Proyecto de estadística: Etapa 3

Sebastián Rojas Vargas

Francisco Soto Quesada

Jairo Pacheco Campos

Jason Barrantes Rodríguez

10/06/2021

# Lista de librerías utilizadas

library(visdat)  
library(ggplot2)  
library(DataExplorer)  
library(datos)  
library(EnvStats)  
library(stests)  
library("ggpubr")  
library(stats)  
library(BSDA)  
library(PASWR2)  
library(TeachingDemos)  
library(mosaicData) # para usar KidsFeet  
library(nortest) # para usar A-D, S-W, K-S-L test  
library(fBasics) # para usar D’Agostino-Pearson  
library(moments) # para custoris y simetría  
library(corrplot)  
library(bbmle)  
library(DALEX)

# I Parte: explicación de los datos

A continuación, se presenta una tabla con los principales aspectos del dataset utilizado.

**Dataset:** Diamantes

**Descripción general:** se optó por la utilización del dataset de diamantes del paquete de datos de R. El motivo de la elección de este es que cuenta con una gran cantidad de datos y columnas. Estas últimas siendo de gran variedad entre datos cualitativos y cuantitativos. Además de todo esto el dataset cuenta con todas sus variables y sus datos ya traducidos al español, facilitando así la comprensión de los datos.

**Filas:** 53 940

**Columnas:** 10

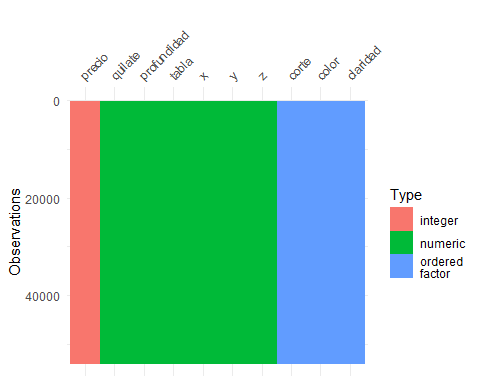
**Resumen del dataset:**

str(diamantes)

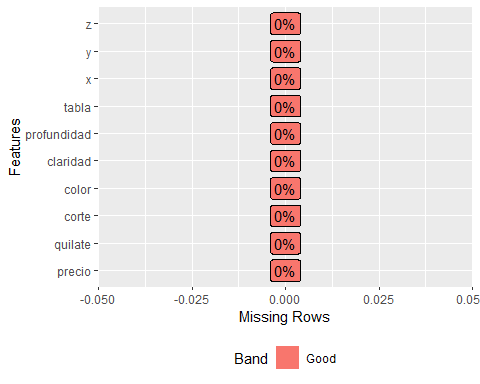
## tibble [53,940 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ precio : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...  
## $ quilate : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...  
## $ corte : Ord.factor w/ 5 levels "Regular"<"Bueno"<..: 5 4 2 4 2 3 3 3 1 3 ...  
## $ color : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...  
## $ claridad : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...  
## $ profundidad: num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...  
## $ tabla : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...  
## $ x : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...  
## $ y : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...  
## $ z : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...

**Resumen de los datos:**

vis\_dat(diamantes)



plot\_missing(diamantes)



En esta última imagen se puede observar que no hay datos faltantes en las columnas del dataset. Esto facilita los cálculos que se vean a realizar más adelante ya que no hay necesidad de filtrar los datos faltantes.

## Resumen de variables seleccionadas

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Tipos | Descripción | Parámetro por estimar (IC) |
| Precio | Cuantitativa | Precio en dólares estadounidenses | Promedio usando distribución z y distribución t |
| Corte | Cualitativa | Calidad del corte (Regular, Bueno, Muy bueno, Premium, Ideal) | Proporción, diferencia de proporciones |
| Profundidad | Cuantitativa | Porcentaje de la profundidad total en milímetros | Diferencia de promedios usando distribución z y distribución t |
| Quilates | Cuantitativa | Peso del diamante | Cociente de varianzas y varianza. |

# II Parte: Análisis Inferencial (IC)

## IC de un promedio usando distribución Z

A continuación, se presentan los promedios usando distribución normal estándar de una población de datos extraídos del dataset diamantes, utilizando los siguientes datos de la variable precio del dataset de diamantes:

set.seed(4562) # semilla  
  
poblacion <- as.data.frame(diamantes)  
poblacion.precios <- poblacion$precio  
  
n1 <- length(poblacion.precios)  
mu1 <- mean(poblacion.precios)  
var1 <- var(poblacion.precios)

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Varianza |
| A | 53940 | 3932.7997219 | 1.5915629^{7} |

**Hipótesis asumidas**

Dado que la muestra es mayor a 30 se puede asumir que la distribución muestral de medias sigue una distribución normal y que se puede aproximar σ₁ mediante s₁.

**Cálculo**

Conociendo los datos se puede aproximar el IC utilizando la función z.test de las librerías stests BSDA y PASWR2.

Para stests los parámetros son los siguientes:

* **x:** vector numérico que representa la muestra
* **sigma2:** varianza de la muestra
* **conf.level:** nivel de confidencia (opcional, se asume 0.95 por defecto)

stests::z.test(  
 x = poblacion.precios,   
 sigma2 = var(poblacion.precios),   
 conf.level = 0.95  
 )$conf.int

## [1] 3899.133 3966.467  
## attr(,"conf.level")  
## [1] 0.95

Los parámetros tanto para BSDA y PASWR2 son los siguientes:

* **x:** la muestra
* **sigma.x:** desviación estándar de la muestra
* **conf.level:** nivel de confidencia (opcional, se asume 0.95 por defecto)

BSDA::z.test(  
 x = poblacion.precios,   
 sigma.x = sqrt(var(poblacion.precios)),   
 conf.level = 0.95  
 )$conf.int

## [1] 3899.133 3966.467  
## attr(,"conf.level")  
## [1] 0.95

IC <- PASWR2::z.test(  
 x = poblacion.precios,   
 sigma.x = sqrt(var(poblacion.precios)),   
 conf.level = 0.95  
 )$conf.int  
  
IC

## [1] 3899.133 3966.467  
## attr(,"conf.level")  
## [1] 0.95

**Conclusiones**

De los resultados obtenidos se puede concluir que la media de precios del dataset diamante se encuentra en el intervalo ]3899.1327134, 3966.4667305[ con un 95% de confianza, resultado que coincide en las tres librerías utilizadas.

## IC de un promedio usando distribución t

A continuación, se presentan los promedios usando distribución t de una población, utilizando muestras de la variable precio del dataset de diamantes.

## Se toman filas random para sacar del dataset  
set.seed(6894)  
filas.random <- sample(1:53940, 20, replace= T)  
  
## Se generan datasets nuevos con las filas random anteriores   
muestra1 <- as.data.frame(diamantes[filas.random,])  
muestra.precios <- muestra1$precio  
n1 <- length(muestra.precios)  
mu1 <- mean(muestra.precios)  
de1 <- sqrt(var(muestra.precios))  
alpha1 <- 0.05

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Desviación estándar |
| A | 20 | 4629.6 | 5083.1543087 |

En este caso dado que el tamaño de la muestra es menor a 30 se ha optado por realizar una inspección visual de los datos para ver si se asemeja a una distribución normal.



**Hipótesis asumidas**

Como se puede observar, la muestra tiene forma de campana, por lo que se puede asumir que sigue una distribución normal.

**Cálculo**

Conociendo esto se puede aproximar el IC mediante la función t.test, la cual tiene la siguiente estructura, con los siguientes parámetros:

* **x:** muestra
* **conf.level:** nivel de confidencia

IC <- t.test(  
 x = muestra.precios,  
 conf.level = 0.95  
 )$conf.int  
  
IC

## [1] 2250.611 7008.589  
## attr(,"conf.level")  
## [1] 0.95

**Conclusión**

De los resultados obtenidos, se puede concluir que la media de precios del dataset de diamantes se encuentra en el intervalo ]2250.6105535, 7008.5894465[ con un 95% de confianza. Además de esto, comparándolo con el IC de la distribución Z se puede observar que este posee menor precisión debido a la poca cantidad de datos que fueron brindados.

## IC de una proporción

A continuación, se presenta la proporción para una población de datos tomados del dataset diamantes, para ello se ha utilizado una muestra de la variable “corte” donde se incluyen todos los diamantes cuyo corte sea de tipo “Premium” o “Ideal”, utilizando una muestra de gama alta de la siguiente manera:

gamaAlta <- diamantes[diamantes$corte == "Premium" | diamantes$corte == "Ideal",]  
length(gamaAlta$corte)

## [1] 35342

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Alta | Premium e Ideal | 35342 |

Como sabemos el dataset diamantes contiene una variable llamada corte, de igual manera, este dataset contiene otra variable llamada color. Estas variables las utilizaremos para saber cuál es la proporción que existe de un diamante de corte premium o ideal que hemos llamado gama alta con respecto al color que este posee. Para ello utilizamos la población de diamantes de gama alta que sean de color tipo D:

exitos.alta <- gamaAlta[gamaAlta$color == "D",]  
length(exitos.alta$color)

## [1] 4437

|  |  |
| --- | --- |
| Color | Tamaño de muestra |
| D | 4437 |

**Hipótesis asumidas**

Antes de realizar las operaciones se debe verificar que np > 5 y nq > 5

Para ello tenemos los siguientes datos:

Como ambos valores son mayores a 5 es posible obtener la proporción. Una vez conocidos estos datos se puede calcular el IC de la proporción, para esto se implementa la función prop.test para una población de la siguiente manera:

**Cálculo**

Una vez conociendo todos los datos, se procede a calcular el intervalo de confianza para una proporción con un nivel de confianza del 95%:

prop.test(x=length(exitos.alta$color), n=length(gamaAlta$corte), conf.level = 0.95)$conf.int

## [1] 0.1221170 0.1290541  
## attr(,"conf.level")  
## [1] 0.95

En donde:

* **x:** vector con el conteo de éxitos de la muestra
* **n:** vector con el número de ensayos la muestra
* **conf.level:** nivel de confianza

**Conclusiones**

De los resultados obtenidos se puede observar que el intervalo resultante es muy pequeño, esto debido a la gran cantidad de datos que se usaron para realizar el cálculo.

