UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

MANUAL DE PRACTICAS DE LA MATERIA

ESTADISTICA AVANZADA

clave 40004

AUTOR

Dr. JUAN IVAN NIETO HIPÓLITO

2022-2

PRÁCTICA 1. Instalación de R y Rstudio

Instrucciones de laboratorio de la práctica 1: RStudio

1) **Instale R y RStudio:**

Primero, deberá instalar R y RStudio. R es el nombre del propio lenguaje de programación y RStudio es una interfaz conveniente.

- *Instale R: vaya a https://cran.r-project.org/ y siga el enlace para su sistema operativo.*
- *Instale RStudio: Vaya a https://www.rstudio.com/products/rstudio/download/ y haga clic en el enlace del instalador para su sistema operativo.*

El siguiente video narra las instrucciones de instalación paso a paso si es necesario.

- *https://www.youtube.com/watch?v=eD07NznguA4*

NOTA: Si ya tiene instalados R y RStudio, aún debe visitar estos enlaces para confirmar que tiene las versiones más recientes de este software. La versión más reciente de R se puede encontrar en la página The R Project for Statistical Computing y la versión más reciente de RStudio se puede encontrar en la página de descarga de RStudio. Instale las versiones más recientes antes de continuar.

2) **Instalar paquetes R:**

Instalar y cargar herramientas de desarrollo:

Usaremos el paquete ***devtools*** para posteriormente instalar el paquete ***statsr*** el cual incluye las funciones estadísticas que usaremos en este curso. Inicie RStudio e ingrese los siguientes comandos en la Consola:

```
install.packages("devtools")

library(devtools)
```

Enseguida instale el resto de los paquetes que usaremos en el curso. Escriba los siguientes comandos en la consola:

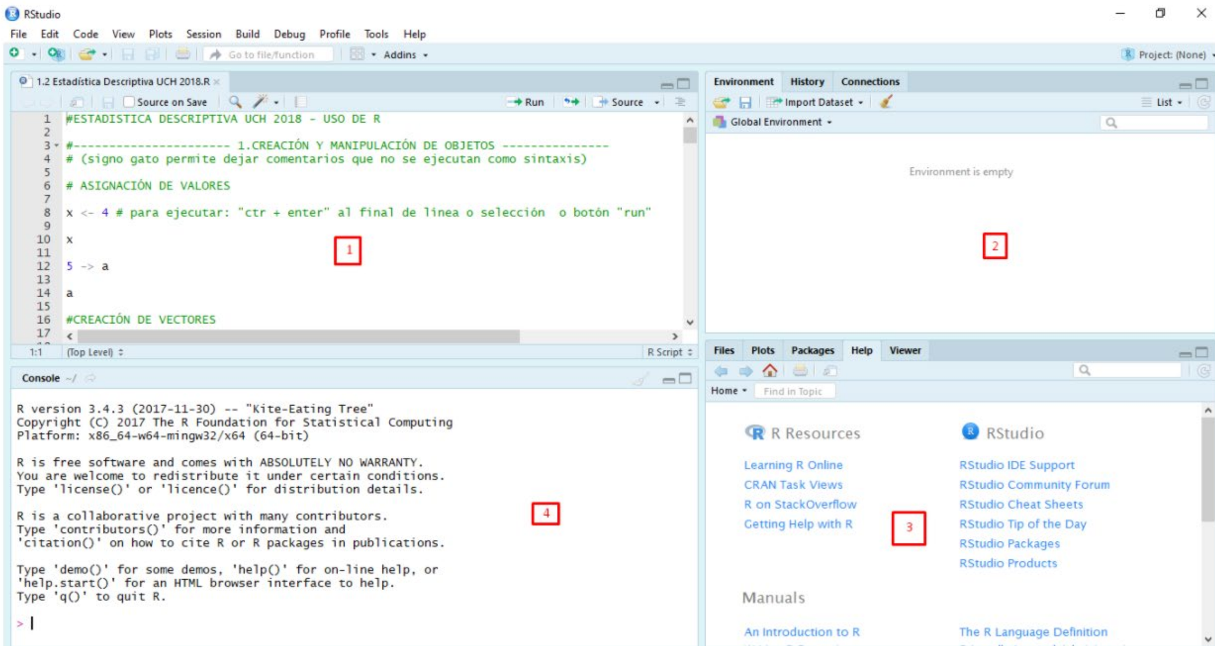
```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("shiny")
```

3) Consulte la información relativa a los paquetes instalados en los puntos 1 y 2.

4) Pruebe su instalación ejecutando los comandos (ejemplos) de los capítulos 2 y 3 del documento “**R para principiantes, autor Emmanuel Paradis**”.

PRACTICA # 2. INTRODUCCION A R

Rstudio



Ventana (1): es el editor de sintaxis: se trata del lugar donde editamos la sintaxis para posteriormente ejecutarla. Al escribir allí no sucederá nada, a no ser que se apriete algún botón para ejecutar los comandos o la tecla **ctrl+enter**.

Ventana (2): es el "entorno de trabajo" del programa: en este lugar se muestra el conjunto de datos y los "objetos" (resultados, variables, gráficos, etc.) que se almacenan al ejecutar diferentes análisis.

Ventana (3) tiene varias sub pestañas: (i) la pestaña files permite ver el historial de archivos trabajados con el programa; (ii) la pestaña plots permite visualizar los gráficos que se generen; (iii) la pestaña packages permite ver los paquetes descargados y guardados en el disco duro así como gestionar su instalación o actualización; (iv) la ventana help permite acceder al CRAN - Comprehensive R Archive Network; (v) la ventana viewer muestra los resultados al construir reportes mediante funcionalidades tipo rmarkdown.

Ventana (4): es la consola. Corresponde a lo que sería el software R en su versión básica. Allí el software ejecuta las operaciones realizadas desde el editor de sintaxis.

Ejecutar los siguientes comandos en la consola.

Introducción de datos

Para construir un vector se utiliza la sentencia c():

Crea un vector

```
> x<-
c(1.2,2.7,3.2,4.5,3.3,4.4,5.6,7,9,2.3,5,4.3,8.1,3.4,5.1,6.8,1.9,1.
8,3,2.5)

#Despliega los valores asignados a x

(<-, símbolo de asignación, no es igual =)
> x
 [1] 1.2 2.7 3.2 4.5 3.3 4.4 5.6 7.0 9.0 2.3
[11] 5.0 4.3 8.1 3.4 5.1 6.8 1.9 1.8 3.0 2.5

Para agregar comentarios se utiliza #

> length(x)# muestra la longitud del vector x
[1] 20

> y<-exp(x)# asigna los valores exponenciales de x a y

> y #valores exponenciales de x
 [1]    3.320117   14.879732   24.532530
 [4]   90.017131   27.112639   81.450869
 [7]  270.426407 1096.633158 8103.083928
[10]    9.974182  148.413159   73.699794
[13] 3294.468075   29.964100  164.021907
[16]  897.847292    6.685894    6.049647
[19]   20.085537   12.182494

> plot(x,y)#grafica x,y básica

> hist(x) #histograma de los valores de x

> ?plot #muestra la ayuda del comando plot

> ?nombre de la función # muestra la ayuda de la función

> 5*8 #multiplicacion, no asigna el resultados a ninguna variable
[1] 40

> x1 <- 5*8 #asigna el resultado de la multiplicación a x1
> x1 # muestra el valor actualde x1
[1] 40

> exp(1) # operación exponencial
[1] 2.718282

> x2 <-exp(1) #asigna el valor a x2


> #operaciones con variables definidas

> x3<-5
> x
```

```
 [1] 1.2 2.7 3.2 4.5 3.3 4.4 5.6 7.0 9.0 2.3 5.0 4.3 8.1
[14] 3.4 5.1 6.8 1.9 1.8 3.0 2.5
>
> x3
[1] 5
> x^3
 [1]   1.728  19.683  32.768  91.125  35.937  85.184
 [7] 175.616 343.000 729.000  12.167 125.000  79.507
[13] 531.441  39.304 132.651 314.432   6.859   5.832
[19]  27.000  15.625
> x3^3
[1] 125

> sqrt(x3) #raiz cuadrada de x3
[1] 2.236068
```

Introducción de datos con scan:

```
> altura_en_metros <- scan() #se introducen los datos mediante teclado, al final se pulsa dos veces
                                   #enter
1: 1.65
2: 1.52
3: 1.83
4: 1.48
5: 1.7
6: 1.73
7: 1.68
8:
Read 7 items
> altura_en_metros

[1] 1.65 1.52 1.83 1.48 1.70 1.73 1.68
```

Para construir series de valores, por ejemplo los números impares comprendidos entre 1 y 65, se utiliza la función seq().

