Estadística I
MGAIE

Prof. Dr. Roberto Muiños

- En nuestra sociedad hay abundancia de información disponible y continuamente debemos leer e interpretar datos.
- Esto nos lleva a la necesidad de aprender a utilizar métodos para extraer conclusiones basadas en los datos

Tres preguntas para responder

- ¿Cómo recolectar los datos?
- ¿Cómo analizar y resumir los datos para producir información o una conclusión?
- ¿Qué grado de confianza puedo tener en mis resultados? o dicho de otro modo ¿Qué exactitud tiene mi resultado?

La estadística provee las herramientas conceptuales y metodológicas para responder estas preguntas

¿Para que sirve la estadística?

- Para organizar, presentar y describir un conjunto de datos

Estadística Descriptiva

- Para poder generalizar los resultados obtenidos en una muestra a la población de la cual se extrajo.

Estadística Inferencial

  - Estimar características poblacionales
  - Probar hipótesis formuladas sobre una población.
  - Construir modelos estadísticos y efectuar predicciones

Análisis estadístico

- Es un proceso total de organización y resumen, procesamiento y obtención de los datos.
- En el análisis estadístico se combinan las metodologías descriptivas y las metodologías inferenciales.

Datos, variables y escalas de medición

- Una unidad de observación o unidad experimental es aquella sobre la cual se efectúan mediciones o se intenta clasificar en categorías.

Las unidades de observación pueden ser personas o grupos de personas, viviendas, etc

- En el proceso de observación se registra para cada unidad experimental alguna característica y esto constituye un DATO.

Por ejemplo: Si el objetivo de una investigación consiste en realizar un estudio sobre ingresos familiares, la unidad de observación podría ser la familia. El ingreso familiar medido es un dato.

Variables

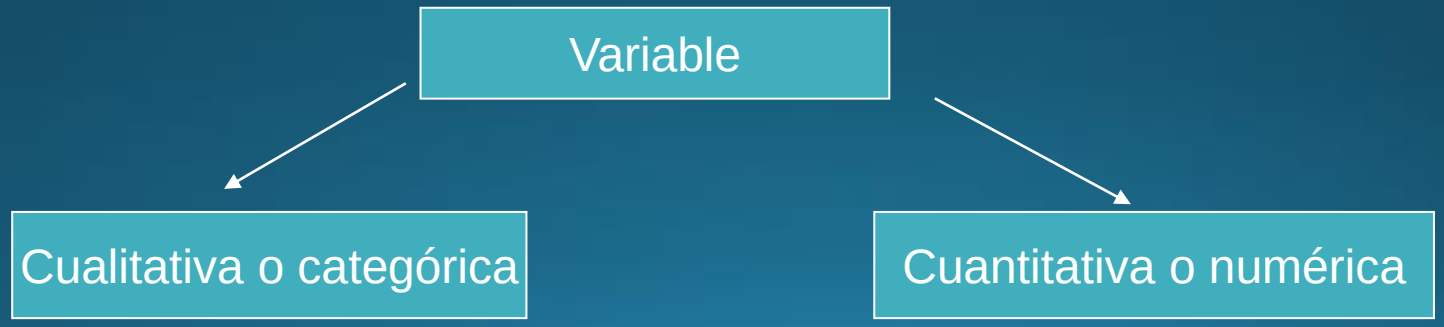
- Una variable es la medición de cualquier característica que varía de una unidad experimental a otra en la población o en una muestra.

- Propiedades de
  - Exhaustividad
  - Exclusividad

Tipos de variables

- Numéricas
- Alfanuméricas

- Las variables numéricas pueden clasificarse en dos grandes grupos.