## IC de una diferencia de proporciones

A continuación, se presenta la diferencia de proporciones para dos poblaciones de datos tomados del dataset diamantes, para ello se han utilizado dos muestras de la variable “corte” donde en la primera se incluyen todos los diamantes cuyo corte sea de tipo “Premium” ó “Ideal” (muestra de gama alta) y la segunda muestra donde se incluyen todos los diamantes cuyo corte sea de tipo “Regular” ó “Bueno” (muestra de gama baja), de la siguiente manera:

gamaAlta <- diamantes[diamantes$corte == "Premium" | diamantes$corte == "Ideal",]  
length(gamaAlta$corte)

## [1] 35342

gamaBaja <- diamantes[diamantes$corte == "Regular" | diamantes$corte == "Bueno",]  
length(gamaBaja$corte)

## [1] 6516

exitos.alta <- gamaAlta[gamaAlta$color == "D",]  
length(exitos.alta$color)

## [1] 4437

exitos.baja <- gamaBaja[gamaBaja$color == "D",]  
length(exitos.baja$color)

## [1] 825

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Alta | Premium e Ideal | 35342 |
| Baja | Regular y Bueno | 6516 |

Como sabemos el dataset diamantes contiene una variable llamada corte, de igual manera, este dataset contiene otra variable llamada color. Estas variables las utilizaremos para saber cuál es la diferencia de proporciones que existen de un diamante de gama alta y de gama baja con respecto al color que este posee. Para ello utilizamos la población de diamantes de gama alta y baja que sean de color tipo D para representar los éxitos:

|  |  |  |
| --- | --- | --- |
| Gama | Color | Tamaño de muestra |
| Alta | D | 4437 |
| Baja | D | 825 |

**Hipótesis asumidas**

Antes de realizar las operaciones se debe verificar que np > 5 y nq > 5 para las dos poblaciones

Para ello tenemos los siguientes datos para la primera población:

Y tenemos los siguientes datos para la segunda población:

Como ambos valores son mayores a 5 significa que si se puede realizar la diferencia de proporciones. Una vez conocidos estos datos se puede calcular el IC para la diferencia de proporciones, para esto se implementa la función prop.test para dos poblaciones de la siguiente manera:

prop.test(x = c(length(exitos.alta$color) , length(exitos.baja$color)), n = c(length(gamaBaja$corte), length(gamaAlta$corte)), conf.level = 0.95)$conf.int

## [1] 0.6460786 0.6691132  
## attr(,"conf.level")  
## [1] 0.95

En donde:

* **x:** vector con el conteo de éxitos de la muestra
* **n:** vector con el número de ensayos la muestra
* **conf.level:** nivel de confianza

De la cual se puede concluir que la proporción de la muestra A es mayor que el de la muestra B.

## IC de una diferencia de promedios usando distribución z

Para este ejemplo se utilizarán dos grupos. En primer lugar, está el grupo de diamantes que se considera que tienen un color de buena calidad, este grupo está compuesto de aquellos diamantes cuyo color es “D”, “E” o “F”. Luego se tiene el grupo diamantes cuyo color es de peor calidad, el cual está compuesto por los diamantes con colores “H”, “I” o “J”. Ambos grupos están distribuidos de la siguiente manera:

buenaCalidad <- diamantes[diamantes$color == "D" | diamantes$color == "E" | diamantes$color == "F",]  
peorCalidad <- diamantes[diamantes$color == "H" | diamantes$color == "I" | diamantes$color == "J",]  
  
length(buenaCalidad$color)

## [1] 26114

length(peorCalidad$color)

## [1] 16534

|  |  |  |
| --- | --- | --- |
| Calidad | Colores | Tamaño de muestra |
| Buena | D, E o F | 26114 |
| Peor | H, I o J | 16534 |

**Hipótesis asumidas**

Dado que ambas muestras son mayores a 30 se puede asumir que la distribución muestral de medias sigue una distribución normal y se puede aproximar σ₁ y σ₂ mediante s₁ y s₂.

**Cálculo**

Una vez obtenidos estos datos se desea calcular la diferencia de promedios de profundidad entre la población de buena calidad y la de peor calidad con un intervalo de confianza del 95%. Para esto se hace uso de las siguientes funciones:

BSDA::z.test(x = buenaCalidad$profundidad, y = peorCalidad$profundidad, sigma.x = sd(buenaCalidad$profundidad), sigma.y = sd(peorCalidad$profundidad), conf.level = 0.95)$conf.int

## [1] -0.1935446 -0.1368928  
## attr(,"conf.level")  
## [1] 0.95

PASWR2::z.test(x = buenaCalidad$profundidad, y = peorCalidad$profundidad, sigma.x = sd(buenaCalidad$profundidad), sigma.y = sd(peorCalidad$profundidad), conf.level = 0.95)$conf.int

## [1] -0.1935446 -0.1368928  
## attr(,"conf.level")  
## [1] 0.95

Ambas funciones aplican para una y dos poblaciones y están compuestas de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **y:** vector numérico que representa la segunda muestra (opcional)
* **sigma.x:** desviación estándar de x (opcional)
* **sigma.y:** desviación estándar de y (opcional)
* **conf.level:** nivel de confianza entre 0 y 1 (opcional, se asume 0.95 por defecto)

**Conclusiones**

De los resultados obtenidos para la diferencia de promedios se puede concluir que la media de profundidad de diamantes con colores de baja calidad es mayor que la diamantes con colores de alta calidad.

## IC de una diferencia de promedios usando distribución t

Para este caso se utilizarán muestras aleatorias sin reemplazo de tamaño 25 de los dos grupos utilizados en el cálculo anterior (diamantes con color de buena calidad y diamantes con color de peor calidad).

# Se sacan los datos de la calidad  
buenaCalidad <- diamantes[diamantes$color == "D" | diamantes$color == "E" | diamantes$color == "F",]  
peorCalidad <- diamantes[diamantes$color == "H" | diamantes$color == "I" | diamantes$color == "J",]  
  
# Se toman las filas para las muestras aleatorias  
set.seed(6894)  
filas.randomMejorCalidad <- sample(1:length(buenaCalidad$color), 25, replace= F)  
filas.randomPeorCalidad <- sample(1:length(peorCalidad$color), 25, replace= F)  
  
muestraMejorCalidad <- as.data.frame(buenaCalidad[filas.randomMejorCalidad,])  
muestraPeorCalidad <- as.data.frame(peorCalidad[filas.randomPeorCalidad,])  
  
length(muestraMejorCalidad$color)

## [1] 25

length(muestraPeorCalidad$color)

## [1] 25

|  |  |  |
| --- | --- | --- |
| Calidad | Colores | Tamaño de muestra |
| Buena | D, E o F | 25 |
| Peor | H, I o J | 25 |

En este caso dado que los tamaños de las muestras son menores a 30 se ha optado por realizar una inspección visual de los datos para ver si se asemejan a una distribución normal





**Hipótesis asumidas**

Como se puede observar ambas muestras tienen forma de campana, por lo que se puede asumir que siguen una distribución normal. Además, también se asume que las varianzas son iguales. Sin embargo, como las muestras son pequeñas no se pueden utilizar s₁ y s₂ para realizar una aproximación σ₁ y σ₂.

**Cálculo**

Una vez obtenidos estos datos se desea calcular la diferencia de promedios de profundidad entre las muestras de buena calidad y las de peor calidad con un intervalo de confianza del 95%. Para esto se hace uso de la siguiente función:

t.test(x=muestraMejorCalidad$profundidad, y=muestraPeorCalidad$profundidad, conf.level = 0.95, var.equal = TRUE)$conf.int

## [1] -0.844178 0.436178  
## attr(,"conf.level")  
## [1] 0.95

Donde:

* **x:** primera muestra
* **y:** segunda muestra
* **conf.level:** nivel de confidencia
* **var.equal:** si las varianzas se asumen iguales

**Conclusión**

A diferencia del caso anterior, dado a que las muestras aquí son más pequeñas, se puede observar que hay tanto valores negativos como positivos, por lo cual no se puede llegar a una conclusión con certeza ya que existe la posibilidad de que ambas muestras sean iguales o que una sea mayor que otra.

## IC de una varianza

A continuación, se presenta IC de la varianza de la variable quilates del dataset diamantes.

|  |  |
| --- | --- |
| Muestra | Tamaño |
| A | 53940 |

Una vez tenemos la muestra, utilizaremos la variable de quilates para obtener el IC de varianza de la muestra, para esto utilizaremos la librería EnvStats y específicamente su método varTest que nos dará el intervalo que estamos buscando.

Antes de realizar el cálculo debemos tener en cuenta lo siguiente:

* El nivel de confianza utilizado será de 95%
* La muestra sigue una distribución normal

**Cálculo**

El código utilizado para este cálculo fue:

P1 <- varTest(diamantes$quilate, conf.level = 0.95)$conf.int  
P1

## LCL UCL   
## 0.2220290 0.2273925   
## attr(,"conf.level")  
## [1] 0.95

Donde:

* **Primer parámetro:** primera muestra
* **conf.level:** nivel de confidencia

**Conclusión**

Dando como resultado el IC del 95% para la varianza de quilates: ]0.222029, 0.2273925[, como podemos ver el intervalo es muy pequeño y esto se debe a que se usaron una gran cantidad de datos para realizar el cálculo, dando un resultado muy preciso.

## IC de un cociente de varianzas

Gracias a la librería stests podemos realizar el cociente de varianza de dos poblaciones, para esto dividiremos el dataset de diamantes de dos poblaciones la primera contendrá los diamantes que tengan una alta claridad, teniendo como la más baja calidad posible para esta población VS1 y una segunda población que estará compuesta por los diamantes que tengan una claridad igual o inferior a VS2.

claridadAlta <- diamantes[diamantes$claridad == "IF" | diamantes$claridad == "VVS1" | diamantes$claridad == "VVS2" |  
diamantes$claridad == "VS1",]  
length(claridadAlta$quilate)

## [1] 18682

clariadadBaja <- diamantes[diamantes$claridad == "VS2" | diamantes$claridad == "SI1" | diamantes$claridad == "SI2" |   
diamantes$claridad == "I1",]  
length(clariadadBaja$quilate)

## [1] 35258

|  |  |  |
| --- | --- | --- |
| Calidad | Claridad | Tamaño de muestra |
| Alta | IF, VVSS1, VVSS2, VS1 | 18682 |
| Baja | VS2, SI1, SI2, I1 | 35258 |

Antes de realizar el cálculo debemos tener en cuenta lo siguiente:

* El nivel de confianza utilizado será de 95%.
* La población 1 y la población 2 se comportan normalmente.

**Cálculo**

El código utilizado para este cálculo fue:

R <- stests :: var.test(claridadAlta$quilate, clariadadBaja$quilate,conf.level = 0.95)$conf.int  
R

## [1] 0.6059488 0.6371236  
## attr(,"conf.level")  
## [1] 0.95

Donde:

* **Primer parámetro:** primera muestra
* **Segundo parámetro:** segunda muestra
* **conf.level:** nivel de confidencia

**Conclusión**

Dando como resultado: ]0.6059488, 0.6371236[

En conclusión, al resultado obtenido podemos ver que las varianzas de estas dos poblaciones no deberían ser iguales ya que el número 1 no pertenece al intervalo, también podemos decir que la varianza de la segunda población es mayor ya que ambos valores son menores que 1.

# III Parte: Análisis Inferencial (pruebas de hipótesis de una y dos poblaciones)

## Resumen de variables seleccionadas

**Una población:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Tipos | Descripción | Parámetro por estimar (IC) |
| Precio | Cuantitativa | Precio en dólares estadounidenses | Promedio usando distribución z y distribución t |
| Corte | Cualitativa | Calidad del corte (Regular, Bueno, Muy bueno, Premium, Ideal) | Proporción |
| Quilates | Cuantitativa | Peso del diamante | Varianza |

**Dos poblaciones:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Tipos | Descripción | Parámetro por estimar (IC) |
| Corte | Cualitativa | Calidad del corte (Regular, Bueno, Muy bueno, Premium, Ideal) | Diferencia de proporciones |
| Profundidad | Cuantitativa | Porcentaje de la profundidad total en milímetros | Diferencia de promedios usando distribución z y t y cociente de varianzas |

## Prueba de hipótesis para un promedio usando distribución Z

Para este ejemplo se utilizarán dos grupos (los mismos utilizados para la prueba del IC para un promedio usando distribución Z). Utilizando todos los datos de la variable precio extraídos del dataset diamantes. Tal está distribuido de la siguiente manera:

set.seed(4562) # semilla  
  
poblacion <- as.data.frame(diamantes)  
poblacion.precios <- poblacion$precio  
  
n1 <- length(poblacion.precios)  
mu1 <- mean(poblacion.precios)  
var1 <- var(poblacion.precios)  
  
n1

## [1] 53940

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Varianza |
| A | 53940 | 3932.7997219 | 1.5915629^{7} |

**Hipótesis asumidas**

Dado que la muestra es mayor a 30, debido a las condiciones del teorema del límite central, se puede asumir que la población sigue una distribución normal y que se puede aproximar σ₁ mediante s₁.