Con rep() es posible repetir un patrón dado, incluso caracteres:

```
> seq(1,65,2) # crea la secuencia de 1 a 65 en saltos de 2

[1] 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49

[26] 51 53 55 57 59 61 63 65

> rep(98,5) # repite 98, 5 veces
```

```
[1] 98 98 98 98 98
> rep(c("sí","no"),3) # repite 3 veces la secuencia de caracteres si,no
[1] "sí" "no" "sí" "no" "sí" "no"
> rep((si,no),3) # error los caracteres se deben encerrar en “”
Error: unexpected ',' in "rep((si,"
> rep(("si","no"),3)  # error los caracteres se deben asignar a un vector c()
Error: unexpected ',' in "rep(("si","
> rep(c("si","no"),3)  # instrucción correcta
[1] "si" "no" "si" "no" "si" "no"
```

LECTURA DE DATOS DE UN ARCHIVO
Especificar con total exactitud la ruta y el nombre del archivo, todo ello entre comillas. Se debe poner el símbolo / en lugar de \.

```
> datos<-read.table("arbuthnost_data.txt.R") # si el archivo es solo texto
```

También, se puede configurar el directorio de trabajo en Rstudio, al sitio donde se encuentra el archivo a cargar.
En el menú de sesion, configurar el directorio de trabajo.
```
setwd("~/2022/Estadistica Avanzada/Practicas")
```

```
> datos<-read.table("arbuthnost_data.txt") # si el archivo es solo texto
```

Si incluye código se puede cargar y ejecutar simultáneamente desde la ventana 3 al seleccionarlo y dar enter.

Con el comando View(arbuthnot arbuthnost_data.txt.R) se pueden ver los datos y comprobar que se cargaron.

```
> datos<-read.table("arbuthnost_data.txt.R") # se asignan los datos cargados a la variable datos
```

carga un archivo con solo tabulaciones. Los nombres de las variables no deben tener espacios, por eso se han colocado puntos. El argumento header=T significa que la primera línea del marco de datos contiene los nombres de las variables.

```
> Costos<-read.table("Costos.txt",header=T)
> Costos
  Costo.unit Costo.mat Costo.mano.de.obra
1     13.59       87              80
2     15.71       78              95
3     15.97       81             106
4     20.21       65             115
5     24.64       51             128
>
```

read.table("clipboard"), copia los datos desde el clipboard y si es necesario, el argumento header=T.

```
> ejemplo_clipboard <-read.table("clipboard")
> View(ejemplo_clipboard)
> ejemplo_clipboard
    V1
1  1629
2  5218
3  4683
4     2
5  1630
6  4858
7  4457
```

Con attach() las variables son accesibles por su nombre en la sesión de R y con names() se obtiene una lista de ellas:

```
> attach(Costos)
> names(Costos)
[1] "Costo.unit" "Costo.mat" "Costo.mano.de.obra"
> Costo.mat
[1] 87 78 81 65 51
```

Otro modo de acceder a una variable (columna), por ejemplo a "Costo.mat" del marco de datos Costos, consiste en seleccionar la correspondiente columna: Costos$Costo.mat.

El símbolo $ se utiliza, en general, para seleccionar elementos de un objeto.
Si queremos conocer de qué tipo es una variable concreta hacemos

```
> class(Costo.mat)
[1] "numeric"
> #Lo que significa que Costo.mat es un vector
```

Con data.frame se crean marco de datos.

```
> num<-c(1,2,3,4,5,6,7,8,910) # crea un vector con siete elementos
> numcua<-num^2
> numcub<-num^3
> A<-data.frame(num,numcua,numcub) # se crea una matriz o marco de datos de 3 x 7.

> A   # la primer columna indica solo el renglón y no es parte del data frame
   num numcua numcub
1   1      1      1
2   2      4      8
3   3      9     27
4   4     16     64
5   5     25    125
6   6     36    216
7   7     49    343
8   8     64    512
9   9     81    729
10  10   100   1000
```

SELECCIÓN DE ELEMENTOS DE UN OBJETO

str(), que permite conocer cuál es la estructura de un determinado objeto.

```
> str(A)
'data.frame':     10 obs. of  3 variables:
 $ num   : num  1 2 3 4 5 6 7 8 9 10
 $ numcua: num  1 4 9 16 25 36 49 64 81 100
 $ numcub: num  1 8 27 64 125 216 343 512 729 1000
```

A las columnas se puede acceder mediante

```
A$num
A$numcua
A$numcub
```

Seleccionar por partes (renglón y columna completos)

```
> A.nuevo<-A[1:4,] # selecciona los primeros cuatro renglones y columnas correspondientes de A y los asigna
                    #a A.nuevo.
                   #indicar las filas (1 a 4) y todas las columnas (espacio en blanco tras la coma)

> A.nuevo_2_columnas <-A[1:4,1:2]
> A.nuevo_2_columnas
  num numcua
1   1      1
2   2      4
3   3      9
4   4     16
```

Para realizar selecciones condicionales.
> A
num numcua numcub
1 1 1 1
2 2 4 8
3 3 9 27
4 4 16 64
5 5 25 125
6 6 36 216
7 7 49 343
> numcua[num<4] # muestra el contenido de las posiciones 1,2, y 3 de numcua
[1] 1 4 9
> numcub[num==7] # muestra el contenido de la posición 7 de numcub
[1] 343

Ejercicio

Los datos de rbuthnot describen los nacimientos de hombres y mujeres en Londres entre 1629 y 1710. John Arbuthnot (1710) utilizó estos datos de series de tiempo para llevar a cabo la primera prueba de significación conocida.

Variables
año: El año, comprendido entre 1629 y 1710.
niños: Número de bautizos (nacimientos) de hombre.
niñas: Número de bautizos (nacimientos) mujeres.

Con esta base de datos realizar:
1.- Graficar en el eje x el año y en el eje y los nacimientos tanto de hombres como de mujeres.
2.- Agregar una cuarta columna que muestre el total de nacimientos, nombre a esta cuarta columna *total_nacimientos.*
3.- Agregar una quinta columna que muestre la proporción de hombres del total de nacimientos, nombre a esta quinta columna *prop_hombres*.
4.- Agregar una quinta columna que muestre la proporción de mujeres del total de nacimientos, nombre a esta quinta columna *prop_mujeres*.
5.- Agregue una sexta columna con la comparación *prop_hombre>prop_mujeres* (R indicara con True cuando se cierto y False cuando sea negativo, es decir realizara una comparación lógica).

Después de realizar esta comparación, ¿cuál es su conclusión o hallazgo?

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Práctica 3. Intervalos de confianza para medias (muestras grandes y pequeñas).

jinh

PRÁCTICA 3. Intervalos de confianza para medias (muestras grandes y pequeñas)

Procedimiento:

1) Revisar conceptos básicos.
2) Distinguir entre muestras grandes y pequeñas.
3) Establecer procedimiento para la elaboración de intervalo.
4) Construir intervalo de confianza.
5) Presentar en el laboratorio los ejercicios completos en R.

6) Suponiendo que los datos provienen de una distribución normal, resuelva.

a) Una empresa del sector eléctrico fabrica focos que tiene una duración de vida que es aproximadamente normal distribuida con una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una vida media de 780 horas, encontrar un intervalo de confianza del 96% para la media poblacional de todos los focos producidos por esta empresa.

b) Reportar la implementación en R del inciso a).

c) Reportar la implementación en R del diagrama de caja del inciso a).

7) Muchos pacientes cardíacos usan un marcapasos implantado para controlar los latidos de su corazón. Un módulo conector de plástico se monta en la parte superior del marcapasos. Suponiendo una desviación estándar de 0.0015 pulgadas y una distribución aproximadamente normal, encuentre un intervalo confianza del 95% para la media de las profundidades de todos los módulos conectores fabricados por una determinada empresa. Una muestra aleatoria de 75 módulos tiene una profundidad promedio de 0.310 pulgadas. Repetir los incisos a, b, y c del problema 1.

8) Los precios de una determinada variedad de arroz, por kilogramo, recolectados de 48 tiendas en Ensenada varían con una media de $3 y una desviación estándar de $1.6.

(a) Construya un intervalo de confianza del 95% para el precio medio.

(b) Con un 95% de confianza, ¿qué podemos afirmar acerca de la tamaño posible de nuestro error si estimamos la media precio del arroz para todas las tiendas en el área como $3?

c) Reporte su implementación en R.

Práctica 1 - Intervalos de confianza para medias (muestras grandes y pequeñas).

Estimación por intervalo: Una estimación por intervalo de un parámetro de la población es un intervalo de forma $O_L < O < O_U$, donde O_L y O_U dependen del valor estadístico para una muestra específica y también la distribución de muestreo de O

Donde O_L y O_U se denominan límites de confianza inferior y superior.

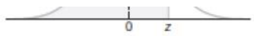
$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha \qquad 0 < \alpha < 1$$

- Cuando $\alpha = 0.05$, tenemos un intervalo de confianza del 95%,
- Cuando $\alpha = 0.01$, obtenemos un intervalo de confianza mas amplio del 99%.
- Cuanto más amplio sea el intervalo de confianza, mas confiaremos en que contiene el parámetro desconocido.
- Es mejor tener un 95% de confianza en que la vida promedio de aparato electrónico esta entre los 6 y los 7 años, que tener un 99% de confianza en que este entre los 3 y los 10 años.
- De manera ideal, se prefiere un intervalo corto con un grado de confianza alto.

