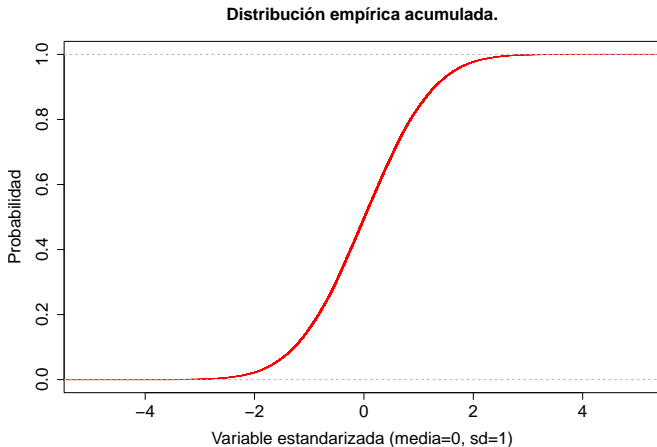
PREDICCIÓN CON DISTRIBUCIÓN NORMAL

- ▶ Si la variable aleatoria tiene una distribución normal, podemos predecir la probabilidad de que la variable tome un determinado valor dentro de un intervalo (ej. 13,59%).

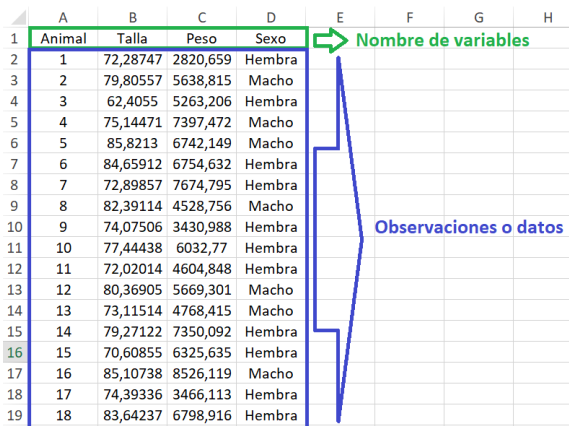


PREDECIR CON DISTRIBUCIÓN ACUMULADA

- La función de distribución empírica acumulada **ecdf()** permite predecir la probabilidad de que la variable aleatoria tome un valor determinado.



FORMATO CORRECTO PARA IMPORTAR A R



	A	B	C	D	E	F	G	H
1	Animal	Talla	Peso	Sexo				
2	1	72,28747	2820,659	Hembra				
3	2	79,80557	5638,815	Macho				
4	3	62,4055	5263,206	Hembra				
5	4	75,14471	7397,472	Macho				
6	5	85,8213	6742,149	Macho				
7	6	84,65912	6754,632	Hembra				
8	7	72,89857	7674,795	Hembra				
9	8	82,39114	4528,756	Macho				
10	9	74,07506	3430,988	Hembra				
11	10	77,44438	6032,77	Hembra				
12	11	72,02014	4604,848	Hembra				
13	12	80,36905	5669,301	Macho				
14	13	73,11514	4768,415	Macho				
15	14	79,27122	7350,092	Hembra				
16	15	70,60855	6325,635	Hembra				
17	16	85,10738	8526,119	Macho				
18	17	74,39336	3466,113	Hembra				
19	18	83,64237	6798,916	Hembra				

Figura 1: Formato correcto de archivo excel para que sea importado a R.

ERRORES EN FORMATO EXCEL

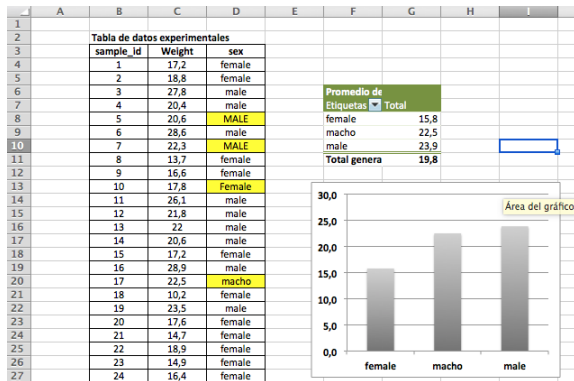


Figura 2: Errores comunes antes de importar a excel.

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	

Figura 3: 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

El paquete **readxl** es muy util para importar datos a R. Pero debe tener cuidado con: separador de columnas, decimales y valores faltantes.

```
library(readxl)
salmon<-read_excel("datos.xlsx",
                   sheet = 1, na = "NA")
```

PRÁCTICA VARIABLES ALEATORIAS

Guía de trabajo programación con R en Rstudio.cloud.



0. RUN



1. STUDY



3. SHARE



4. IMPROVE

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 empírica acumulada.
- ▶ Es importante identificar la naturaleza que tiene nuestra variable en estudio, y así evitar errores en los análisis estadísticos que llevemos a cabo. No siempre será normal.