**Cálculo**

Sea:

* : la media del precio de los diamantes

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de las siguientes funciones:

stests::z.test(  
 x = poblacion.precios,   
 sigma2 = var(poblacion.precios),   
 mu = 4000, alternative = "greater"  
)

##   
## One Sample z-test  
##   
## data: poblacion.precios  
## Z = -3.9121, p-value = 1  
## alternative hypothesis: true mean is greater than 4000  
## 95 percent confidence interval:  
## 3904.545 Inf  
## sample estimates:  
## mean of poblacion.precios   
## 3932.8

BSDA::z.test(  
 x = poblacion.precios,   
 sigma.x = sd(poblacion.precios),   
 mu = 4000, alternative = "greater"  
)

##   
## One-sample z-Test  
##   
## data: poblacion.precios  
## z = -3.9121, p-value = 1  
## alternative hypothesis: true mean is greater than 4000  
## 95 percent confidence interval:  
## 3904.545 NA  
## sample estimates:  
## mean of x   
## 3932.8

PASWR2::z.test(  
 x = poblacion.precios,   
 sigma.x = sd(poblacion.precios),   
 mu = 4000, alternative = "greater"  
)

##   
## One Sample z-test  
##   
## data: poblacion.precios  
## z = -3.9121, p-value = 1  
## alternative hypothesis: true mean is greater than 4000  
## 95 percent confidence interval:  
## 3904.545 Inf  
## sample estimates:  
## mean of x   
## 3932.8

Las funciones BSDA y PASWR2 aplican para una y dos poblaciones y están compuestas de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **sigma.x:** desviación estándar de x (opcional)
* **mu:** el valor de la media o la diferencia de medias en la hipótesis nula
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)

En cambio, las funciones de stests solo aplican para una población, la cual cuenta con los parámetros donde:

* **x:** vector numérico que representa la primera muestra
* **sigma2:** varianza de x (opcional)
* **mu:** el valor de la media o la diferencia de medias en la hipótesis nula
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad | No corresponde |
| Estadístico de prueba |  |
| Región de aceptación | ]3904.545, [ |
| Región de rechazo | ], 3904.545[ |
| Nivel de confianza | 95% |

**Conclusión**

No se encontró evidencia en contra por lo cual se puede asumir que la media de los precios de diamantes es menor o igual a 4000 dólares estadounidenses. Esto concordando también con los resultados del valor P, donde se tiene una aceptación fuerte de .

## Prueba de hipótesis para un promedio usando distribución t

Para este caso se utilizarán muestras las mismas muestras del caso anterior (precio de los diamantes), pero esta vez con una muestra de 20 datos.

## Se toman filas random para sacar del dataset  
set.seed(6894)  
filas.random <- sample(1:53940, 20, replace= T)  
  
## Se generan datasets nuevos con las filas random anteriores   
muestra1 <- as.data.frame(diamantes[filas.random,])  
muestra.precios <- muestra1$precio  
n1 <- length(muestra.precios)  
mu1 <- mean(muestra.precios)  
de1 <- sqrt(var(muestra.precios))  
alpha1 <- 0.05

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Desviación estándar |
| A | 20 | 4629.6 | 5083.1543087 |

En este caso dado que el tamaño de la muestra es menor a 30 se ha optado por realizar una inspección visual de los datos para ver si se asemeja a una distribución normal.



**Hipótesis asumidas**

Como se puede observar, la muestra tiene forma de campana, por lo que se puede asumir que sigue una distribución normal.

**Cálculo**

Sea:

* : la media del precio de los diamantes

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de la siguiente función:

t.test(  
 x = muestra.precios,   
 mu = 4000, alternative = "greater"  
 )

##   
## One Sample t-test  
##   
## data: muestra.precios  
## t = 0.55392, df = 19, p-value = 0.293  
## alternative hypothesis: true mean is greater than 4000  
## 95 percent confidence interval:  
## 2664.219 Inf  
## sample estimates:  
## mean of x   
## 4629.6

La función está compuesta de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **mu:** el valor de la media o la diferencia de medias en la hipótesis nula
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad |  |
| Estadístico de prueba |  |
| Región de aceptación | ]2664.219, [ |
| Región de rechazo | ], 2664.219[ |
| Nivel de confianza | 95% |

**Conclusión**

No se encontró evidencia en contra por lo cual se puede asumir que la media del precio de los diamantes es menor o igual a 4000 dólares estadounidenses. Esto concordando también con los resultados del valor P, donde se tiene una aceptación de .

## Prueba de hipótesis para una proporción

Basado en la experiencia, un joyero afirma que mínimo el 12% de los diamantes de gama alta son de color tipo D.

¿Podemos aceptar la afirmación del joyero?

Para realizar la prueba, se consideran las siguientes hipótesis:

dónde p representa la proporción de que unos diamantes de gama alta elegidos al azar, sean de color tipo D. La prueba a realizar es de cola izquierda.

En R se está implementado la función prop.test, que además también sirve para contrastar dos proporciones por medio de muestras independientes grandes.

Su sintaxis es:

prop.test(x, n, p = ..., alternative = ..., conf.level = ...)

donde:

* **n** es el tamaño de la muestra.
* **x** es el número de éxitos de la muestra.
* **p** es la proporción poblacional que contrastamos.
* **alternative** especifica la hipótesis alternativa, debe ser “two.sided”, “greater” o “less”.
* **conf.level** es el nivel de confianza.

Para obtener el valor n se tomaron los datos del dataset “Diamantes” donde se utilizó una muestra de la variable “corte” incluyéndose todos los diamantes cuyo corte sea de tipo “Premium” o “Ideal”, siendo esta una muestra de diamantes de gama alta, de la siguiente manera:

gamaAlta <- diamantes[diamantes$corte == "Premium" | diamantes$corte == "Ideal",]  
gamaAlta

## # A tibble: 35,342 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 326 0.23 Ideal E SI2 61.5 55 3.95 3.98 2.43  
## 2 326 0.21 Premium E SI1 59.8 61 3.89 3.84 2.31  
## 3 334 0.29 Premium I VS2 62.4 58 4.2 4.23 2.63  
## 4 340 0.23 Ideal J VS1 62.8 56 3.93 3.9 2.46  
## 5 342 0.22 Premium F SI1 60.4 61 3.88 3.84 2.33  
## 6 344 0.31 Ideal J SI2 62.2 54 4.35 4.37 2.71  
## 7 345 0.2 Premium E SI2 60.2 62 3.79 3.75 2.27  
## 8 345 0.32 Premium E I1 60.9 58 4.38 4.42 2.68  
## 9 348 0.3 Ideal I SI2 62 54 4.31 4.34 2.68  
## 10 355 0.24 Premium I VS1 62.5 57 3.97 3.94 2.47  
## # ... with 35,332 more rows

tamañoMuestraGama <- length(gamaAlta$corte)  
tamañoMuestraGama

## [1] 35342

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Alta | Premium e Ideal | 35342 |

Para obtener el valor x se incorporó la variable color del dataset “Diamantes” para saber cuál es el número de diamantes que hay de corte “premium” o “Ideal” (que hemos llamado gama alta) de color tipo D, de la siguiente manera:

exitos.alta <- gamaAlta[gamaAlta$color == "D",]  
exitos.alta

## # A tibble: 4,437 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 404 0.22 Premium D VS2 59.3 62 3.91 3.88 2.31  
## 2 552 0.3 Premium D SI1 62.6 59 4.23 4.27 2.66  
## 3 552 0.3 Ideal D SI1 62.5 57 4.29 4.32 2.69  
## 4 552 0.3 Ideal D SI1 62.1 56 4.3 4.33 2.68  
## 5 2762 0.71 Ideal D SI2 62.3 56 5.73 5.69 3.56  
## 6 2764 0.71 Ideal D SI1 61.9 59 5.69 5.72 3.53  
## 7 2767 0.71 Ideal D SI2 61.6 55 5.74 5.76 3.54  
## 8 2768 0.71 Premium D SI2 61.7 59 5.71 5.67 3.51  
## 9 2770 0.71 Premium D VS2 62.5 60 5.65 5.61 3.52  
## 10 2770 0.76 Ideal D SI2 62.4 57 5.78 5.83 3.62  
## # ... with 4,427 more rows

tamañoMuestraColor <-length(exitos.alta$color)  
tamañoMuestraColor

## [1] 4437

|  |  |  |  |
| --- | --- | --- | --- |
| Gama | Corte | Color | Tamaño de muestra |
| Alta | Premiun e Ideal | D | 4437 |

Antes de realizar las operaciones se debe verificar que np > 5 y nq > 5

Para ello tenemos los siguientes datos:

p = x/n = 4437/35342 = 0.1255446777

np = 35342 \* 0.1255446777 = 4437

q = 1-p = 1 - 0.1255446777 = 0.8744553223

nq = 35342 \* 0.8744553223 = 30905

Como ambos valores son mayores a 5 se procede a utilizar la función prop.test con sus respectivos valores

prop.test(x=length(exitos.alta$color), n=length(gamaAlta$corte), p = 0.12, alternative = "less", conf.level = 0.95)

##   
## 1-sample proportions test with continuity correction  
##   
## data: length(exitos.alta$color) out of length(gamaAlta$corte), null probability 0.12  
## X-squared = 10.237, df = 1, p-value = 0.9993  
## alternative hypothesis: true p is less than 0.12  
## 95 percent confidence interval:  
## 0.0000000 0.1284867  
## sample estimates:  
## p   
## 0.1255447

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado | = 3.199531216 |
| Grados de libertad | no corresponde |
| Estadístico de prueba | = 0.1255447 |
| Región de aceptación | ]0.0000000, 0.1284867[ |
| Región de rechazo | ]0.1284867, 1[ |
| Nivel de confianza | 0.95 |

**Conclusión:**

no puede ser rechazado dado que no se encontró evidencia suficiente en contra, ya que según lo observado el 0.12 se encuentra dentro del intervalo de aceptación y el valor p es mayor que el valor de alfa, por lo que lo ubica en la región de aceptación, y por ende se asume como válida la afirmación del joyero.

## Prueba de hipótesis para una diferencia de proporciones

Para las pruebas de contraste de proporciones para muestras grandes, se utilizará la función prop.test

Su sintaxis es:

prop.test(x, n, p = ..., alternative = ..., conf.level = ...)

donde:

* **n** es un vector de dos entradas de los respectivos tamaños de las muestras.
* **x** es el vector de dos números naturales que representan los éxitos ambas muestras.