Obtener el valor del a:

Para obtener el valor de Alfa necesitamos buscarlo a través de las tablas de Áreas bajo la curva normal.

Tabla A.3 Áreas bajo la curva normal



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Para muestras pequeñas:

Se le considera una muestra pequeña a una muestra de n<=29 elementos

Si se conoce la desviación estándar poblacional, se utiliza:

$$\mu : \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

sino se conoce, entonces:

$$\mu : \bar{x} \pm t \frac{s}{\sqrt{n}}$$

Para muestras grandes:

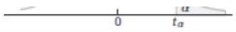
Se le considera una muestra pequeña a una muestra de n>=30 elementos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Para cuando se desconoce la varianza: En este caso se utiliza la distribucion *t* student con *n-1* grados de libertad.

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Tabla A.4 Valores críticos de la distribución *t*



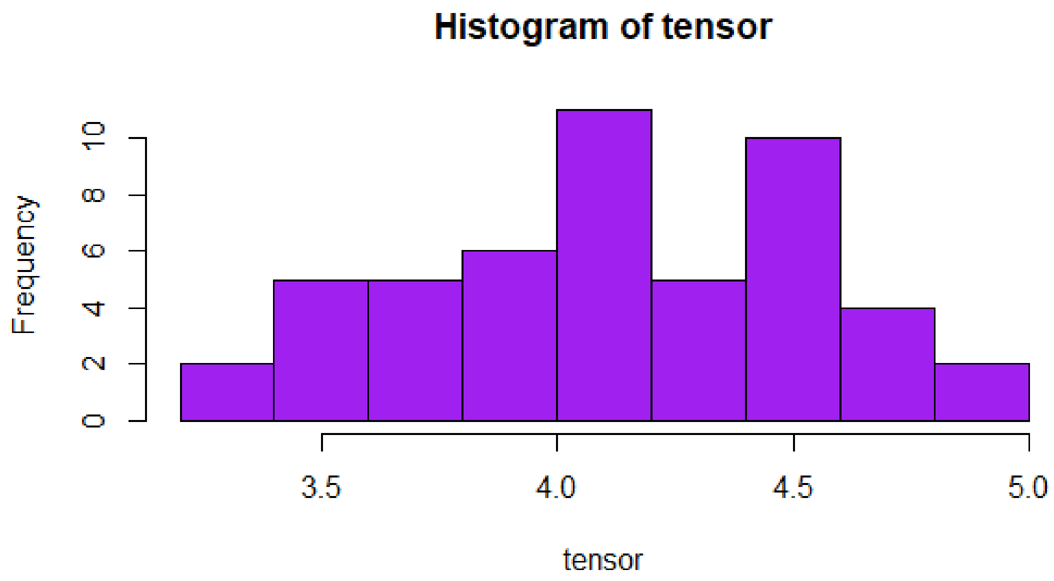
	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

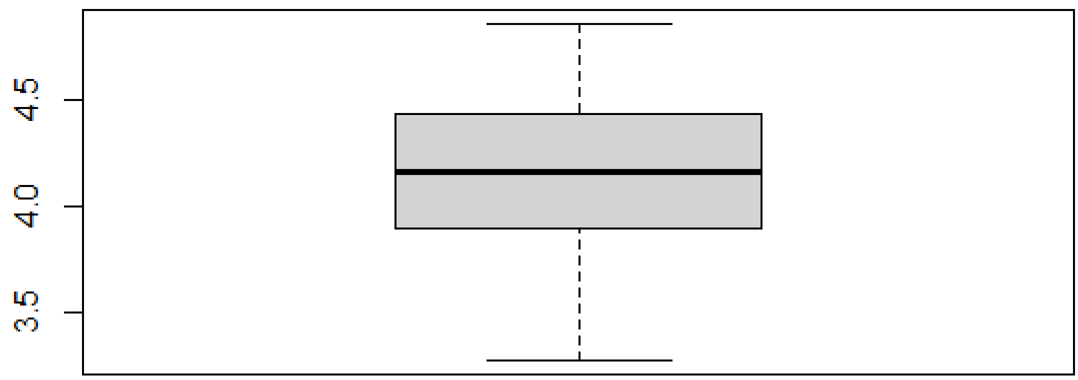
Ejercicio:

```r
datos<-read.table("tension_de_rotura.txt")
class(datos)
str(datos)

tensor<-datos$V1
tensor

lenght(tensor)
sort(tensor)
sqrt(50)                      #el 7 va a ser el num de intervalos
(max(tensor)-min(tensor))/7 #el resultado es 0.227 pero tomaremos 0.3
#se crea una variable límites para utilizarlo en el histograma
#incluye los valores min y max de tensor en saltos de 0.3
limites<- c(3, 5.1, 0.3)

hist(tensor, col= "purple")
hist(tensor, limites, col= "purple")

boxplot(tensor)

summary(tensor)

#tabla de frecuencias
#numero de intervalos
```

```
> datos<-read.table("tension_de_rotura.txt")
> class(datos)
[1] "data.frame"
> str(datos)
'data.frame':	50 obs. of  1 variable:
 $ V1: num  4.05 4.58 4.42 4.2 4.41 4.64 4.76 4.58 3.95 4.17 ...
>
> tensor<-datos$V1
> tensor
 [1] 4.05 4.58 4.42 4.20 4.41 4.64 4.76 4.58 3.95 4.17 4.56 3.51 3.27 3.80 3.59 4.70
[17] 3.77 3.80 4.27 3.94 3.96 4.86 4.39 4.04 4.36 3.72 4.00 3.46 4.01 4.08 3.40 3.89
[33] 4.46 4.38 4.41 4.33 4.16 4.58 4.03 3.76 4.05 4.17 4.46 3.60 4.76 3.99 4.43 4.15
[49] 3.54 4.84
```

```
Error in lenght(tensor) : could not find function "lenght"
> sort(tensor)
 [1] 3.27 3.40 3.46 3.51 3.54 3.59 3.60 3.72 3.76 3.77 3.80 3.80 3.89 3.94 3.95 3.96
[17] 3.99 4.00 4.01 4.03 4.04 4.05 4.05 4.08 4.15 4.16 4.17 4.17 4.20 4.27 4.33 4.36
[33] 4.38 4.39 4.41 4.41 4.42 4.43 4.46 4.46 4.56 4.58 4.58 4.58 4.64 4.70 4.76 4.76
[49] 4.84 4.86
> sqrt(50)                      #el 7 va a ser el num de intervalos
[1] 7.071068
> (max(tensor)-min(tensor))/7 #el resultado es 0.227 pero tomaremos 0.3
[1] 0.2271429
> #se crea una variable límites para utilizarlo en el histograma
> #incluye los valores min y max de tensor en saltos de 0.3
> limites<- c(3, 5.1, 0.3)
```


Histogram of tensor
Frequency
0
2
4
6
8
10
3.5
4.0
4.5
5.0
tensor


3.5
4.0
4.5

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Práctica 4. Intervalos de confianza para proporciones y varianza.

Procedimiento:

1. Revisa conceptos básicos.
2. Diferencia entre proporciones y varianzas.
3. Establece procedimiento para la elaboración del intervalo correspondiente.
4. Construye intervalo de confianza.
5. Presentar en el laboratorio los ejercicios completos en R.

jinh 2022-2

Práctica 4. Intervalos de confianza para proporciones y varianza.

Intervalo de confianza: Es un intervalo (rango) de valores calculados a partir de una muestra en el cual se encuentra el valor verdadero del parámetro, con una probabilidad determinada.

- La probabilidad de que valor verdadero del parámetro sen encuentre en el intervalo calculado se denomina **nivel de confianza** y se define como **(1-α).**
- La probabilidad de equivocación (diferente a tener un error) se denomina **nivel de significancia** y se simboliza con **α.**
- **Un nivel de confianza (1-α) = 95%** corresponde a una **significancia α=5%**

Cálculos:

Cálculo de la media poblacional cuando se conoce la varianza σ^2 (desviación estándar σ). Se supone una distribución normal.	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Si usamos $\bar{x}$ como una estimación de μ, podemos tener $100(1-\alpha)\%$ de confianza en que el error no excederá a una cantidad especifica e cuando el tamaño de la muestra sea:	$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$ $e=(\bar{x} - \mu)$

Si se desconoce la varianza:

Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población normal de la que se **desconoce** la varianza σ^2, un intervalo de confianza del $100(1-\alpha)$ % para μ es:	$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$
Cuando no sea posible suponer la normalidad, se **desconozca** σ y $n \geq 30$, σ se puede reemplazar con s para poder utilizar el intervalo de confianza:	$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

Intervalo de confianza para una proporción:

En este caso, interesa construir un intervalo de confianza para una proporción o un porcentaje p poblacional.