Tipos de variables

- Las variables categóricas son aquellas que están definidas por las clases o categorías que la componen

Ejemplo:

- Las personas pueden ser clasificadas de acuerdo al color de sus cabellos: rubias, morochas o pelirrojas.
- Las rocas pueden ser clasificadas en sedimentarias, ígneas o metamórficas

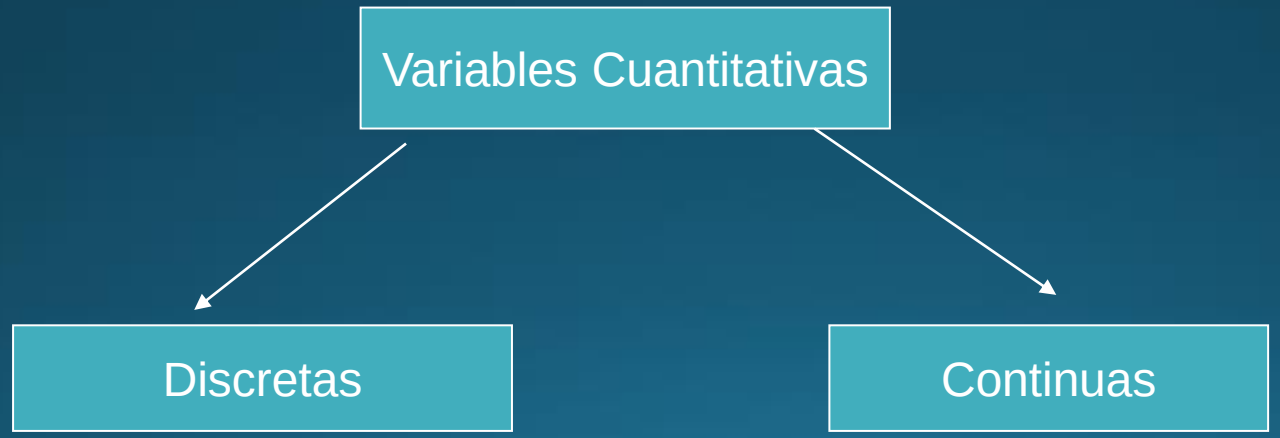
La clasificación mas simple para este tipo de variable es aquella que tiene solo dos categorías. La caracterización se debe a la presencia o ausencia de una cualidad dada. Estas variables se llaman DICOTÓMICAS O BINARIAS

Ejemplo:

- Una persona puede o no estar empleada.
- Una planta puede o no tener flores

Tipos de variables

- Las variables cuantitativas son aquellas que se expresan por un número implica cantidad. Las variables cuantitativas se clasifican en:



Tipos de variables

- Las variables discretas son aquellas que surgen como conteos o por asignación de ciertos códigos numéricos a las categorías de las variables cualitativas.

Ejemplos: Cantidad de árboles frutales, cantidad de alumnos en un curso, cantidad de personas por vivienda, etc

- Las variables continuas son aquellas que surgen de mediciones efectuadas sobre cada unidad experimental. Estas variables pueden tomar infinitos valores en un rango dado.

Ejemplo: Longitud de las espigas de trigo, altura de un individuo, temperatura media mensual, etc.

Escalas de Medición

Las variables estadísticas también pueden clasificarse según su nivel de medición

- Nominales
- Ordinales
- De intervalo
- De razón

La aplicación de los métodos estadísticos depende del nivel de medición de las variables!!!

Escalas de Medición

NOMINALES

Es el nivel más primitivo

Con él se clasifican objetos, personas, características.

Ejemplos

- Lugar de nacimiento
- Estado civil
- Tipo de religión

Escalas de Medición

ORDINALES

Las categorías de las variables poseen una relación de orden.

Ejemplos

- Nivel de educación
- Nivel socioeconómico
- Grado militar

Escalas de Medición

DE INTERVALO

El cero de estas variables es arbitrario.