Es este ejercicio se contrastará si la proporción de diamantes de gama alta de color tipo D es la misma que la proporción de diamantes de gama baja de color tipo D

Para realizar esta prueba, se considerarán las siguientes hipótesis:

Para obtener el valor n se tomaron dos muestras de la variable “corte”, en la primera muestra se incluyen todos los diamantes cuyo corte sea de tipo “Premium” ó “Ideal” (muestra de gama alta) y la segunda muestra donde se incluyen todos los diamantes cuyo corte sea de tipo “Regular” ó “Bueno” (muestra de gama baja), de la siguiente manera:

gamaAlta <- diamantes[diamantes$corte == "Premium" | diamantes$corte == "Ideal",]  
gamaAlta

## # A tibble: 35,342 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 326 0.23 Ideal E SI2 61.5 55 3.95 3.98 2.43  
## 2 326 0.21 Premium E SI1 59.8 61 3.89 3.84 2.31  
## 3 334 0.29 Premium I VS2 62.4 58 4.2 4.23 2.63  
## 4 340 0.23 Ideal J VS1 62.8 56 3.93 3.9 2.46  
## 5 342 0.22 Premium F SI1 60.4 61 3.88 3.84 2.33  
## 6 344 0.31 Ideal J SI2 62.2 54 4.35 4.37 2.71  
## 7 345 0.2 Premium E SI2 60.2 62 3.79 3.75 2.27  
## 8 345 0.32 Premium E I1 60.9 58 4.38 4.42 2.68  
## 9 348 0.3 Ideal I SI2 62 54 4.31 4.34 2.68  
## 10 355 0.24 Premium I VS1 62.5 57 3.97 3.94 2.47  
## # ... with 35,332 more rows

gamaBaja <- diamantes[diamantes$corte == "Regular" | diamantes$corte == "Bueno",]  
gamaBaja

## # A tibble: 6,516 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 327 0.23 Bueno E VS1 56.9 65 4.05 4.07 2.31  
## 2 335 0.31 Bueno J SI2 63.3 58 4.34 4.35 2.75  
## 3 337 0.22 Regular E VS2 65.1 61 3.87 3.78 2.49  
## 4 339 0.3 Bueno J SI1 64 55 4.25 4.28 2.73  
## 5 351 0.3 Bueno J SI1 63.4 54 4.23 4.29 2.7   
## 6 351 0.3 Bueno J SI1 63.8 56 4.23 4.26 2.71  
## 7 351 0.3 Bueno I SI2 63.3 56 4.26 4.3 2.71  
## 8 402 0.23 Bueno F VS1 58.2 59 4.06 4.08 2.37  
## 9 402 0.23 Bueno E VS1 64.1 59 3.83 3.85 2.46  
## 10 402 0.31 Bueno H SI1 64 54 4.29 4.31 2.75  
## # ... with 6,506 more rows

tamañoMGA <- length(gamaAlta$corte)  
tamañoMGA

## [1] 35342

tamañoMGB <- length(gamaBaja$corte)  
tamañoMGB

## [1] 6516

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Alta | Premium e Ideal | 35342 |

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Baja | Regular y Bueno | 6516 |

Para obtener el valor x se incorporó la variable color del dataset “Diamantes”, con el fin de calcular cuantos diamantes de gama alta y baja son de color tipo D, de la siguiente manera:

exitos.alta <- gamaAlta[gamaAlta$color == "D",]  
exitos.alta

## # A tibble: 4,437 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 404 0.22 Premium D VS2 59.3 62 3.91 3.88 2.31  
## 2 552 0.3 Premium D SI1 62.6 59 4.23 4.27 2.66  
## 3 552 0.3 Ideal D SI1 62.5 57 4.29 4.32 2.69  
## 4 552 0.3 Ideal D SI1 62.1 56 4.3 4.33 2.68  
## 5 2762 0.71 Ideal D SI2 62.3 56 5.73 5.69 3.56  
## 6 2764 0.71 Ideal D SI1 61.9 59 5.69 5.72 3.53  
## 7 2767 0.71 Ideal D SI2 61.6 55 5.74 5.76 3.54  
## 8 2768 0.71 Premium D SI2 61.7 59 5.71 5.67 3.51  
## 9 2770 0.71 Premium D VS2 62.5 60 5.65 5.61 3.52  
## 10 2770 0.76 Ideal D SI2 62.4 57 5.78 5.83 3.62  
## # ... with 4,427 more rows

exitos.baja <- gamaBaja[gamaBaja$color == "D",]  
exitos.baja

## # A tibble: 825 x 10  
## precio quilate corte color claridad profundidad tabla x y z  
## <int> <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 403 0.26 Bueno D VS2 65.2 56 3.99 4.02 2.61  
## 2 403 0.26 Bueno D VS1 58.4 63 4.19 4.24 2.46  
## 3 2788 0.71 Bueno D VS2 63.3 56 5.64 5.68 3.58  
## 4 2816 0.71 Bueno D VS1 63.4 55 5.61 5.69 3.58  
## 5 2827 0.72 Bueno D VS2 64 54 5.68 5.7 3.64  
## 6 2848 0.75 Regular D SI2 64.6 57 5.74 5.72 3.7   
## 7 2858 0.71 Regular D VS2 56.9 65 5.89 5.84 3.34  
## 8 2872 0.7 Bueno D SI1 62.5 56.7 5.59 5.62 3.51  
## 9 2885 0.9 Regular D SI2 66.9 57 6.02 5.9 3.99  
## 10 2891 0.71 Bueno D VS2 62.3 61 5.7 5.73 3.56  
## # ... with 815 more rows

exitosAlta <- length(exitos.alta$color)  
exitosAlta

## [1] 4437

exitosBaja <- length(exitos.baja$color)   
exitosBaja

## [1] 825

|  |  |  |
| --- | --- | --- |
| Gama | Color | Tamaño de muestra |
| Alta | D | 4437 |

|  |  |  |
| --- | --- | --- |
| Gama | Color | Tamaño de muestra |
| Baja | D | 825 |

Antes de realizar las operaciones se debe verificar que np > 5 y nq > 5 para las dos poblaciones

Para ello tenemos los siguientes datos para la primera población:

p1 = x1/n1 = 4437/35342 = 0.1255446777

n1p1 = 35342 \* 0.1255446777 = 4437

q1 = 1-p1 = 1 - 0.1255446777 = 0.8744553223

n1q1 = 35342 \* 0.8744553223 = 30905

Y tenemos los siguientes datos para la segunda población:

p2 = x2/n2 = 825/6516 = 0.126611418

n2p2 = 6516 \* 0.126611418 = 825

q2 = 1-p2 = 1 - 0.126611418 = 0.873388582

n2q2 = 6516 \* 0.873388582 = 5691

Como ambos valores son mayores a 5 se procede a utilizar la función prop.test con sus respectivos valores

prop.test(x = c(length(exitos.alta$color) , length(exitos.baja$color)), n = c(length(gamaAlta$corte), length(gamaBaja$corte)))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(length(exitos.alta$color), length(exitos.baja$color)) out of c(length(gamaAlta$corte), length(gamaBaja$corte))  
## X-squared = 0.047669, df = 1, p-value = 0.8272  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.009939703 0.007806222  
## sample estimates:  
## prop 1 prop 2   
## 0.1255447 0.1266114

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado | = -0.0010667 |
| Grados de libertad | No corresponde |
| Estadístico de prueba | - = 0 |
| Región de aceptación | ]-0.009939703, 0.007806222[ |
| Región de rechazo | ]-1, -0.009939703[ ]0.007806222, +1[ |
| Nivel de confianza | 0.95 |

**Conclusión:**

no puede ser rechazado dado que no se encontró evidencia suficiente para rechazarlo, ya que según lo observado el 0 se encuentra dentro del intervalo de aceptación y el valor p es mayor que el valor de alfa por lo que lo ubica en la región de aceptación, y por ende se asume que la proporción de diamantes de gama alta con color de tipo D se puede considerar similar a la proporción de diamantes de gama baja con color de tipo D.

## Prueba de hipótesis para una diferencia de promedios usando distribución z

Para este ejemplo se utilizarán dos grupos (los mismo utilizados para la prueba del IC). En primer lugar, está el grupo de diamantes que se considera que tienen un color de buena calidad, este grupo está compuesto de aquellos diamantes cuyo color es “D”, “E” o “F”. Luego se tiene el grupo diamantes cuyo color es de peor calidad, el cual está compuesto por los diamantes con colores “H”, “I” o “J”. Ambos grupos están distribuidos de la siguiente manera:

buenaCalidad <- diamantes[diamantes$color == "D" | diamantes$color == "E" | diamantes$color == "F",]  
peorCalidad <- diamantes[diamantes$color == "H" | diamantes$color == "I" | diamantes$color == "J",]  
  
length(buenaCalidad$color)

## [1] 26114

length(peorCalidad$color)

## [1] 16534

|  |  |  |
| --- | --- | --- |
| Calidad | Colores | Tamaño de muestra |
| Buena | D, E o F | 26114 |
| Peor | H, I o J | 16534 |

**Hipótesis asumidas**

Dado que ambas muestras son mayores a 30 se puede asumir que ambas siguen una distribución normal y que se puede aproximar σ₁ y σ₂ mediante s₁ y s₂. Además, ambos grupos son independientes.

**Cálculo**

Sea:

* : la media de la profundidad de los diamantes de buena calidad
* : la media de la profundidad de los diamantes de peor calidad

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de las siguientes funciones:

BSDA::z.test(x = buenaCalidad$profundidad, y = peorCalidad$profundidad, sigma.x = sd(buenaCalidad$profundidad), sigma.y = sd(peorCalidad$profundidad), mu = 0, alternative = "greater")

##   
## Two-sample z-Test  
##   
## data: buenaCalidad$profundidad and peorCalidad$profundidad  
## z = -11.432, p-value = 1  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.1889905 NA  
## sample estimates:  
## mean of x mean of y   
## 61.68331 61.84853

PASWR2::z.test(x = buenaCalidad$profundidad, y = peorCalidad$profundidad, sigma.x = sd(buenaCalidad$profundidad), sigma.y = sd(peorCalidad$profundidad), mu = 0, alternative = "greater")

##   
## Two Sample z-test  
##   
## data: buenaCalidad$profundidad and peorCalidad$profundidad  
## z = -11.432, p-value = 1  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.1889905 Inf  
## sample estimates:  
## mean of x mean of y   
## 61.68331 61.84853

Ambas funciones aplican para una y dos poblaciones y están compuestas de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **y:** vector numérico que representa la segunda muestra (opcional)
* **sigma.x:** desviación estándar de x (opcional)
* **sigma.y:** desviación estándar de y (opcional)
* **mu:** el valor de la media o la diferencia de medias en la hipótesis nula
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad | No corresponde |
| Estadístico de prueba | - = 0 |
| Región de aceptación | ]-0.1889905, +[ |
| Región de rechazo | ]-, -0.1889905[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

No se encontró evidencia en contra por lo cual se puede asumir que la media de la profundidad de diamantes con color de mejor calidad es menor a la de los diamantes con color de peor calidad. Esto concordando también con los resultados del valor P, donde se tiene una aceptación fuerte de .

## Prueba de hipótesis para un cociente de varianzas

Para este caso se utilizarán muestras aleatorias sin reemplazo de tamaño 25 de los dos mismos grupos utilizados en cálculo del IC para diferencia de promedios con distribución t (diamantes con color de buena calidad y diamantes con color de peor calidad).