Proporción de la muestra

$\hat{p}$=x/n, x representa el número de éxitos en una muestra de tamaño n. % de éxitos

$\hat{q} = 1 - (\hat{p}\ \ \% \ de\ fracasos)$

Si el tamaño de la muestra es grande, por el teorema del límite central:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Ejercicio:

```
"Funcion prop.test
Se utiliza para calcular intervalos de confianza para la proporcion y
diferencia de proporciones, us argumentos son
prop.test(x,n, p=NULL, alternatives= c(`two.sised`, `less´, greater`),
conf.level=0.95, correct=TRUE)"

"Intervalo de confianza bilateral para la proporcion p:
Se neceitan 3 argumentos: x, considera el conteo de exitos
n, indica el numero de evento o de forma equivalente o corresponde a la
longitud de la variable que se quiere analizar y conf.level= nvl de conf"

"El gerente de una estacion de TV debe determinar en la ciuadad que porc
de casas tienen mas de una TV. Una muestra aleatoria de 500 casas revela
que 275 tienen dos televisores o ma. ¿Cual es el intervalo de confianza
de 90% para estimar la proporcion de todas las casas que tienen 2 o mas
tvs"

prop.test(x=340, n=500, conf.level=0.95)$conf.int
```

```
> prop.test(x=340, n=500, conf.level=0.95)$conf.int
[1] 0.6368473 0.7203411
attr(,"conf.level")
[1] 0.95
>
```

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Práctica 5. Intervalo de confianza para dos poblaciones.

Procedimiento:
1. Revisa conceptos básicos.
2. Clasifica el tipo de variable.
3. Establece procedimiento para la elaboración del intervalo correspondiente a la variable.
4. Construye intervalo de confianza.
5. Presentar en el laboratorio los ejercicios completos en R.

jinh 2022-2

Práctica 5. Intervalo de confianza para dos poblaciones.

Dos muestras: estimación de la diferencia entre dos medias

Si tenemos dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2, respectivamente, el estadístico que da un estimador puntual de la diferencia entre μ_1 y μ_2 es $\bar{X}_1 - \bar{X}_2$. Por lo tanto, para obtener una estimación puntual de $\mu_1 - \mu_2$, se seleccionan dos muestras aleatorias independientes, una de cada población, de tamaños n_1 y n_2, y se calcula $\bar{x}_1 - \bar{x}_2$, la diferencia de las medias muestrales. Evidentemente, debemos considerar la distribución muestral de $\bar{X}_1 - \bar{X}_2$.

Cálculos:

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños n_1 y n_2, de poblaciones que tienen varianzas conocidas σ_1^2 y σ_2^2, respectivamente, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha

Varianzas desconocidas pero iguales:

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes con tamaños n_1 y n_2, respectivamente, tomadas de poblaciones más o menos normales con varianzas iguales pero desconocidas, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

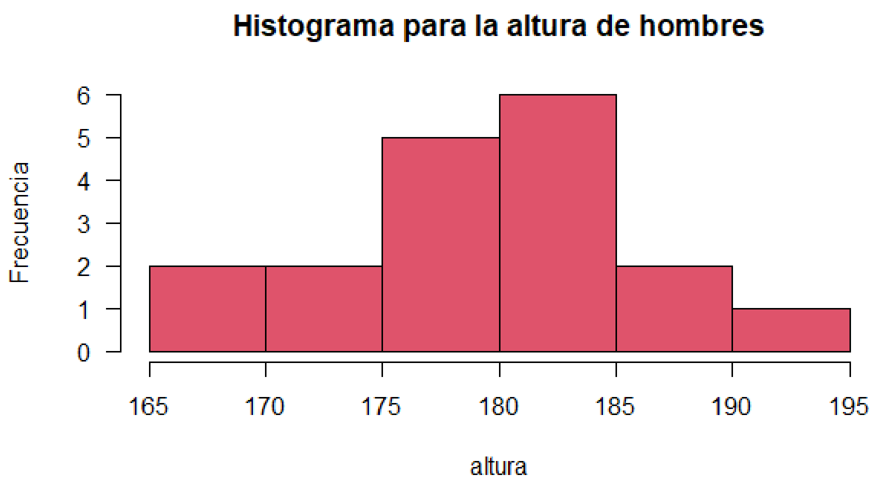
$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

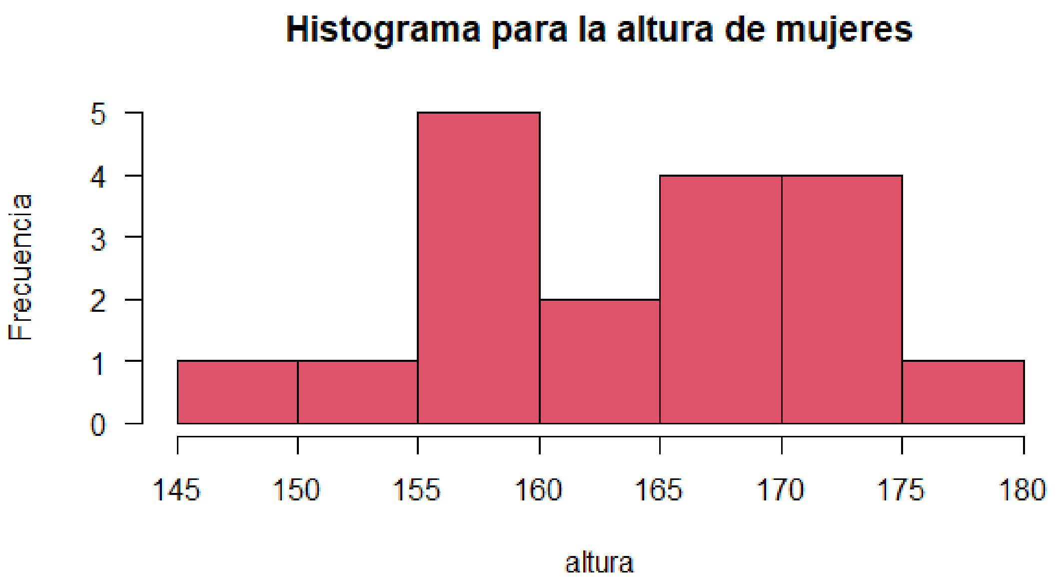
donde s_p es la estimación agrupada de la desviación estándar de la población y $t_{\alpha/2}$ es el valor t con $v = n_1 + n_2 - 2$ grados de libertad, que deja una área de $\alpha/2$ a la derecha.

Ejercicio:

```r
"Ejemplo: se requiere saber si existe diferencia significativo entre alturas de
hombres y mujeres, utilizar un intervalo de confianza de 95% para la
diferencia de las alturas promedio de honmbres y mujeres(m1-m2)"

#Importar archivo datos<-real.table(file="medidas_cuerpo.txt")

datos<-read.table(file = "medidas_cuerpo.txt")
data.class(datos)
str(datos)

#Primero analizar normalidad
Hombre<-datos[datos$sexo=="Hombre", ]
hist(Hombre$altura, las=1, xlab='altura',ylab='Frecuencia', col= 84,
     main='Histograma para la altura de hombres')
Mujer<-datos[datos$sexo=="Mujer", ]
hist(Mujer$altura, las=1, xlab='altura',ylab='Frecuencia', col= 84,
     main='Histograma para la altura de mujeres')

t.test(x=Hombre$altura, y=Mujer$altura, paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)$conf.int
# #[1] 10.05574   20.03315
# #attr(,"conf.level")
# #0.95

# Mediana
mean(Hombre$altura)
median(Hombre$altura)
mean(Mujer$altura)
median(Mujer$altura)

summary(Hombres$altura)
summary(Mujer$altura)
#AHORA CAMBIAMOS EL ORDEN DE LAS VARIABLES
t.test(x=Mujer$altura, y=Hombre$altura, paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)$conf.int
```

```
> datos<-read.table(file = "medidas_cuerpo.txt")
> data.class(datos)
[1] "data.frame"
> str(datos)
'data.frame':   36 obs. of  6 variables:
 $ edad  : int  43 65 45 37 55 33 25 35 28 26 ...
 $ peso  : num  87.3 80 82.3 73.6 74.1 85.9 73.2 76.3 65.9 90.9 ...
 $ altura: num  188 174 176 180 168 ...
 $ sexo  : chr  "Hombre" "Hombre" "Hombre" "Hombre" ...
 $ muneca: num  12.2 12 11.2 11.2 11.8 12.4 10.6 11.3 10.2 12 ...
 $ biceps: num  35.8 35 38.5 32.2 32.9 38.5 38.3 35 32.1 40.4 ...
> 
```






```
> t.test(x=Hombre$altura, y=Mujer$altura, paired = FALSE,
+        var.equal = FALSE, conf.level = 0.95)$conf.int
[1] 10.05574 20.03315
attr(,"conf.level")
[1] 0.95
>
> mean(Hombre$altura)
[1] 179.0778
> median(Hombre$altura)
[1] 179.7
> mean(Mujer$altura)
[1] 164.0333
> median(Mujer$altura)
[1] 163.6
>
> summary(Hombre$altura)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  167.6   176.8   179.7   179.1   182.3   190.5 
> summary(Mujer$altura)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  147.2   159.2   163.6   164.0   170.7   176.2 
> 
```

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 6. Investigación del Método de mínimos cuadrados (Regresión simple)

Procedimiento:

1. Realiza una búsqueda de información del método de mínimos cuadrados para regresión múltiple.
2. Identifica los conceptos estadísticos para el análisis estadístico.
3. Realiza un tutorial de resolución de problemas en software estadístico.
4. Presenta en el laboratorio un ejercicio de mínimos cuadrados en R.