Las diferencias pueden compararse

Ejemplos

- Temperatura
- Longitud geográfica
- Calendario

Escalas de Medición

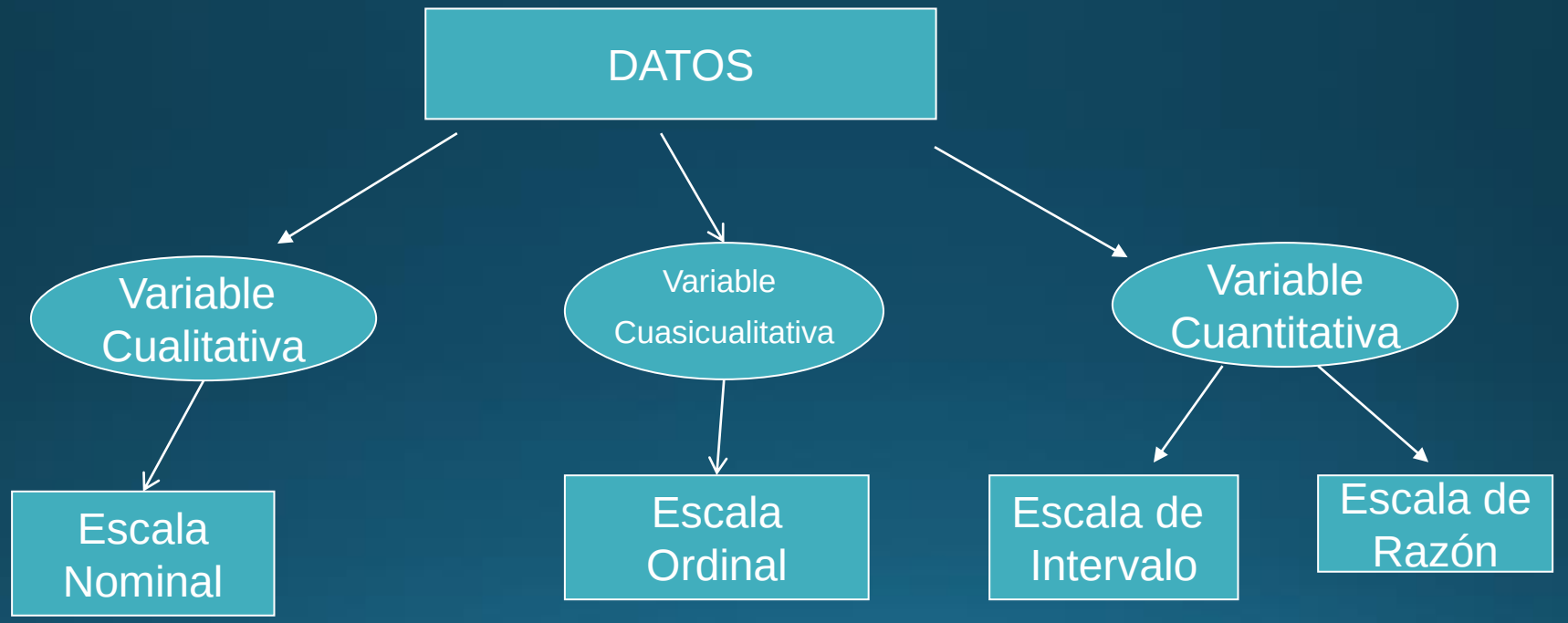
DE RAZON

El cero de estas variables significa "ausencia" de lo que se está midiendo.

Ejemplos

- Peso
- Altura
- Distancia

Resumiendo....


DATOS
Variable Cualitativa
Variable Cuasicualitativa
Variable Cuantitativa
Escala Nominal
Escala Ordinal
Escala de Intervalo
Escala de Razón

Organización de los datos.

- Para su análisis estadístico, los datos deben estar disponibles en una **tabla de datos**
- Los softwares estadísticos trabajan con tablas de datos
  - Pueden generarlas directamente
  - Pueden exportarlas de otros softwares, por ejemplo, Excel, Spss, etc.