# Se sacan los datos de la calidad  
buenaCalidad <- diamantes[diamantes$color == "D" | diamantes$color == "E" | diamantes$color == "F",]  
peorCalidad <- diamantes[diamantes$color == "H" | diamantes$color == "I" | diamantes$color == "J",]  
  
# Se toman las filas para las muestras aleatorias  
set.seed(6894)  
filas.randomMejorCalidad <- sample(1:length(buenaCalidad$color), 25, replace= F)  
filas.randomPeorCalidad <- sample(1:length(peorCalidad$color), 25, replace= F)  
  
muestraMejorCalidad <- as.data.frame(buenaCalidad[filas.randomMejorCalidad,])  
muestraPeorCalidad <- as.data.frame(peorCalidad[filas.randomPeorCalidad,])  
  
length(muestraMejorCalidad$color)

## [1] 25

length(muestraPeorCalidad$color)

## [1] 25

|  |  |  |
| --- | --- | --- |
| Calidad | Colores | Tamaño de muestra |
| Buena | D, E o F | 25 |
| Peor | H, I o J | 25 |

Previamente ya se había realizado una prueba de normalidad de estas muestras en el cálculo del IC para diferencia de promedio, sin embargo, aquí se adjunta la prueba de nuevo:





**Hipótesis asumidas**

* El nivel de confianza utilizado será de 95%.
* La población 1 y la población 2 se comportan normalmente.

**Cálculo**

Sea:

* : la media de la profundidad de los diamantes de buena calidad
* : la media de la profundidad de los diamantes de peor calidad

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de la siguiente función:

stests::var.test(x = muestraMejorCalidad$profundidad, y = muestraPeorCalidad$profundidad)

##   
## F test to compare two variances  
##   
## data: muestraMejorCalidad$profundidad and muestraPeorCalidad$profundidad  
## F = 2.8255, num df = 24, denom df = 24, p-value = 0.0137  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.245114 6.411864  
## sample estimates:  
## ratio of variances   
## 2.825509

Donde:

* **x:** primera muestra
* **y:** segunda muestra (opcional)
* **alternative:** si es de cola derecha, doble cola o cola izquierda (opcional)
* **null.value:** por defecto es 1 considerando que la relación entre las varianzas es 1. (opcional)
* **conf.level:** nivel de confidencia (opcional)

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad |  |
| Estadístico de prueba |  |
| Región de aceptación | ]1.245114, 6.411864[ |
| Región de rechazo | ]-, 1.245114[]6.411864, +[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

Se llegó a encontrar evidencia contra por lo cual se puede asumir que la varianza de la profundidad de diamantes con color de mejor calidad es distinta a la de los diamantes con color de peor calidad. Esto concordando también con los resultados del valor P, donde se tiene un rechazo fuerte de .

## Prueba de hipótesis para una diferencia de promedios usando distribución t

Para este caso se utilizarán muestras las mismas muestras del ejemplo anterior (diamantes con color de buena calidad y diamantes con color de peor calidad).

# Se sacan los datos de la calidad  
buenaCalidad <- diamantes[diamantes$color == "D" | diamantes$color == "E" | diamantes$color == "F",]  
peorCalidad <- diamantes[diamantes$color == "H" | diamantes$color == "I" | diamantes$color == "J",]  
  
# Se toman las filas para las muestras aleatorias  
set.seed(6894)  
filas.randomMejorCalidad <- sample(1:length(buenaCalidad$color), 25, replace= F)  
filas.randomPeorCalidad <- sample(1:length(peorCalidad$color), 25, replace= F)  
  
muestraMejorCalidad <- as.data.frame(buenaCalidad[filas.randomMejorCalidad,])  
muestraPeorCalidad <- as.data.frame(peorCalidad[filas.randomPeorCalidad,])  
  
length(muestraMejorCalidad$color)

## [1] 25

length(muestraPeorCalidad$color)

## [1] 25

|  |  |  |
| --- | --- | --- |
| Calidad | Colores | Tamaño de muestra |
| Buena | D, E o F | 25 |
| Peor | H, I o J | 25 |

**Hipótesis asumidas**

Por el ejemplo anterior se puede asumir que ambas muestras siguen una distribución normal y que las varianzas son distintas.

**Cálculo**

Sea:

* : la media de la profundidad de los diamantes de buena calidad
* : la media de la profundidad de los diamantes de peor calidad

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de las siguientes funciones:

t.test(x = muestraMejorCalidad$profundidad, y = muestraPeorCalidad$profundidad, mu = 0, alternative = "greater", var.equal = F)

##   
## Welch Two Sample t-test  
##   
## data: muestraMejorCalidad$profundidad and muestraPeorCalidad$profundidad  
## t = -0.64071, df = 39.097, p-value = 0.7373  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.7404251 Inf  
## sample estimates:  
## mean of x mean of y   
## 61.576 61.780

La función está compuesta de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **y:** vector numérico que representa la segunda muestra (opcional)
* **mu:** el valor de la media o la diferencia de medias en la hipótesis nula
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)
* **var.equal** indica si las varianzas se asumen iguales

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad | 39 |
| Estadístico de prueba | - = 0 |
| Región de aceptación | ]-0.7404251, +[ |
| Región de rechazo | ]-, -0.7404251[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

No se encontró evidencia en contra por lo cual se puede asumir que la media de la profundidad de diamantes con color de mejor calidad es menor a la de los diamantes con color de peor calidad. Esto concordando también con los resultados del valor P, donde se tiene una aceptación fuerte de .

## Prueba de hipótesis para una varianza

Para este caso de hipótesis se estará usando como muestra la variable de quilates del dataset diamantes

quilates <- diamantes$quilate

|  |  |
| --- | --- |
| Muestra | Tamaño |
| Quilates | 53940 |

**Hipótesis asumidas**

Antes de realizar el cálculo es importante asumir que la muestra sigue una distribución normal

**Cálculo**

Sea:

* : la varianza de la variable quilates de la muestra

Y dadas las siguientes hipótesis para la prueba:

Una vez obtenidos estos datos se procede a realizar la prueba de hipótesis haciendo uso de las siguientes funciones:

sigma.test(x = quilates, sigma = 0.47401, alternative = "two.sided", conf.level = 0.95)

##   
## One sample Chi-squared test for variance  
##   
## data: quilates  
## X-squared = 53939, df = 53939, p-value = 0.9977  
## alternative hypothesis: true variance is not equal to 0.2246855  
## 95 percent confidence interval:  
## 0.2220290 0.2273925  
## sample estimates:  
## var of quilates   
## 0.2246867

La función está compuesta de los siguientes parámetros, en donde:

* **x:** vector numérico que representa la primera muestra
* **sigma:** es la desviación estándar de la muestra
* **alternative:** indica si la prueba es de cola izquierda (less), derecha (greater) o de dos colas (two.sided)
* **conf.level:** Indica el nivel de confianza con el que se realiza la prueba

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Valor observado |  |
| Grados de libertad |  |
| Estadístico de prueba |  |
| Región de aceptación | ]0.2220290, 0.2273925[ |
| Región de rechazo | ]-, 0.2220290[ ]0.2273925, +[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

No se encontró evidencia en contra de por lo cual se puede asumir que la varianza de la muestra si es de 0.2246867, esto debido a que el valor de p es muy cercano a 1 y porque qué valor dado por la función sobre la varianza muestral si se encuentra entre el intervalo ]0.2220290, 0.2273925[

# IV Parte: Otras pruebas de hipótesis en R

## Caso 1:

Para este caso se utilizará la base de datos KidsFeet del paquete mosaicData:

str(KidsFeet)

## 'data.frame': 39 obs. of 8 variables:  
## $ name : Factor w/ 36 levels "Abby","Alisha",..: 10 24 36 20 23 34 13 4 14 8 ...  
## $ birthmonth: int 5 10 12 1 2 3 2 6 5 9 ...  
## $ birthyear : int 88 87 87 88 88 88 88 88 88 88 ...  
## $ length : num 24.4 25.4 24.5 25.2 25.1 25.7 26.1 23 23.6 22.9 ...  
## $ width : num 8.4 8.8 9.7 9.8 8.9 9.7 9.6 8.8 9.3 8.8 ...  
## $ sex : Factor w/ 2 levels "B","G": 1 1 1 1 1 1 1 2 2 1 ...  
## $ biggerfoot: Factor w/ 2 levels "L","R": 1 1 2 1 1 2 1 1 2 2 ...  
## $ domhand : Factor w/ 2 levels "L","R": 2 1 2 2 2 2 2 2 2 1 ...

feetsplit <- split(KidsFeet$length, KidsFeet$sex)  
  
str(feetsplit)

## List of 2  
## $ B: num [1:20] 24.4 25.4 24.5 25.2 25.1 25.7 26.1 22.9 27.5 24.8 ...  
## $ G: num [1:19] 23 23.6 26 23.7 24 24.7 26.7 25.5 24 24.5 ...

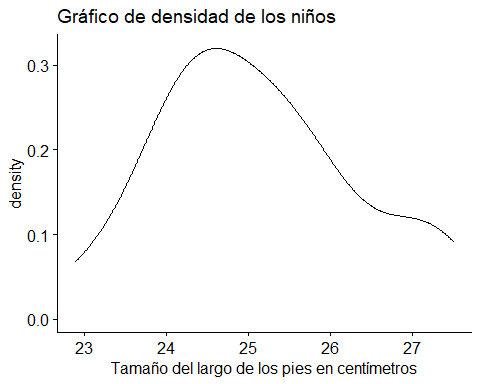
Como pudimos observar, la función split dividió la base de datos en dos, con B para los niños y G para las niñas, se procederá a almacenarla en las siguientes variables:

boys.feets <- feetsplit$B  
girls.feets <- feetsplit$G

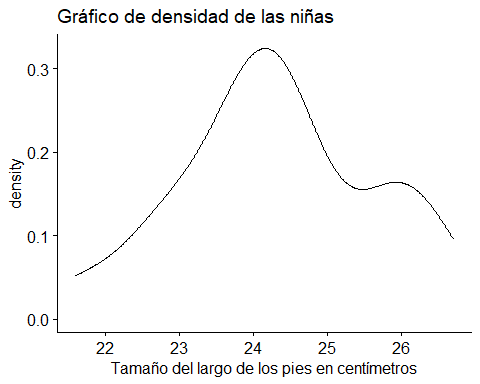
### 1. Gráficos.

Debido a que los tamaños de las muestras de los niños y niñas son de 20 y 19 respectivamente, se pueden crear distintas gráficas para determinar si dicha muestra sigue una distribución normal. A continuación, se presentan los gráficos de densidad.

ggdensity(  
 boys.feets,   
 main = "Gráfico de densidad de los niños",  
 xlab = "Tamaño del largo de los pies en centímetros"  
)



ggdensity(  
 girls.feets,   
 main = "Gráfico de densidad de las niñas",  
 xlab = "Tamaño del largo de los pies en centímetros"  
)



mean(boys.feets)

## [1] 25.105

mean(girls.feets)

## [1] 24.32105

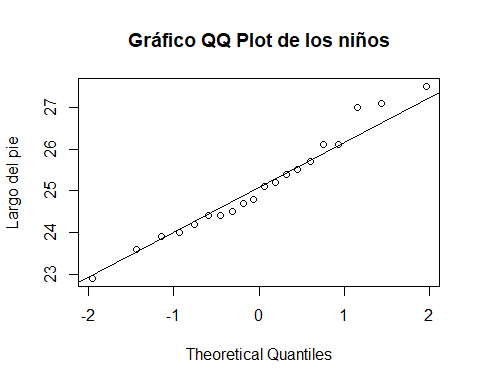
**¿Se puede intuir una posible normalidad para los datos?**

Como se puede observar, probablemente exista normalidad en ambas muestras de datos, ya que parece ser que la gráfica tiene forma de campana, además, en ambos casos la punta de la gráfica se encuentra muy cerca de la media.