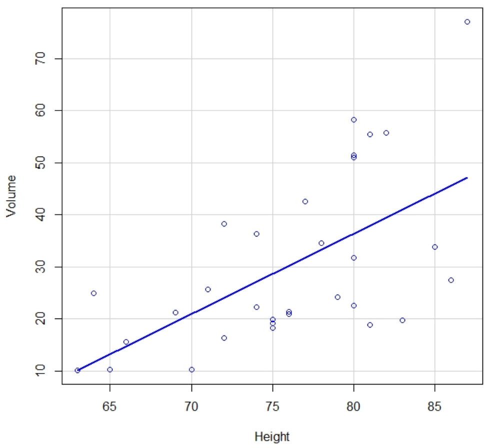
jinh 2022-2

Práctica 6. Investigación del Método de mínimos cuadrados (Regresión simple)

El análisis de la regresión lineal se utiliza para predecir el valor de una variable según el valor de otra.

La variable que desea predecir se denomina variable dependiente Yi.

La variable que está utilizando para predecir el valor de la otra variable se denomina variable independiente Xi



Cálculos:

$$a = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{N\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Sumatoria de residuos al cuadrado RSS. En el método del gradiente se elegirá la recta cuyo valor de RSS sea menor.

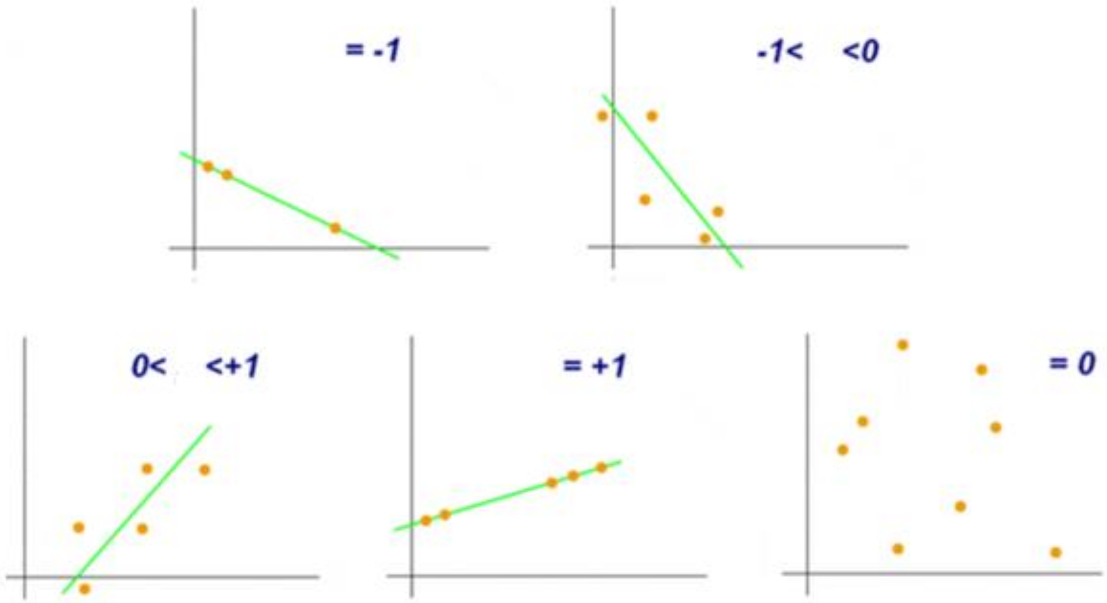
Coeficiente de Correlación de Pearson (r) y el Coeficiente de Determinación r Cuadrado (r^2 o R^2):

Coeficiente de correlación de Pearson r mide la dependencia lineal entre dos variables.

$$r_{xy} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cuando el coeficiente r de Pearson se eleva al cuadrado, se conoce como el coeficiente de determinación, r al cuadrado o r^2, e indica el porcentaje de la variación de una variable debido a la variación de la otra y viceversa.

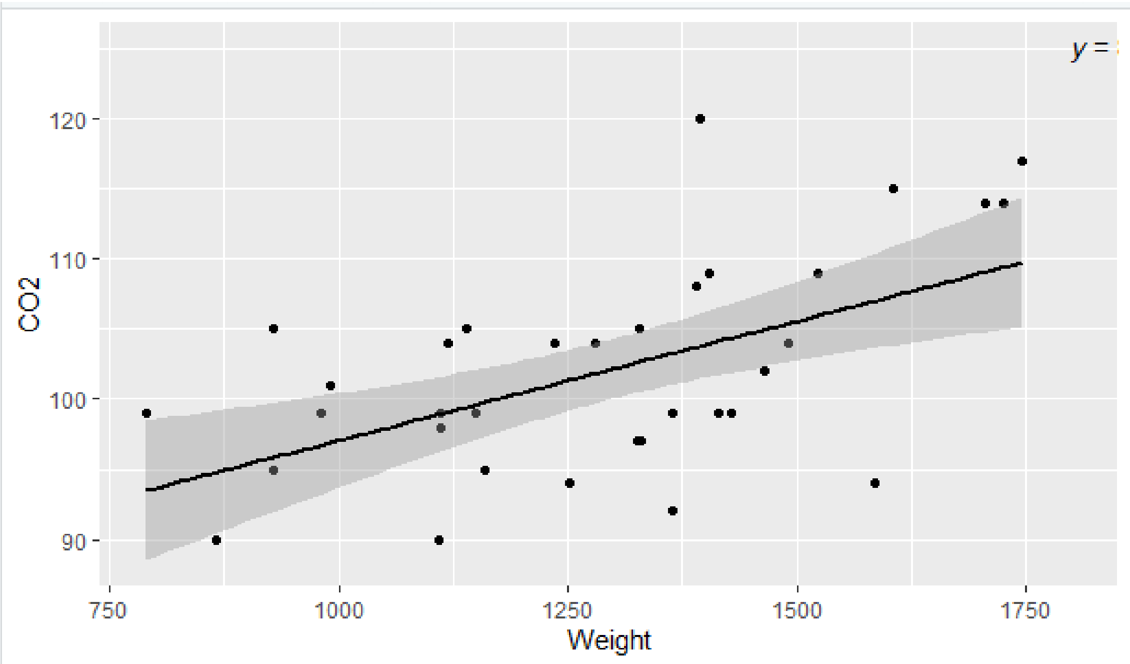
Es decir, el, es la proporción de la variación en Y explicada por X. Puede adoptar cualquier valor entre 0 y 1.



$$r^2 = R^2 = \frac{(\sum x\,y - n\,\bar{x}\,\bar{y})^2}{(\sum x^2 - n\,\bar{x}^2)(\sum y^2 - n\,\bar{y}^2)}$$

Ejercicio:

```
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)

cars<- read.csv("cars.csv")
view(cars)
summary(cars)
hist(cars$CO2)

plot(CO2 ~ Volume, data = cars)

CO2.Volume.ml<-lm(CO2 ~ Volume, data = cars)
summary(CO2.Volume.ml)

#agregar linea de regresion a los datos
#1ro. se crea la grafica
CO2.grafica<-ggplot(cars, aes(x=Volume, y=CO2))+geom_point()
CO2.grafica
#2do. agregar la linea de tendencia a la grafica
CO2.grafica <- CO2.grafica + geom_smooth(method="lm", col="black")
CO2.grafica
#3ro. agregar la ecuacion de la linea de regresion
CO2.grafica <- CO2.grafica + stat_regline_equation(label.x = 2500, label.y = 120)

CO2.grafica


```



UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 7. Estimación de modelos de predicción lineal.

Procedimiento:

1. Analiza los casos prácticos a resolver.
2. Resuelve los ejercicios con apoyo de software estadístico.
3. Concluye la eficiencia del modelo de predicción mediante el coeficiente de determinación.
4. Estima nuevos parámetros.
5. Realiza un ejemplo en la sesión de laboratorio correspondiente.