Tabla de datos

Las filas corresponden a los individuos

C:/Users/Roberto/OneDrive/doctorado2016/DATOS_Y_ANALISIS - RStudio Source Editor

vgrales

Filter

	HCL	Modo	Genero	Edad	Ocupacion	EstadoCivil	MotivoConsulta	Terapeuta	sexo
1	155218	Caduco	F	52	Asistente terapéutico	Soltero/a	Trastorno de ansiedad	Sansone, Maria Solange	1
2	201579	Caduco	M	40	contador publico	Soltero/a	Problemas de pareja	Melchiorre, Roberto	0
3	237013	Caduco	M	47	Comerciante	Soltero/a	Otros	Tommasi, Mauro	0
4	255134	Respondido	F	35	Administrativa	Soltero/a	Otros	Bedecarats, Jesica Marlene	1
5	265775	Caduco	F	44	contadora pública	Soltero/a	Otros	Braccia, Laura Alejandra	1
6	267513	Caduco	M	31	Empleado administrativo	Casado/a	Otros	Pellegrini, Paula	0
7	268144	Caduco	M	26	empleado administrativo	Soltero/a	Otros	Mascolo, Martha Lidia	0
8	271191	Caduco	F	54	empleada de galeno	Divorciado/a	Problemas de familia	Donato, Tomas	1
9	271258	Respondido	M	55	Procurement Marketing	Concubino/a	Trastornos somaticos y stress	Lazzati, Alejandra	0
10	278063	Caduco	M	30	Administrativa	Soltero/a	Otros	Rosarios, Lucia	0
11	279188	Respondido	M	36	Director de cine	Casado/a	Otros	Pando, Manuel Matias	0
12	281313	Caduco	F	31	empleada administrativa	Soltero/a	Otros	Adler, Julieta Eliana	1
13	281686	Respondido	F	33	maestra	Concubino/a	Problemas de pareja	Adanez, Paz	1
14	282494	Caduco	F	26	trabajadora social	Soltero/a	Trastorno del estado de animo	Mascolo, Martha Lidia	1
15	286034	Caduco	F	25	Empleada Administrativa	Soltero/a	Trastorno del estado de animo	Gulisano, Jeronimo	1
16	287581	Respondido	F	60	Medica	Divorciado/a	Problemas de familia	Yañez, Fabiana	1
17	287784	Caduco	M	29	Empleado	Soltero/a	Otros	Mendez, Lucía	0
18	288413	Respondido	F	23	Estudiante	Soltero/a	Otros	Ormaechea, Valeria	1
19	289119	Respondido	M	27	Ingeniero en petroleo	Soltero/a	Problemas de pareja	Larcapide, Hernan	0
20	289834	Caduco	F	21	estudiante	Soltero/a	Otros	Pinieri Amato, Antonela	1
21	289839	Caduco	M	21	estudiante de cine	Soltero/a	Trastorno de ansiedad	Katz, Ezequiel	0

Showing 1 to 25 of 535 entries, 9 total columns

Las columnas corresponden a las variables

Tabla de datos

- Las variables deben tener un nombre
- Las variables numéricas se pueden analizar estadísticamente
- Si una variable no es numérica se debe codificar
- Los individuos deben estar identificados unívocamente

Codificación de variables

- Cada modalidad debe tener un código diferente
- Debe haber un código de dato faltante
- Debe haber un código de respuesta múltiple
- Estos códigos deben estar fuera del rango de valores de la variable