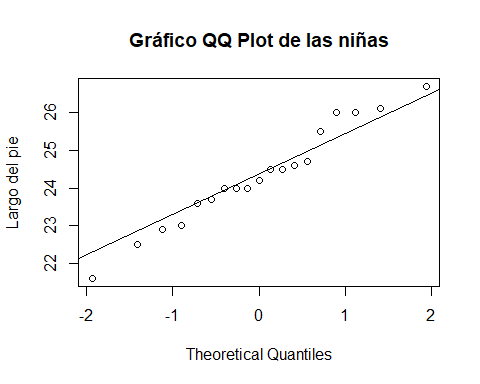
### 2. Gráfico QQ-plot.

A continuación se presentan los gráficos de QQ-plot;

qqnorm(boys.feets, ylab = "Largo del pie", main = "Gráfico QQ Plot de los niños")  
qqline(boys.feets)



qqnorm(girls.feets, ylab = "Largo del pie", main = "Gráfico QQ Plot de las niñas")  
qqline(girls.feets)



En ambos casos, parece ser que el conjunto de datos no parece seguir una distribución normal, especialmente cerca de las colas.

### 3. Pruebas formales de normalidad S-W test, A-D test, K-S-L test.

**Test de Shapito-Wilks o S-W test**

boys.sw <- shapiro.test(boys.feets)  
boys.sw

##   
## Shapiro-Wilk normality test  
##   
## data: boys.feets  
## W = 0.97195, p-value = 0.7954

girls.sw <- shapiro.test(girls.feets)  
girls.sw

##   
## Shapiro-Wilk normality test  
##   
## data: girls.feets  
## W = 0.97303, p-value = 0.835

Para ambos casos, no se encontró evidencia en contra para asumir normalidad, debido a que el valor p es mayor que el nivel de significancia (0.05).

**Test de normalidad de Anderson-Darling o A-D test**

boys.ad <- ad.test(boys.feets)  
boys.ad

##   
## Anderson-Darling normality test  
##   
## data: boys.feets  
## A = 0.24361, p-value = 0.7311

girls.ad <- ad.test(girls.feets)  
girls.ad

##   
## Anderson-Darling normality test  
##   
## data: girls.feets  
## A = 0.26721, p-value = 0.6464

En ambos casos, no se encontró evidencia en contra para asumir normalidad, debido a que el valor p es mayor que el nivel de significancia (0.05).

**Test de Kolmogorov-Smirnov-Lilliefors o K-S-L test**

boys.ksl <- lillie.test(boys.feets)  
boys.ksl

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: boys.feets  
## D = 0.098964, p-value = 0.8746

girls.ksl <- lillie.test(girls.feets)  
girls.ksl

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: girls.feets  
## D = 0.12471, p-value = 0.6094

En ambos casos, no se encontró evidencia en contra para asumir normalidad, debido a que el valor p es mayor que el nivel de significancia (0.05).

### 4. D’Agostino-Pearson.

Para realizar esta prueba, es necesario saber que el tamaño de la muestra obligatoriamente debe ser 20 o superior, por ende, no es posible realizar la prueba para el caso del tamaño de las niñas debido a que dicha muestra es de tamaño 19. Sabiendo esto, se procede a realizar la prueba para el tamaño de la muestra en general y luego para la muestra de los niños:

dagoTest(KidsFeet$length)

##   
## Title:  
## D'Agostino Normality Test  
##   
## Test Results:  
## STATISTIC:  
## Chi2 | Omnibus: 0.0206  
## Z3 | Skewness: 0.1242  
## Z4 | Kurtosis: 0.0722  
## P VALUE:  
## Omnibus Test: 0.9897   
## Skewness Test: 0.9012   
## Kurtosis Test: 0.9425   
##   
## Description:  
## Mon Jun 14 20:36:11 2021 by user: fran-

boys.dap <- dagoTest(boys.feets)  
boys.dap

##   
## Title:  
## D'Agostino Normality Test  
##   
## Test Results:  
## STATISTIC:  
## Chi2 | Omnibus: 0.6384  
## Z3 | Skewness: 0.7831  
## Z4 | Kurtosis: -0.1584  
## P VALUE:  
## Omnibus Test: 0.7267   
## Skewness Test: 0.4335   
## Kurtosis Test: 0.8742   
##   
## Description:  
## Mon Jun 14 20:36:11 2021 by user: fran-

No se encontró evidencia en contra para asumir normalidad, debido a que el valor p general (Omnibus) es mayor que el nivel de significancia (0.05).

### 5. Análisis de custoris y simetría de los datos.

A continuación, se mostrarán los análisis de custoris en niños y niñas para el tamaño de la planta del pie:

boys.kurtosis <- moments::kurtosis(boys.feets) # curtosis en niños  
boys.kurtosis

## [1] 2.46336

girls.kurtosis <- moments::kurtosis(girls.feets) # curtosis en niñas  
girls.kurtosis

## [1] 2.496249

Como podemos observar, ambos resultados fueron mayores que cero, por lo que se puede concluir que ambos tienden a ser leptocúrtica, es decir, su campana es alargada y muy punteada, con unas colas muy bajas.

A continuación se mostrarán los análisis de simetría en niños y niñas para el tamaño de la planta del pie:

boys.skewness <- moments::skewness(boys.feets) # simetria en niños  
boys.skewness

## [1] 0.3526679

girls.skewness <- moments::skewness(girls.feets) # simetria en niñas  
girls.skewness

## [1] -0.04344937

De la simetría se puede concluir que ambas gráficas son simétricas, aunque en el caso de los niños, al ser positivo, su media está un poco más inclinada hacia la izquierda, y en las niñas ocurre lo contrario, al ser negativo pero acercándose a 0, su media está ligeramente más inclinada hacia la derecha.

## Caso 2:

Para este caso se utilizará la base de datos medidas\_cuerpo de la siguiente url:

url <- 'https://raw.githubusercontent.com/fhernanb/datos/master/medidas\_cuerpo'   
body <- read.table(file=url, header=T)  
  
str(body)

## '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 ...

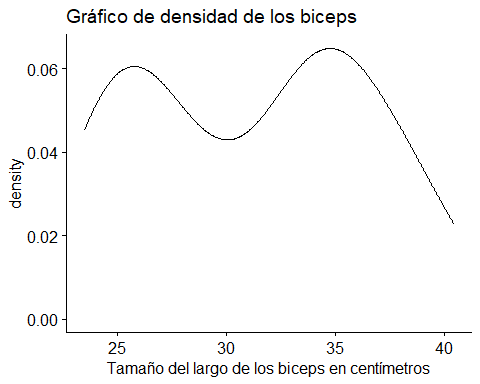
biceps <- body$biceps  
str(biceps)

## num [1:36] 35.8 35 38.5 32.2 32.9 38.5 38.3 35 32.1 40.4 ...

### 1. Gráficos.

A continuación, se presentan los gráficos de densidad para el tamaño de los bíceps:

ggdensity(  
 biceps,   
 main = "Gráfico de densidad de los biceps",  
 xlab = "Tamaño del largo de los biceps en centímetros"  
)

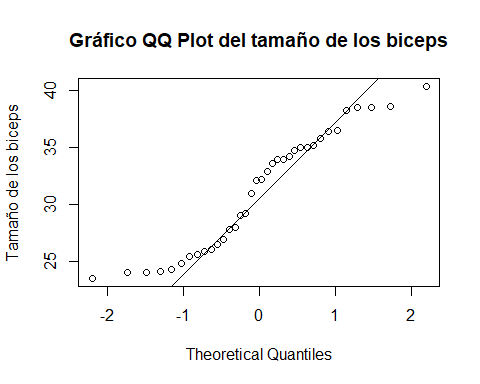


Se puede observar que la gráfica no tiene forma de campana y cerca de la media no existe un punto máximo, por lo que se concluye que dicha muestra no sigue una distribución normal.

### 2. Gráfico QQ-plot.

A continuación, se presentan los gráficos de QQ-plot, es importante saber que se utilizarán las edades para que así el gráfico del tamaño de los bíceps tenga sentido:

#qqplot(biceps, body$edad, xlab = "Tamaño de los biceps", ylab = "Edad de la persona", main = "Gráfico QQ Plot de los biceps")  
qqnorm(biceps, ylab = "Tamaño de los biceps", main = "Gráfico QQ Plot del tamaño de los biceps")  
qqline(biceps)



Se puede observar que el conjunto de datos no parece seguir una distribución normal, especialmente en las colas y en el centro de la gráfica.

### 3. Pruebas formales de normalidad S-W test, A-D test, K-S-L test.

**Test de Shapito-Wilks o S-W test**

biceps.sw <- shapiro.test(biceps)  
biceps.sw

##   
## Shapiro-Wilk normality test  
##   
## data: biceps  
## W = 0.92377, p-value = 0.01627

Se encontró evidencia en contra para asumir normalidad, debido a que el valor p es menor que el nivel de significancia de 0.05.

**Test de normalidad de Anderson-Darling o A-D test**

biceps.ad <- ad.test(biceps)  
biceps.ad

##   
## Anderson-Darling normality test  
##   
## data: biceps  
## A = 0.94278, p-value = 0.0151

Se encontró evidencia en contra para asumir normalidad, debido a que el valor p es menor que el nivel de significancia de 0.05.

**Test de Kolmogorov-Smirnov-Lilliefors o K-S-L test**

biceps.ksl <- lillie.test(biceps)  
biceps.ksl

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: biceps  
## D = 0.12583, p-value = 0.1588

Se encontró evidencia en contra para asumir normalidad, debido a que el valor p es menor que el nivel de significancia de 0.05.

### 4. D’Agostino-Pearson.

biceps.dap <- dagoTest(biceps)  
biceps.dap

##   
## Title:  
## D'Agostino Normality Test  
##   
## Test Results:  
## STATISTIC:  
## Chi2 | Omnibus: 13.6459  
## Z3 | Skewness: 0.0162  
## Z4 | Kurtosis: -3.694  
## P VALUE:  
## Omnibus Test: 0.001088   
## Skewness Test: 0.9871   
## Kurtosis Test: 0.0002208   
##   
## Description:  
## Mon Jun 14 20:36:12 2021 by user: fran-

Se encontró evidencia en contra para asumir normalidad, debido a que el valor p general (Omnibus) es menor que el nivel de significancia de 0.05.

### 5. Análisis de custoris y simetría de los datos

A continuación, se mostrarán el análisis de custoris para el tamaño de los bíceps:

biceps.kurtosis <- moments::kurtosis(biceps) # curtosis de los biceps  
biceps.kurtosis

## [1] 1.626789

Como se puede observar, el resultado fue mayor a cero, por lo que se puede concluir que tiende a ser leptocúrtica, es decir, su campana es alargada y muy punteada, con unas colas muy bajas.

A continuación, se mostrará el análisis de simetría para el tamaño de los bíceps:

biceps.skewness <- moments::skewness(biceps) # simetria en medida de biceps  
biceps.skewness

## [1] 0.005754543

De la simetría se puede concluir que la gráfica es simétrica, aunque, al ser positivo, su media está un poco más inclinada hacia la izquierda.