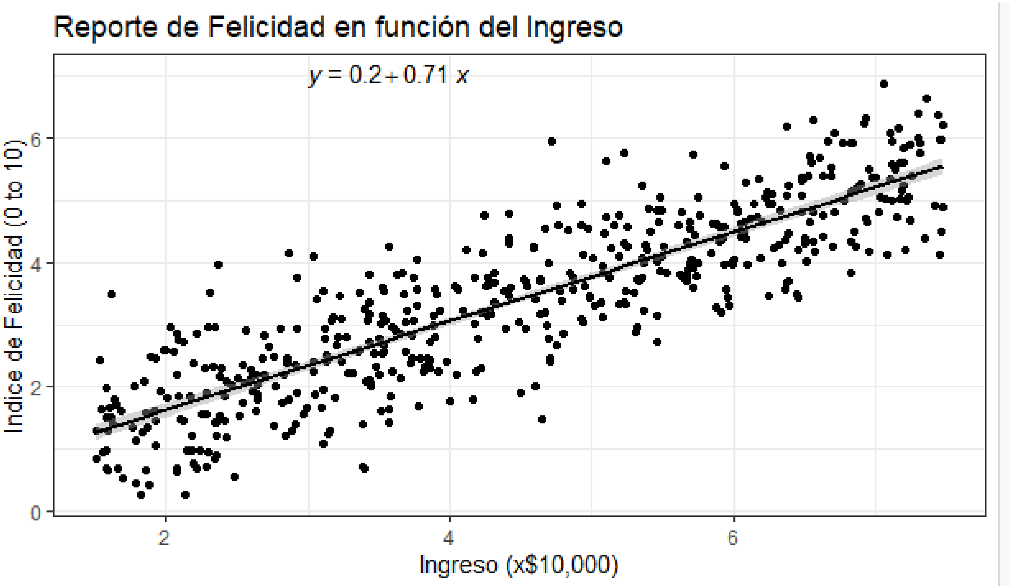
jinh 2022-2

Práctica 7. Estimación de modelos de predicción lineal.

Ejercicio:

```r
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
income.data <- read.csv("income.data.csv")
View(income.data)
summary(income.data)
hist(income.data$happiness)
plot(happiness ~ income, data = income.data)
#Analisis de regresion
income.happiness.lm <- lm(happiness ~ income, data = income.data)
#
summary(income.happiness.lm)

#agregar linea de regresion a los datos
#1ro. se crea la grafica
income.graph<-ggplot(income.data, aes(x=income, y=happiness))+geom_point()
income.graph
#2do. agregar la linea de tendencia a la grafica
income.graph <- income.graph + geom_smooth(method="lm", col="black")
income.graph
#3ro. agregar la ecuacion de la linea de regresion
income.graph <- income.graph +
  stat_regline_equation(label.x = 3, label.y = 7)

income.graph
#grafica final con titulos
income.graph +
  theme_bw() +
  labs(title = "Reporte de Felicidad en función del Ingreso",
       x = "Ingreso (x$10,000)",
       y = "Indice de Felicidad (0 to 10)")
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.20427    0.08884   2.299   0.0219 *  
income       0.71383    0.01854  38.505   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493,	Adjusted R-squared:  0.7488 
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```



Conclusiones:

El valor obtenido de R2 fue de 0.749, lo que no alcanza el estandar de 0.9 para poder afirmar que es una estimación determinante. Pero, a su vez, nos ayuda a visualizar la tendencia que siguen los resultados.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 8. Investigación del Método de mínimos cuadrados (Regresión múltiple).

Procedimiento:

1. Realiza una búsqueda de información del método de mínimos cuadrados.
2. Identifica los conceptos estadísticos para el análisis estadístico.
3. Realiza un tutorial de resolución de problemas en software estadístico.
4. Realiza un reporte de investigación, que cumpla con introducción, desarrollo, conclusiones y referencias.

jinh 2022-2

Práctica 8. Investigación del Método de mínimos cuadrados (Regresión múltiple).

En la regresión lineal múltiple vamos a utilizar más de una variable explicativa; esto nos va a ofrecer la ventaja de utilizar más información en la construcción del modelo y, consecuentemente, realizar estimaciones más precisas.

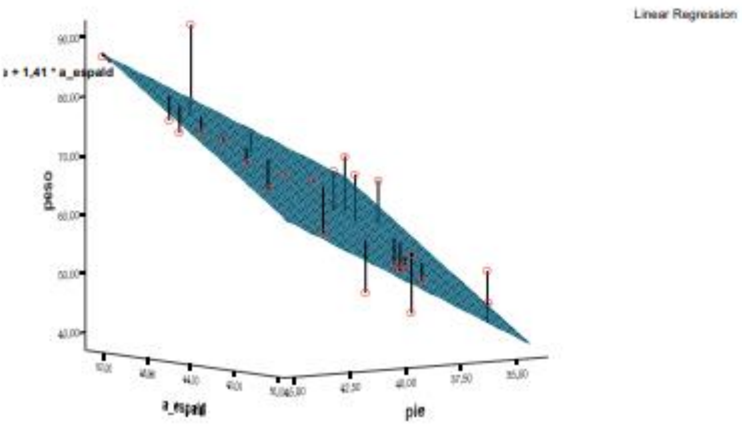
Al tener más de una variable explicativa (no se debe de emplear el término independiente) surgirán algunas diferencias con el modelo de regresión lineal simple. Una cuestión de gran interés será responder a la siguiente pregunta: de un vasto conjunto de variables explicativas: x1, x2, …, xk, cuáles son las que más influyen en la variable dependiente Y.

En definitiva, y al igual que en regresión lineal simple, vamos a considerar que los valores de la variable dependiente Y han sido generados por una combinación lineal de los valores de una o más variables explicativas y un término aleatorio:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + u$$

Los coeficientes son elegidos de forma que la suma de cuadrados entre los valores observados y los pronosticados sea mínima, es decir, que se va a minimizar la varianza residual.

Esta ecuación recibe el nombre de **hiperplano**, pues cuando tenemos dos variables explicativas, en vez de recta de regresión tenemos un plano:



El método de los mínimos cuadrados se utiliza para calcular la recta de regresión lineal que minimiza los residuos, esto es, las diferencias entre los valores reales y los estimados por la recta. Se revisa su fundamento y la forma de calcular los coeficientes de regresión con este método.

$$Min \sum (y_j - \hat{y}_j)^2$$

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 9. Estimación de modelos de regresión lineal múltiple.

Procedimiento:

1. Realiza una búsqueda de un modelo de regresión múltiple que describa el comportamiento de algún sistema.
2. Define la variable independiente y las variables independientes del modelo.
3. Utiliza al menos dos herramientas de software para la estimación del modelo usando los métodos vistos en clase y muestra los resultados en la sesión de laboratorio correspondiente.
4. Realiza una tabla comparativa de los resultados.

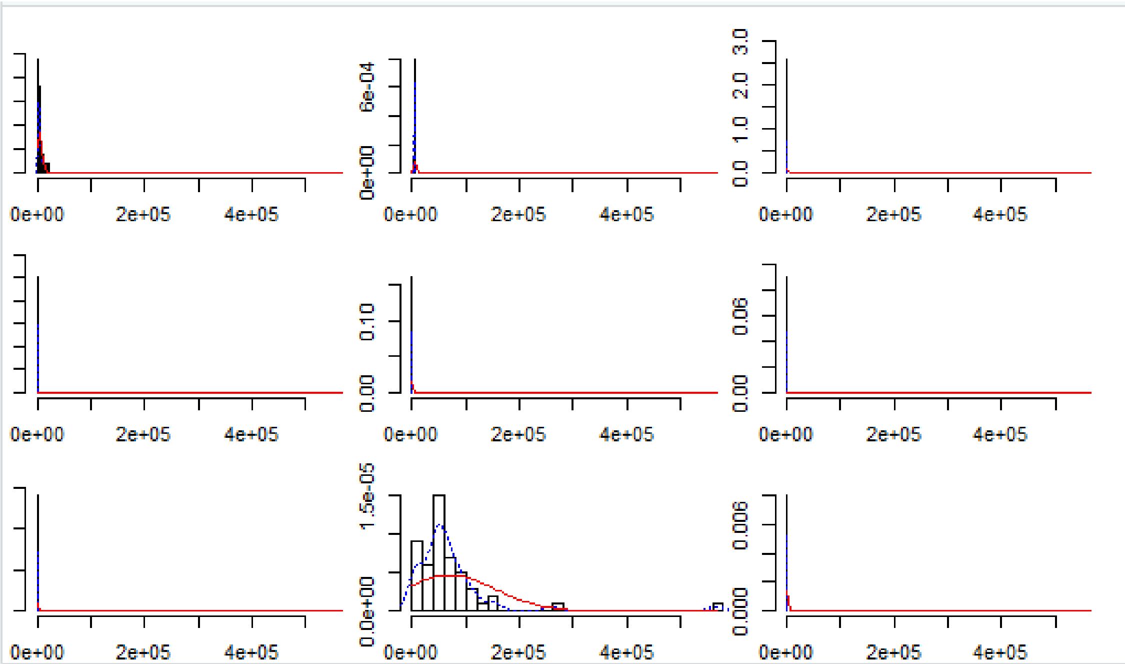
jinh 2022-2

Práctica 9. Estimación de modelos de regresión lineal múltiple.