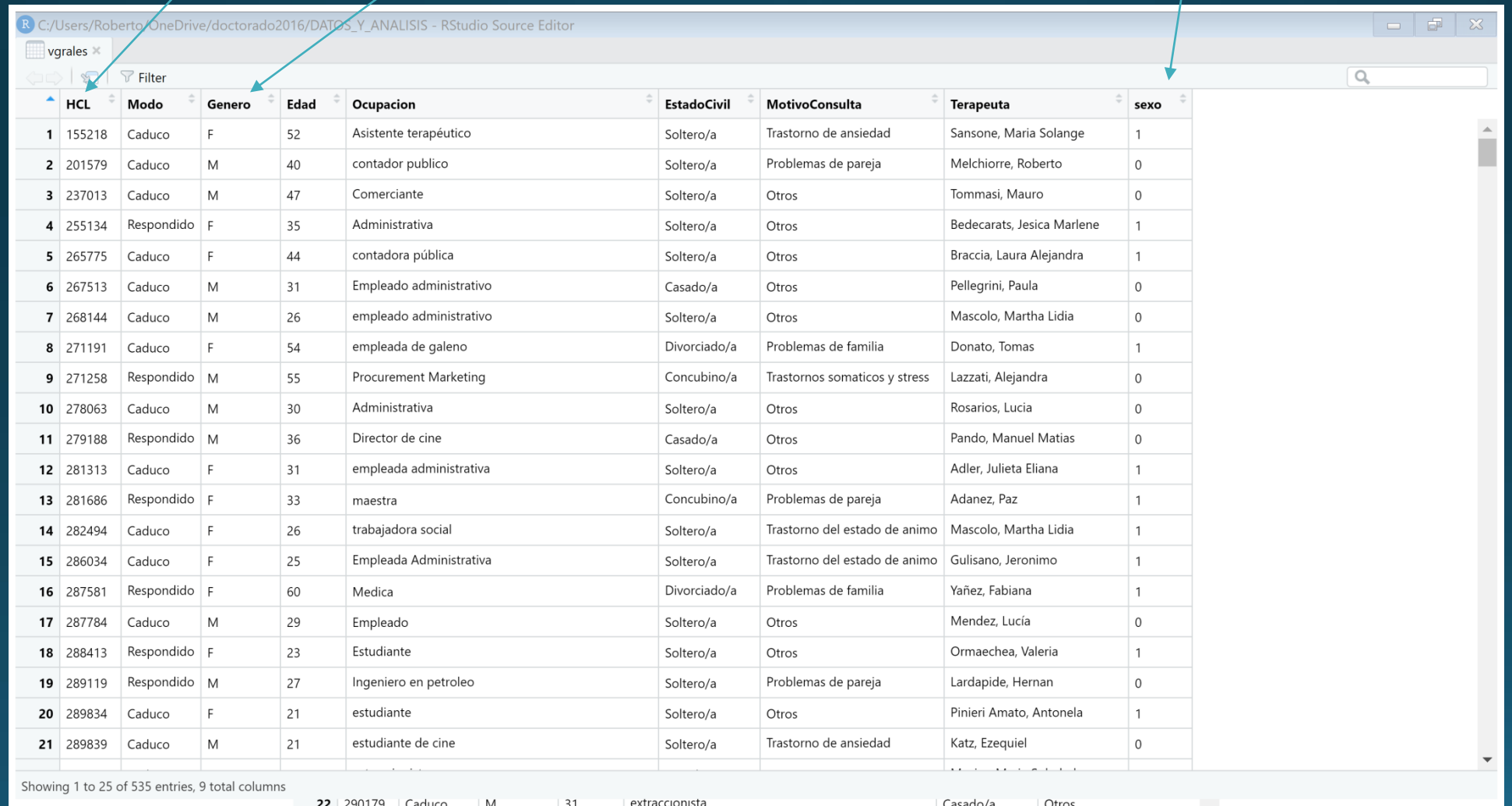
Identificación de casos

- Cada individuo debe tener una identificación preferentemente numérica
- Esa identificación debe ser unívoca
- Se pueden tener otras variables identificatorias no numéricas adicionales