## Resumen casos 1 y 2

|  |  |  |  |
| --- | --- | --- | --- |
| Tipo de prueba | Niños | Niñas | Medidas de bíceps |
| Función de densidad versus curva normal | Parece tener forma de campana | Parece tener forma de campana | No tiene forma de campana |
| Conclusión | Se concluye que sigue una dist. normal | Se concluye que sigue una dist. normal | Se concluye que no sigue una dist. normal |
| QQ-Plot | En las colas y el centro de la gráfica los datos están más alejados de la línea. | En las colas y el centro de la gráfica los datos están más alejados de la línea. | En las colas y el centro de la gráfica los datos están más alejados de la línea. |
| Conclusión | No parece seguir una distribución normal | No parece seguir una distribución normal | No parece seguir una distribución normal |
| S-W test | Valor P=0.7954 | Valor P=0.835 | Valor P=0.0163 |
| Conclusión | Se asume normalidad; valor p > 0.05 (significancia) | Se asume normalidad; valor p > 0.05 (significancia) | NO se asume normalidad; valor p < 0.05 (significancia) |
| A-D test | Valor P=0.7311 | Valor P=0.6464 | Valor P=0.0151 |
| Conclusión | Se asume normalidad; valor p > 0.05 (significancia) | Se asume normalidad; valor p > 0.05 (significancia) | NO se asume normalidad; valor p < 0.05 (significancia) |
| D’Agostino-Pearson | Valor P=0.7267 | Valor P=NA | Valor P=0.00108 |
| Conclusión | Se asume normalidad; valor p > 0.05 (significancia) | No aplica, n=19 y se requiere mínimo 20 | NO se asume normalidad; valor p < 0.05 (significancia) |
| Curtosis y simetría | 2.4634 y 0.3527 | 2.4962 y -0.0434 | 1.6268 y 0.0058 |
| Conclusión | Leptocúrtica y simétrica | Leptocúrtica y simétrica | Leptocúrtica y simétrica |
| Conclusión general sobre normalidad de los datos | Se asume normalidad ya que a pesar de que gráficamente no lo parezca, el valor p es un dato más acertado que el factor visual | Se asume normalidad ya que a pesar de que gráficamente no lo parezca, el valor p es un dato más acertado que el factor visual | NO se asume normalidad |

## Caso 3:

La siguiente tabla resume los datos de obtenidos de víctimas de crímenes elegidas al azar (según datos del Departamento de Justicia de USA):

|  |  |  |  |
| --- | --- | --- | --- |
| Descripción | HOMICIDIO | ROBO | ASALTO |
| **El criminal era un extraño** | 12 | 379 | 727 |
| **El criminal era un conocido o pariente** | 39 | 106 | 642 |

Con los datos anteriores, ¿sería posible considerar que el tipo de delito es independiente de la condición del delincuente?, o por el contrario, ¿existe alguna relación entre el tipo de delito con respecto al quien comete el acto? Se asumen un nivel de significancia de 5%.

Para este caso se plantea una prueba de independencia con las siguientes hipótesis:

:| El tipo de delito es independiente de la condición del delincuente

:| No existe independencia entre el tipo de delito y la condición del delincuente

Para las pruebas de independencia se utilizará la función chisq.test

Su sintaxis es:

chisq.test(matrix(c(valores matriz), tamaño matriz))

Y se procede a utilizar la función chisq.test con los respectivos valores de la matriz que se presentó al inicio y su tamaño el cual es 2x3 (2 filas y 3 columnas)

chisq.test(matrix(c(12,379,727,39,106,642),2,3,byrow = TRUE))

##   
## Pearson's Chi-squared test  
##   
## data: matrix(c(12, 379, 727, 39, 106, 642), 2, 3, byrow = TRUE)  
## X-squared = 119.33, df = 2, p-value < 2.2e-16

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Tipo de prueba: | Independencia |
| Valor observado | 119.33 |
| Grados de libertad | 2, representan el total de filas -1 multiplicado por el total de columnas -1 |
| Valor P | 2.2e-16 |
| : | El tipo de delito es independiente de la condición del delincuente |
| : | No existe independencia entre el tipo de delito y la condición del delincuente |

**Conclusión**

Se cuenta con suficiente evidencia para rechazar ya que el valor p es muy cercano a 0, por lo tanto, no se puede aceptar independencia del tipo de delito con respecto a la condición del delincuente

## Caso 4:

Considerando la siguiente situación

La seguridad de los automóviles se determina mediante diversas pruebas. Una de ellas consiste en hacer chocar un automóvil contra una barrera fija a 35 𝑚𝑖/ℎ con un maniquí colocado en el asiento del conductor. A una de las medidas utilizadas para cuantificar el impacto del choque sobre el conductor se le conoce como **Desaceleración de pecho** y se mide en unidades de fuerza de gravedad (𝑔). Los valores más grandes indican mayores cantidades de desaceleración, las cuáles pueden provocar lesiones graves en los conductores. La siguiente tabla muestra mediciones de desaceleraciones de pecho obtenidas a partir de pruebas de choques de diferentes tipos de vehículos:

|  |  |  |
| --- | --- | --- |
| Autos compactos | Autos medianos | Autos grandes |
| 44 | 41 | 32 |
| 43 | 49 | 37 |
| 44 | 43 | 38 |
| 54 | 41 | 45 |
| 38 | 47 | 37 |
| 43 | 44 | 33 |
| 42 | 37 | 38 |
| 45 | 34 | 45 |
| 50 |  | 43 |
|  |  | 42 |

Con los datos anteriores, ¿es posible considerar que el tamaño del automóvil puede variar en cuanto a la seguridad de sus pasajeros o por el contrario, es igualmente riesgoso? Se asume un nivel de significancia del 5%.

Para este caso se plantean las siguientes hipótesis:

Sea:

X: las unidades de fuerza de gravedad para cuantificar el impacto del choque (Variable cuantitativa)

Y: tipos de autos utilizados en las pruebas de choque (Variable cualitativa)

μ1, μ2 y μ3 respectivamente las unidades de fuerza de gravedad promedio en los choques de cada tipo de auto.

: μ1=μ2=μ3.

: al menos dos de las medias no son iguales.

Creamos un archivo csv con los datos de la tabla anterior y lo cargamos en la variable “datos”

library(readr)  
  
datos <- read.csv("caso4.csv", sep = ";")   
head(datos, 5)

## gravedad autos  
## 1 44 compactos  
## 2 43 compactos  
## 3 44 compactos  
## 4 54 compactos  
## 5 38 compactos

Aplicamos la prueba de varianza con la función aov y datos del csv (el summary es para hacer un resumen de los resultados):

summary(aov(gravedad ~ autos, data = datos))

## Df Sum Sq Mean Sq F value Pr(>F)   
## autos 2 158.5 79.26 3.552 0.0445 \*  
## Residuals 24 535.6 22.31   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Resumen de la prueba**

|  |  |
| --- | --- |
| Dato | Valor |
| Tipo de prueba: | ANOVA |
| Valor observado | 3.552 |
| Grados de libertad autos | 2, cantidad de tipos de autos - 1 |
| Grados de libertad residuals | 24, representan el total de filas(datos de fuerza de gravedad) -1 |
| SSE | 535.6 |
| SSA | 158.5 |
| SST | 694.1 |
| S1 | 79.26 |
| S2 | 22.31 |
| Valor P | 0.0445 |
| : | μ1=μ2=μ3 |
| : | al menos dos de las medias no son iguales |

**Conclusión**

En esta prueba el p-valor está dado por el resultado Pr(>F), que en este caso al ser menor que α, se rechaza la hipótesis nula, la cual indica que μ1, μ2 y μ3 son iguales, por lo que si es posible asegurar que el tamaño del automóvil puede variar en cuanto a la seguridad de sus pasajeros

**4.2**

Para decir en cuáles tipos de autos existe una verdadera diferencia significativa realizaremos una prueba Tuckey

Para despejar esta duda, se recomienda utilizar la prueba Tuckey que permite calcular la diferencia de impacto de cada valor en cada variable independiente. La prueba Tuckey la realizamos de la siguiente manera:

# Almacenamos la prueba de ANOVA en una nueva variable  
AnovaCaso4 <- aov(gravedad ~ autos, data = datos)  
  
#Pasamo la prueba ANOVa a la prueba Tuckey  
TukeyHSD(AnovaCaso4)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = gravedad ~ autos, data = datos)  
##   
## $autos  
## diff lwr upr p adj  
## grandes-compactos -5.777778 -11.198045 -0.3575109 0.0351275  
## medianos-compactos -2.777778 -8.510004 2.9544486 0.4588276  
## medianos-grandes 3.000000 -2.595725 8.5957252 0.3881731

Del resumen anterior, debe comprenderse lo siguiente:

diff: corresponde a la resta de los promedios muestrales entre las dos categorías comparadas

lwr: extremo izquierdo del IC

upr: extremo derecho del IC

p adj: valor P ajustado (para efectos prácticos, el valor P de la prueba)

**Conclusión**

La prueba muestra que hay una diferencia significativa entre autos “grandes” y “compactos” donde valor P=0.0351275 < α = 0.05, no así en el resto de comparaciones.

# V Parte: Modelos de regresión lineal

## Caso 5:

Primeramente, se cargan los datos de **EdadPesoGrasas.txt** como se muestra a continuación:

grasas <- read.table('http://verso.mat.uam.es/~joser.berrendero/datos/Eda  
dPesoGrasas.txt', header = TRUE)

Con estos datos se proceden a realizar los siguientes puntos de un análisis de regresión lineal:

### 1. Análisis de correlación entre todas la variables.

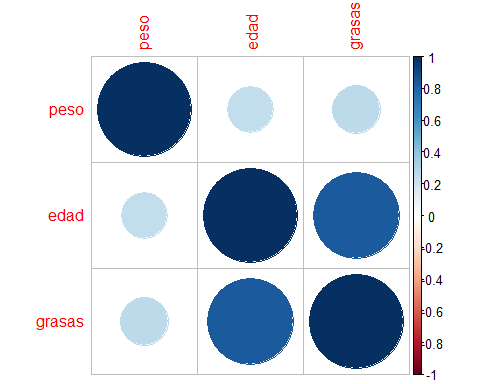
Para esto se hace uso de la función cor del paquete stats de R, con la cual se puede resumir de manera muy sencilla todos los posibles casos de correlación en una matriz:

c <- cor(grasas)  
c

## peso edad grasas  
## peso 1.0000000 0.2400133 0.2652935  
## edad 0.2400133 1.0000000 0.8373534  
## grasas 0.2652935 0.8373534 1.0000000

Adicional a esta matriz, también se puede hacer uso de la función corrplot del paquete del mismo nombre. Esta función permite representar los datos de la matriz anterior de una manera más intuitiva.

corrplot(c)



De este gráfico se puede observar que conforme los datos están más lejos 0 los cículos son más grandes, esto indica que existe una correlación entre las variables. En caso de que el color sea azul significa que es una correlación positiva y si es rojo es una correlación negativa. Por ejemplo en el caso de las variables edad y grasas se tiene que hay correlación positiva que indica que el aumento de una se debe en gran parte debido a la otra.

### 2. Generación de modelo RLS de mejor ajuste para dos variables con mayor coeficiente de correlación

Para este punto se hará uso de las variables edad y grasas que fueron las que tuvieron un mayor coeficiente de variación según los datos del punto anterior.