```r
library(dplyr)
library(psych)
#Un estudio quiere generar un modelo que permita predecir la
#esperanza de vida media de los habitantes de una ciudad en función
#de diferentes variables. Se dispone de información sobre: habitantes,
#analfabetismo, ingresos, esperanza de vida, asesinatos, universitarios,
#heladas, área y densidad poblacional.
#state.x77 viene en los datos de R
#write.table(datos, file="state.x77") # guarda los datos
datos <- as.data.frame(state.x77)
datos <- rename(habitantes = Population, analfabetismo = Illiteracy,
                ingresos = Income, esp_vida = `Life Exp`, asesinatos = Murder,
                universitarios = `HS Grad`, heladas = Frost, area = Area,
                .data = datos)
datos <- mutate(.data = datos, densidad_pobl = habitantes * 1000 / area)
#Guarda los datos renombrados
#write.table(datos, file="state.x77_renombrados")

#Histogramas de las variables

multi.hist(x = datos, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
           main = "")

#Generacion del modelo
modelo <- lm(esp_vida ~ habitantes + ingresos + analfabetismo + asesinatos +
               universitarios + heladas + area + densidad_pobl, data = datos )
summary(modelo)

#El modelo con todas las variables introducidas como predictores tiene
#un R2 alta (0.7501), es capaz de explicar el 75,01% de la variabilidad
#observada en la esperanza de vida.
#El p-value del modelo es significativo (3.787e-10) por
#lo que se puede aceptar que el modelo no es por azar.
```

- Variable dependiente: Esperanza de vida
- Variables independientes: Habitantes, Ingresos, Analfabetismo, Asesinatos, Universitarios, Heladas, Area, Densidad de Población.

```
Call:
lm(formula = esp_vida ~ habitantes + ingresos + analfabetismo +
    asesinatos + universitarios + heladas + area + densidad_pobl,
    data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47514 -0.45887 -0.06352  0.59362  1.21823

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.995e+01  1.843e+00  37.956  < 2e-16 ***
habitantes      6.480e-05  3.001e-05   2.159   0.0367 *
ingresos        2.701e-04  3.087e-04   0.875   0.3867
analfabetismo   3.029e-01  4.024e-01   0.753   0.4559
asesinatos     -3.286e-01  4.941e-02  -6.652 5.12e-08 ***
universitarios  4.291e-02  2.332e-02   1.840   0.0730 .
heladas        -4.580e-03  3.189e-03  -1.436   0.1585
area           -1.558e-06  1.914e-06  -0.814   0.4205
densidad_pobl  -1.105e-03  7.312e-04  -1.511   0.1385
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7337 on 41 degrees of freedom
Multiple R-squared:  0.7501,    Adjusted R-squared:  0.7013
F-statistic: 15.38 on 8 and 41 DF,  p-value: 3.787e-10
```



UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 10. Análisis de regresión lineal múltiple.

Procedimiento:

1. Realiza la búsqueda de tablas de datos de fuentes de información confiables y pertinentes para la aplicación de métodos de análisis de regresión múltiple.
2. Realiza la búsqueda de un modelo de regresión lineal múltiple que describa el comportamiento de los datos encontrados.
3. Utiliza al menos dos herramientas de software para encontrar en las tablas de datos, la relación entre una variable independiente y un conjunto de variables independientes.
4. Utiliza al menos dos herramientas de software para encontrar en el modelo de regresión lineal múltiple, la relación entre una variable independiente y un conjunto de variables independientes.
5. Realiza tablas y gráficas para comparar los resultados obtenidos.
6. Resuelve un ejercicio en la sesión de laboratorio correspondiente.

Práctica 10. Análisis de regresión lineal múltiple.

Ejemplo: En un estudio sobre la población de un parásito se hizo un recuento de parásitos en 15 localizaciones con diversas condiciones ambientales. Los datos obtenidos son los siguientes:

Temperatura	Humedad	Recuento
15	70	156
16	65	157
24	71	177
13	64	145
21	84	197
16	86	184
22	72	172
18	84	187
20	71	157
16	75	169
28	84	200
27	79	193
13	80	167
22	76	170
23	88	192

Excel:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,956022932
Coeficiente de determinación R^2	0,913979846
R^2 ajustado	0,899643154
Error típico	5,350557268
Observaciones	15

ANÁLISIS DE VARIANZA

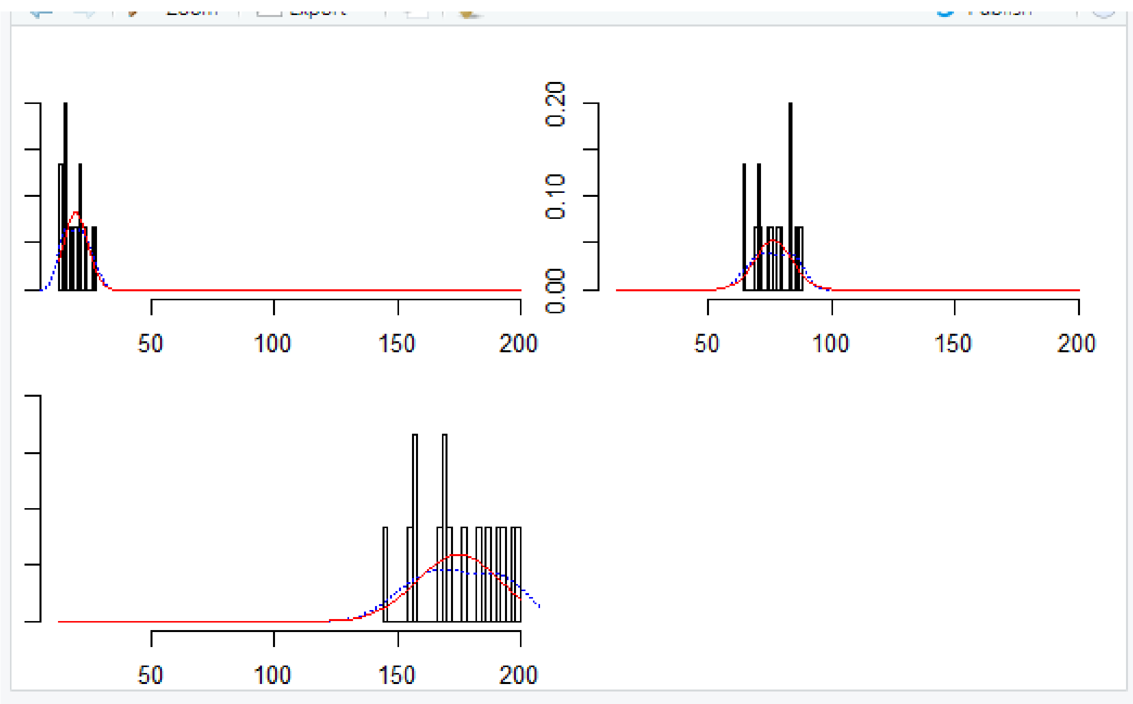
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	3650,191776	1825,095888	63,75109566	4,05136E-07
Residuos	12	343,5415569	28,62846308		
Total	14	3993,733333			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	25,71153011	14,3724855	1,788941106	0,098875902	-5,603425691	57,02648592	-5,603425691	57,02648592
Variable X 1	1,581817976	0,32026301	4,939121688	0,000342592	0,884024822	2,27961113	0,884024822	2,27961113
Variable X 2	1,542447836	0,199504179	7,731406155	5,32418E-06	1,107765571	1,9771301	1,107765571	1,9771301

RStudio:

```
library(dplyr)
library(psych)

datos <- read.table(file="bacterias.txt")
str(datos)

multi.hist(x = datos, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
           main = "")
#generacion de modelo

modelo <- lm(Recuento ~ Humedad + Temperatura, data = datos )
summary(modelo)
```