Variable sin codificar

Variable codificada

Identificador



	HCL	Modo	Genero	Edad	Ocupacion	EstadoCivil	MotivoConsulta	Terapeuta	sexo
1	155218	Caduco	F	52	Asistente terapéutico	Soltero/a	Trastorno de ansiedad	Sansone, Maria Solange	1
2	201579	Caduco	M	40	contador publico	Soltero/a	Problemas de pareja	Melchiorre, Roberto	0
3	237013	Caduco	M	47	Comerciante	Soltero/a	Otros	Tommasi, Mauro	0
4	255134	Respondido	F	35	Administrativa	Soltero/a	Otros	Bedecarats, Jesica Marlene	1
5	265775	Caduco	F	44	contadora pública	Soltero/a	Otros	Braccia, Laura Alejandra	1
6	267513	Caduco	M	31	Empleado administrativo	Casado/a	Otros	Pellegrini, Paula	0
7	268144	Caduco	M	26	empleado administrativo	Soltero/a	Otros	Mascolo, Martha Lidia	0
8	271191	Caduco	F	54	empleada de galeno	Divorciado/a	Problemas de familia	Donato, Tomas	1
9	271258	Respondido	M	55	Procurement Marketing	Concubino/a	Trastornos somaticos y stress	Lazzati, Alejandra	0
10	278063	Caduco	M	30	Administrativa	Soltero/a	Otros	Rosarios, Lucia	0
11	279188	Respondido	M	36	Director de cine	Casado/a	Otros	Pando, Manuel Matias	0
12	281313	Caduco	F	31	empleada administrativa	Soltero/a	Otros	Adler, Julieta Eliana	1
13	281686	Respondido	F	33	maestra	Concubino/a	Problemas de pareja	Adanez, Paz	1
14	282494	Caduco	F	26	trabajadora social	Soltero/a	Trastorno del estado de animo	Mascolo, Martha Lidia	1
15	286034	Caduco	F	25	Empleada Administrativa	Soltero/a	Trastorno del estado de animo	Gulisano, Jeronimo	1
16	287581	Respondido	F	60	Medica	Divorciado/a	Problemas de familia	Yañez, Fabiana	1
17	287784	Caduco	M	29	Empleado	Soltero/a	Otros	Mendez, Lucía	0
18	288413	Respondido	F	23	Estudiante	Soltero/a	Otros	Ormaechea, Valeria	1
19	289119	Respondido	M	27	Ingeniero en petroleo	Soltero/a	Problemas de pareja	Lardapide, Hernan	0
20	289834	Caduco	F	21	estudiante	Soltero/a	Otros	Pinieri Amato, Antonela	1
21	289839	Caduco	M	21	estudiante de cine	Soltero/a	Trastorno de ansiedad	Katz, Ezequiel	0

Showing 1 to 25 of 535 entries, 9 total columns

22 290179 Caduco M 31 extraccionista Casado/a Otros

Documentación de las tablas de datos

- Las tablas de datos deben acompañarse de un archivo que para cada variable
  - Nombre
  - Descripción
  - Nivel de medición
  - Código de dato faltante
  - Código de dato duplicado
  - Descripción y Código de cada modalidad

Ejemplo

Supongamos que tenemos los datos de 5 sujetos medidos en 4 variables:

Sujeto	Sexo	Inteligencia	Nivel cultural	Estrés
1	0	101	2	4
2	1	83	1	5
3	1	95	2	6
4	0	89	1	4
5	0	107	2	7

- Variable: Sexo
- Nivel de medición: nominal
  - 0 : mujer
  - 1 : varón
  - 9 : dato faltante o no respuesta
  - 8 : respuesta múltiple

Ejemplo

Supongamos que tenemos los datos de 5 sujetos medidos en 4 variables:

Sujeto	Sexo	Inteligencia	Nivel cultural	Estrés
1	0	101	2	4
2	1	83	1	5
3	1	95	2	6
4	0	89	1	4
5	0	107	2	7

- Variable: Nivel cultural
- Nivel de medición: nominal
  - 1 : Bajo
  - 2 : Alto
  - 9 : dato faltante o no respuesta
  - 8 : respuesta múltiple

Introducción al Software Estadístico R

Software R

R es un software de libre uso y distribución para programar análisis estadístico y gráfico

fue creado en 1993 en el Departamento de Estadística de la Universidad de Auckland-Nueva Zelanda

desde 1997 se desarrolla con aportes de diversas partes del mundo, bajo la coordinación de un equipo principal de desarrollo (R Core Team Development) (R Project).