```
Call:
lm(formula = Recuento ~ Humedad + Temperatura, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8617 -2.0406  0.4319  2.9881  8.5047 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  25.7115    14.3725   1.789 0.098876 .  
Humedad       1.5424     0.1995   7.731 5.32e-06 ***
Temperatura   1.5818     0.3203   4.939 0.000343 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.351 on 12 degrees of freedom
Multiple R-squared:  0.914,	Adjusted R-squared:  0.8996 
F-statistic: 63.75 on 2 and 12 DF,  p-value: 4.051e-07
```



Conclusión:

Con ambas herramientas de software se pudo realizar el modelo de regresión lineal múltiple. Sin embargo en el caso de Excel, los calculos realizados arrojan un coeficiente de determinación (R2) un poco mayor. Aün así, la diferencia entre estos dos datos no es significativa para poder descartar la opción de RStudio.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 11. Elaboración y análisis de Tablas de contingencia.

Procedimiento:

1. Elabora una tabla de contingencia a partir de una serie de datos.
2. Aplica el análisis probabilístico a través de la probabilidad condicional.
3. Aplica el estadístico de contraste.
4. Entrega hoja de cálculo o documento con ejercicios completos.

jinh 2022-2

Práctica 11. Elaboración y análisis de Tablas de contingencia.

En estadística las tablas de contingencia se emplean para registrar y analizar la asociación entre dos o más variables, habitualmente de naturaleza cualitativa (nominales u ordinales).

Una variable cualitativa nominal presenta modalidades no numéricas que no admiten un criterio de orden. Por ejemplo: El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo.

Una variable ordinal es un tipo de variable estadística de tipo cualitativo que expresa con palabras una cualidad de naturaleza ordenable. Es decir, una variable ordinal es una variable que puede ser ordenada.

Ejemplo

Suponiendo que se tienen dos variables, la primera el género (Masculino - Femenino) y la segunda recoge si el individuo es zurdo o diestro. Se ha observado esta pareja de variables en una muestra aleatoria de 100 individuos. Se puede emplear una tabla de contingencia para expresar la relación entre estas dos variables:

	Diestro	Zurdo	Total
Hombre	43	9	52
Mujer	44	4	48
Total	87	13	100

Las cifras en la columna de la derecha y en la fila inferior reciben el nombre de *frecuencias marginales* y la cifra situada en la esquina inferior derecha es el gran total.

La tabla nos permite ver de un vistazo que la proporción de hombres diestros es aproximadamente igual a la proporción de mujeres diestras.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 12. Pruebas de bondad y ajuste.

Procedimiento:

1. Realiza una búsqueda bibliográfica sobre las diferentes pruebas de bondad y ajuste.
2. Genere una tabla con la información relevante de cada una.
3. Resuelva una serie de ejercicios aplicando la prueba indicada para probar hipótesis.
4. Justifique sus resultados en forma clara.
5. Entrega hoja de cálculo o documento con ejercicios completos

jinh 2022-2

Práctica 12. Pruebas de bondad y ajuste.

Las pruebas de bondad de ajuste son pruebas de hipótesis para verificar si los datos observados en una muestra aleatoria se ajustan con algún nivel de significancia a determinada distribución de probabilidad (uniforme, exponencial, normal, poisson, u otra cualquiera). n La hipótesis nula Ho indica la distribución propuesta, mientras que la hipótesis alternativa H1 , nos indica que la variable en estudio tiene una distribución que no se ajusta a la distribución propuesta.

Ejemplo:

Para estudiar la dependencia entre la práctica de algún deporte y la depresión, se seleccionó a una muestra aleatoria de 100 jóvenes, con lo siguientes resultados

obtenidos			
	Sin depresión	Con depresión	Total
Deportista	38	9	47
No deportista	31	22	53
			100

Determine si existe independencia entre la actividad del sujeto y su condición. Nivel de significancia de 5%

Desarrollo:

obtenidos			
	Sin depresión	Con depresión	Total
Deportista	38	9	47
No deportista	31	22	53
			100

esperados			
	Sin Depresión	Con depresión	
Deportista	32,43	14,57	47
No deportista	36,57	16,43	53
	69	31	100

X^2= $\frac{(38-32,43)^2}{32,43} + \frac{(9-14,57)^2}{14,57} + \frac{(31-36,57)^2}{36,57} + \frac{(22-16,43)^2}{16,43} = 5,82$

El valor obtenido fue de 5.82, siendo un valor mayor que el de la tabla

Tabla A.5 (*continuación*) Valores críticos de la distribución chi cuadrada

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815

Conclusión:

El valor calculado es mayor que el valor crítico de la tabla , concluimos que podemos rechazar la hipótesis nula de independencia, es decir; Sí existe relación entre la depresión y los hábitos deportista en los jóvenes.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Ingeniero en Software y Tecnologías Emergentes

Estadística Avanzada

Práctica 13. Análisis de varianza.

Procedimiento:

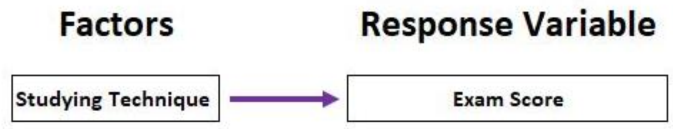
1. Realiza búsqueda bibliográfica sobre conceptos presentes en práctica.
2. Resuelve serie de ejercicios con el nivel se significancia indicados.
3. Indique sus conclusiones.
4. Entrega en la sesión de laboratorio hoja de cálculo o documento con ejercicios completos.

jinh 2022-2

Práctica 13. Análisis de varianza.

Análisis de Varianza (ANOVA) de un solo factor.

De ***k*** poblaciones se seleccionan muestras aleatorias de tamaño ***n***. Las ***k*** poblaciones diferentes se clasifican con base en un criterio único.



Se supone que las ***k*** poblaciones son independientes y que están distribuidas en forma normal con medias μ_1, μ_2,..., μ_k, y varianza común σ^2. Estas suposiciones son más aceptables mediante la aleatoriedad.

Se tiene como objetivo:

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k,$$
$$H_1: \ \text{Al menos dos de las medias no son iguales.}$$

Arreglo de los datos:

Tratamiento:	**1**	**2**	···	***i***	···	***k***	
	y_{11}	y_{21}	···	y_{i1}	···	y_{k1}	
	y_{12}	y_{22}	···	y_{i2}	···	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	···	y_{in}	···	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	···	$Y_{i.}$	···	$Y_{k.}$	$Y_{..}$
Media	$\bar{y}_{1.}$	$\bar{y}_{2.}$	···	$\bar{y}_{i.}$	···	$\bar{y}_{k.}$	$\bar{y}_{..}$

y_{ij} denote la j-ésima observación del i-ésimo,
Y_i es el total de todas las observaciones de la muestra, del i-ésimo tratamiento,
$\bar{y}_i$ es la media de todas las observaciones en la muestra del i-ésimo tratamiento, Y es el total de todas las nk observaciones,
$\bar{y}$ es la media de todas las nk observaciones.

Modelo de ANOVA para un solo factor

Cada observación puede escribirse en la forma $Y_{ij} = \mu_i + \varepsilon_{ij}$,

donde ε_{ij} mide la desviación que tiene la observación j-ésima de la i-ésima muestra, con respecto de la media del tratamiento correspondiente (factor bajo observación).

$Y_{ij} = \mu_i + \varepsilon_{ij}$, es la ecuación de la recta, con pendiente ε_{ij} y en consecuencia puede resolverse con el método de mínimos cuadrados.

Análisis de varianza para el ANOVA de un solo factor

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
Tratamientos	SCT	$k-1$	$s_1^2 = \frac{SCT}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SCE	$N-k$	$s^2 = \frac{SCE}{N-k}$	
Total	STC	$N-1$		

$$STC = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{suma total de cuadrados,}$$

$$SCT = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{suma de los cuadrados del tratamiento,}$$

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{suma de los cuadrados del error.}$$

$$STC = SCT + SCE$$

Los valores críticos dentro de la tabla se comparan con el estadístico F de una prueba F.

Si el **estadístico F es mayor que el valor crítico encontrado en la tabla**, entonces puede rechazar la hipótesis nula de la prueba F y concluir que los resultados de la prueba son estadísticamente significativos.

Ejemplo:

	A	B	C	D	E	F
1		C1	C2	C3	C4	C5
2		551	595	639	417	563
3		457	580	615	449	631
4		450	508	511	517	522
5		731	583	573	438	613
6		499	633	648	415	656
7		632	517	677	555	679
8	TotalColumna	3320	3416	3663	2791	3664
9	MediaColumna	553,3333333	569,3333333	610,5	465,1667	610,6667
10						
11						
12		71,68444444	56,75111061	2371,69	9338,001	2387,951
13		SumadelosCuadradosdelTratamiento SCT				
14						
15					SCT	85356,47

K	L	M	N	O	P
	SumaTotaldeCuadrados STC				
116,64	1102,24	5959,84	20967,04	1,44	28147,2
10983,04	331,24	2830,24	12723,84	4788,64	31657
12499,24	2894,44	2580,64	2007,04	1584,04	21565,4
28628,64	449,44	125,44	15326,44	2621,44	47151,4
3943,84	5069,44	7430,44	21550,24	8873,64	46867,6
4928,04	2007,04	13271,04	46,24	13735,84	33988,2
61099,44	11853,84	32197,64	72620,84	31605,04	
			STC=	**209376,8**	
					209376,8

85356,47 la suma de las medias de las columnas por la cantidad de elementos

SCE = 209377 – 85356 = 124021.

$v_1 = k - 1 = 5 - 1 = 4$

$v_2 = N - k = 30 - 5 = 25$

$\alpha = 0.05$

$s_1^2 = (853563.5)/4 = 21339.1$

$s^2 = 124021/25 = 4960.8$ $f = 21339.1/4960.8 = 4.3$ el valor calculado es mayor al valor f de la tabla

$f(\alpha = 0.05, v_1 = 4, v_2 = 29) = 2.7$ se rechaz; las medias no son iguales

El valor calculado es mayor que el valor crítico de la tabla , concluimos que podemos rechazar la hipótesis nula de independencia.

Como el f calculado es mayor al valor de la tabla, rechazamos Ho y concluimos que los 5 experimentos tienen medias diferentes, al menos 2 son diferentes.