R funciona con paquetes de programación

están disponibles en una Red de Archivos R (Comprehensive R Archive Network, CRAN) en sitios web llamados MIRROR desde los cuales los usuarios finales pueden descargarlos.

Diferencias entre R y otros softwares

En general

Tengo una tabla de datos

Elijo un procedimiento y me sale en otra ventana una importante salida de resultados que, inclusive, puede ampliarse.

En R, tengo un espacio de trabajo con objetos (variables, funciones, tablas de datos, etc.)

Aplico funciones a los objetos y obtengo otros objetos que, a su vez, pueden generar objetos adicionales

Los resultados de la aplicación de las funciones, son otros objetos

Diferencias entre SPSS y R

La asistencia en R

No es tan asistida como en otros softwares

La ayuda debe buscarse permanentemente

En otros softwares no es necesario tener siempre perfectamente claro todo lo que quiero hacer.

En R, es indispensable.

Software R

El paquete de instalación de R, nos permite realizar análisis estadísticos y gráficos básicos

para realizar otros más complejos es necesario instalar paquetes adicionales

Objetos y Funciones

Objetos

nombre_objeto <- valor

Los nombres de los objetos deben comenzar con una letra y solo pueden contener letras, números, _ y .

Es mejor que los nombres sean descriptivos.

Si se necesita usar más de una palabra, usar guion_bajo o punto.

Funciones

nombre_funcion(arg1 = val1, arg2 = val2, ...)

Las funciones siempre tienen paréntesis ()

Pueden o no tener argumentos que completar

Principales objetos en R

- Vectores
- Matrices
- Data frames
- Factores
- Funciones

Vectores

- Corresponden a un conjunto de valores considerados como una única entidad

Ejemplo:

La función c() para generar vectores

- x <- c(2, 4, 6)
- y <- c(3, 7)
- c(x, y)
- Resultado: 2 4 6 3 7

Matrices

- Las matrices son también vectores pero con dos atributos adicionales: número de filas y número de columnas.

- matrix(1:3, nrow = 2, ncol = 3)

```
##      [,1] [,2] [,3]
## [1,]    1    3    2
## [2,]    2    1    3
```

- El primer argumento es un vector que incluye los elementos de la matriz y los otros dos corresponden al número de filas y columnas. Los elementos se van colocando por columnas. Observar cómo se ha aplicado la regla del reciclaje en el ejemplo anterior.

Data frames

- Un data frame es la estructura que se usa en R para las tablas de datos correspondientes a una serie de variables medidas en cada sujeto o unidad bajo estudio.

- Para crearlos se utiliza la función data.frame().
- Ejemplo

```
x <- 7:9                    # Se crea un vector con los valores 7, 8 y 9
y <- c("a", "b", "c")       # Se crea un vector con los valores a, b, c
mifichero <- data.frame(edad = x, grupo = y)    # se crea un data frame con dos variables, edad y grupo

mifichero
##   edad grupo
## 1   7    a
## 2   8    b
## 3   9    c
```

Factores

Los factores son las estructuras que se utilizan para manejar las variables cualitativas en los análisis estadísticos

Se crean usando la función factor()

Funciones

- Corresponden a comandos que pueden realizar determinadas acciones
- Se reconocen por tener () **en todos los casos**
- Tienen argumentos que pueden ajustar su funcionamiento
- Existen centenares de funciones disponibles
- Pero también pueden ser generadas por los usuarios

Los paquetes en R

Un paquete es una colección de funciones, datos y documentación que permite extender las capacidades del R base.

Los paquetes son clave para usar R de manera exitosa.

Los paquetes deben ser cargados al inicio del script

Sintaxis de instalación

install.packages("*nombre del paquete*")

Scripts

Es una secuencia de comandos que se guardan como un archivo reutilizable con mínimas modificaciones

Se trata de un archivo de texto por lo que puede ser editado

Se recomienda armar Scripts con toda la secuencia de actividades a realizar e incluir comentarios que faciliten la lectura

Incluir:

- La carga de los paquetes a utilizar
- La instancia de importación de los datos
- Las acciones de acondicionamiento de la tabla de datos
- Los procesos estadísticos a realizar

Ejemplo de Script en RStudio

PRUEBA - RStudio Source Editor

cigars.R

Source on Save

```
1  library(tidyverse)
2  library(rstatix)
3  library(olsrr)
4
5  library(haven)
6  p081 <- read_sav("p081.sav")
7  View(p081)
8
9  model <- lm(SALES ~ AGE+HS+INCOME+BLACK+FEMALE+PRICE , data = p081)
10 summary(model)
11
12 # CORRELACION DE CADA PREDICTOR CON LA DEPENDIENTE
13 ols_correlations(model)
14
15
16 # DIAGNOSTICOS DE COLINEALIDAD
17 ols_vif_tol(model)
18
19 # LOS MEJORES MODELOS DE CADA TAMAÑO
20 ols_step_best_subset(model)
21
22 kkk <- ols_step_both_p(model)
23 plot(kkk)
24 kkk
25
26 model1 <- lm(SALES ~ INCOME+PRICE , data = p081)
```

RStudio para el uso de R

RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux).[3] RStudio tiene la misión de proporcionar el entorno informático estadístico R. Permite un análisis y desarrollo para que cualquiera pueda analizar los datos con R.

CLASE2 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | CLASE2

ejercicio_clase.R | bajopeso.R | LOWBWT | maschico | fumadores | new.R

Source on Save | Run | Source

```
1  # libreria que contiene la función que importa archivos excel
2  library(readxl)
3
4  # importación de la tabla de datos
5  ejercicio <- read_excel("bajopesoalnacer.xlsx")
6
7  str(ejercicio)
8
9  #para ver el archivo de datos
10 View(bajopesoalnacer)
11
12 # etiquetas y transformación en factores de las cualitativas
13
14 ejercicio$Sexo <- factor(ejercicio$Sexo,labels=c("F","M"))
15
16
```

1:1 (Top Level) R Script

Environment | History | Connections | Tutorial

Import Dataset | 137 MiB | List

R | Global Environment

Data

fumadoras_mas...	74 obs. of 3 variables
fumadores	74 obs. of 11 variables
LOWBWT	189 obs. of 14 variables
maschico	189 obs. of 3 variables

Console | Terminal | Background Jobs

R 4.2.2 · C:/Users/rmuin/OneDrive/CLINICAS_2021/CLASE2/

```
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from C:/Users/rmuin/OneDrive/CLINICAS_2021/CLASE2/.RData]

>
```

Files | Plots | Packages | Help | Viewer | Presentation

Folder | Blank File | Delete | Rename

C: > Users > rmuin > OneDrive > CLINICAS_2021 > CLASE2

Name	Size	Modified
..		
.RData	6.5 KB	Mar 15, 2023, 4:22 PM
.Rhistory	19.7 KB	Mar 15, 2023, 4:22 PM
anova2f.R	930 B	Mar 15, 2023, 1:18 PM
bajopeso.R	1.6 KB	Mar 15, 2023, 3:13 PM
BD diabetes y nutricion.xls	541.5 KB	Aug 12, 2021, 6:56 PM
bland.R	900 B	Oct 22, 2021, 2:11 PM

Buscar ESP 14:07 16/3/2023 36

Proyectos de trabajo en R

Creación de un Proyecto en RStudio

- Crear una carpeta
- Poner en ella los archivos de datos que se utilizarán
- Crear un proyecto en Rstudio asociado a dicha carpeta
- Todo lo que se genere en las sesiones de trabajo de dicho Proyecto se guardará en la carpeta asociada