De esto se tiene que grasas será la variable Y de respuesta y que edad será la la variable X explicativa. A partir de estos datos se procede a generar el RLS haciendo uso de la función lm del paquete stats de R

modelo <- lm(grasas ~ edad, data = grasas)  
modelo

##   
## Call:  
## lm(formula = grasas ~ edad, data = grasas)  
##   
## Coefficients:  
## (Intercept) edad   
## 102.575 5.321

De esto se tiene que el modelo de RLS sería grasas = 102.575 + 5.321\*edad

### 3. Análisis de la calidad del modelo generado

Para esto se hará uso del coeficiente de correlación y el coeficiente de determinación. Se utilizará la función summary para extraer el parámetro r.squared del modelo calculado previamente y posteriormente se hará el análisis de los coeficientes.

## Cálculo del coeficiente de determinación  
summary(modelo)$r.squared

## [1] 0.7011607

## Cálculo del coeficiente de correlación  
sqrt(summary(modelo)$r.squared)

## [1] 0.8373534

Se tiene que , esto significa que la correlación lineal entre grasas y edad es positiva y moderada. Por otra parte, se tiene que , de esto se interpreta que el 70.11% de variación en las grasas se debe a la edad y el restante 29.89% es por otros factores.

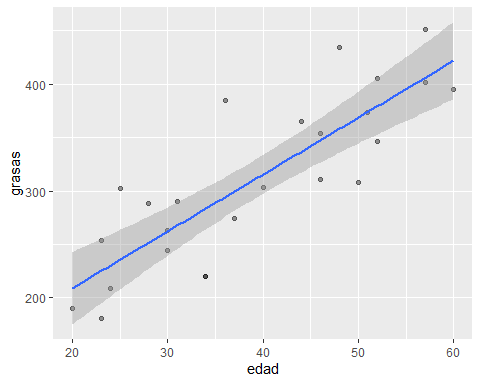
De todo esto se puede llegar a la conclusión de que el modelo tiene una calidad de moderada ya que hay un buen porcentaje otros factores que explican la variación de grasas y además el nivel de correlación no supera el 90%.

### 4. Gráfico de dispersión y recta de mejor ajuste

En este punto se genera un gráfico de RLS mediante el uso de la función geom\_point del paquete ggplot2 de R:

graficoGrasas <- ggplot(grasas, aes(edad, grasas))  
graficoGrasas + geom\_point(alpha = 0.4) + geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



Para un IC del 95% se puede visualizar que la calidad del modelo se encuentra en un nivel moderado, ya que a pesar de que hay varios puntos dentro de la zona del IC hay una cantidad similar que se encuentran por fuera. Esto concordando también con lo visualizado mediante los coeficiente de correlación y determinación.

### 5. Prueba de normalidad de residuos

Con el fin de verificar si el los datos pueden ajustarse de manera lineal, se procede a realizar una prueba para verificar si la distribución de los errores o residuos se comporta de manera normal. Para ello se hace uso de la función lillie.test del paquete nortest de R.

b0 <- 102.575  
b1 <- 5.321  
  
grasas.estimada <- b0 + b1 \* grasas$edad  
errores <- grasas.estimada - grasas$grasas  
lillie.test(errores)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: errores  
## D = 0.10553, p-value = 0.6712

Como el valor P es grande, no existe evidencia suficiente para rechazar la normalidad de los errores, por lo tanto se puede asumir la normalidad de los mismos, por lo que se cumple el principio.

### 6. Intervalo de confianza de 95% para los coeficientes del modelo

Con el objetivo de hallar los IC de los parámetros y , se hace uso de la función confint del paquete stats:

confint(modelo,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 41.265155 163.885130  
## edad 3.822367 6.818986

Entonces se tiene que está entre ]41.265155, 163.885130[ y ]3.822367, 6.818986[

### 7. Para una edad de 27 años, se determinará un IC de 95% para 𝜇𝑌|𝑥=27 y un intervalo de predicción para los valores de Y asociados a dicha edad.

Para el cálculo de estos dos parámetros se hace uso de la función predict.lm del paquete stats, la cual funciona tanto para intervalos de confidencia como de predicción.

Primero se calcula el intervalo de confidencia:

x0 <- data.frame(edad = 27)  
predict.lm(modelo, x0, interval = "confidence", level = 0.95)

## fit lwr upr  
## 1 246.2334 220.6777 271.7891

Y se tiene que se encuentra entre ]220.6777, 271.7891[ con un nivel de confianza del 95%. Y significa que se espera que la media de grasa para una edad de 27 años se encuentre en ese intervalo el 95% de las veces.

Luego se calcula el intervalo de predicción:

x0 <- data.frame(edad = 27)  
predict.lm(modelo, x0, interval = "prediction", level = 0.95)

## fit lwr upr  
## 1 246.2334 152.7653 339.7015

Y se tiene que el intervalo de predicción con IC del 95% es ]152.7653, 339.7015[, lo que significa que se espera que el valor de grasa para una edad de 27 años se encuentre en ese intervalo el 95% de las veces.

### 8. Verificación de la linealidad entre las variables y de dependencia lineal entre estas.

Primero se tiene las siguientes hipótesis:

* (no hay linealidad)

Para probar la siguiente hipótesis se hará uso de la función summary, esto con el obtener un resumen de los datos del modelo.

summary(modelo)

##   
## Call:  
## lm(formula = grasas ~ edad, data = grasas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -63.478 -26.816 -3.854 28.315 90.881   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 102.5751 29.6376 3.461 0.00212 \*\*   
## edad 5.3207 0.7243 7.346 1.79e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.46 on 23 degrees of freedom  
## Multiple R-squared: 0.7012, Adjusted R-squared: 0.6882   
## F-statistic: 53.96 on 1 and 23 DF, p-value: 1.794e-07

Como en la pendiente el valor P es menor a 0.05 se rechaza por lo que se puede asumir que la pendiente es distinta de 0 lo que significa que hay linealidad y que hay dependencia lineal entre las variables grasas y edad.

## Caso 6:

Supóngase que el departamento de ventas de una empresa quiere estudiar la influencia que tiene la publicidad a través de distintos medios de comunicación, sobre el número de ventas de un producto. Se dispone de un conjunto de datos que contiene los ingresos (en millones) conseguido por ventas en 200 regiones, así como la cantidad de presupuesto, también en millones, destinado a anuncios por radio, TV y periódicos en cada una de ellas.

En primer lugar, se cargan los datos del documento DatosVentas.csv

ventasdb <- read.csv("DatosVentas.csv", header = TRUE)

### Generación de los modelos.

### 1. Regresión lineal simple.

En primer lugar, se creará un modelo con las siguientes variables y = ventas y x = tv

mod0 <- lm(ventas ~ tv, data = ventasdb)  
mod0

##   
## Call:  
## lm(formula = ventas ~ tv, data = ventasdb)  
##   
## Coefficients:  
## (Intercept) tv   
## 7.03259 0.04754

Tras generar el modelo se obtiene la ecuación:

ventas = 7.03259 + 0.04754 \* tv

### 2. Primer modelo de regresión lineal múltiple.

Para este punto se crea un modelo con las siguientes variables: y = ventas, x1 = tv, x2 = radio y x3 = periódico.

mod1 <- lm(ventasdb$ventas ~ ventasdb$tv+ventasdb$radio+ventasdb$periodico)

Y se tiene que la ecuación de este es:

ventas = 2.938889 + 0.045765 \* tv + 0.188530 \* radio + -0.001037 \* periódico

Posteriormente se procede a hacer un resumen con el objetivo de poder identificar los coeficientes que sean significativos para el modelo:

summary(mod1)

##   
## Call:  
## lm(formula = ventasdb$ventas ~ ventasdb$tv + ventasdb$radio +   
## ventasdb$periodico)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8277 -0.8908 0.2418 1.1893 2.8292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.938889 0.311908 9.422 <2e-16 \*\*\*  
## ventasdb$tv 0.045765 0.001395 32.809 <2e-16 \*\*\*  
## ventasdb$radio 0.188530 0.008611 21.893 <2e-16 \*\*\*  
## ventasdb$periodico -0.001037 0.005871 -0.177 0.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.686 on 196 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956   
## F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

Tras obtener los coeficientes y realizar un summary al modelo podemos concluir que los que aportan o son significantes para a modelo son los de la variable tv, radio y el intercepto, mientras que la variable periódico no es significativa para el modelo.

### 3. Segundo modelo de regresión lineal múltiple.

En este punto se crea un nuevo modelo con las variables significativas. En este caso: y = ventas y x1 = tv y x2 = radio

mod2 <- lm(ventasdb$ventas ~ ventasdb$tv+ventasdb$radio)  
mod2

##   
## Call:  
## lm(formula = ventasdb$ventas ~ ventasdb$tv + ventasdb$radio)  
##   
## Coefficients:  
## (Intercept) ventasdb$tv ventasdb$radio   
## 2.92110 0.04575 0.18799

De esto se obtiene la ecuación:

ventas = 2.92110 + 0.04575 \* tv + 0.18799 \* radio

Y luego se vuelve a realizar un summary con el objetivo de ver si las variables son significativas.

summary(mod2)

##   
## Call:  
## lm(formula = ventasdb$ventas ~ ventasdb$tv + ventasdb$radio)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7977 -0.8752 0.2422 1.1708 2.8328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.92110 0.29449 9.919 <2e-16 \*\*\*  
## ventasdb$tv 0.04575 0.00139 32.909 <2e-16 \*\*\*  
## ventasdb$radio 0.18799 0.00804 23.382 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.681 on 197 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962   
## F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

En este caso tras realizar el summary del modelo, podemos ver todas las variables son significativas para el modelo.

### 4. Modelo de regresión no lineal múltiple.

Utilizando los mismos datos se procede a realizar un RNLM.

mod3 <- lm(ventasdb$ventas ~ ventasdb$tv + ventasdb$radio + (ventasdb$tv\*ventasdb$radio))  
mod3

##   
## Call:  
## lm(formula = ventasdb$ventas ~ ventasdb$tv + ventasdb$radio +   
## (ventasdb$tv \* ventasdb$radio))  
##   
## Coefficients:  
## (Intercept) ventasdb$tv   
## 6.750220 0.019101   
## ventasdb$radio ventasdb$tv:ventasdb$radio   
## 0.028860 0.001086

Y se obtiene la ecuación:

ventas = 6.750220 + 0.019101 \* tv + 0.028860 \* radio + 0.001086(tv\*radio)

Y se verifica si las variables son significativas:

summary(mod3)

##   
## Call:  
## lm(formula = ventasdb$ventas ~ ventasdb$tv + ventasdb$radio +   
## (ventasdb$tv \* ventasdb$radio))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3366 -0.4028 0.1831 0.5948 1.5246   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 \*\*\*  
## ventasdb$tv 1.910e-02 1.504e-03 12.699 <2e-16 \*\*\*  
## ventasdb$radio 2.886e-02 8.905e-03 3.241 0.0014 \*\*   
## ventasdb$tv:ventasdb$radio 1.086e-03 5.242e-05 20.727 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9435 on 196 degrees of freedom  
## Multiple R-squared: 0.9678, Adjusted R-squared: 0.9673   
## F-statistic: 1963 on 3 and 196 DF, p-value: < 2.2e-16

Tras obtener los resultados y realizar el summary podemos ver que todos los coeficientes obtenidos son significativos y aportan al modelo.

### Seleccionando el mejor modelo.

Para esta sección, se pretende seleccionar el mejor modelo entre los 4 ya generados en la I Parte de esta prueba.

## Cálculo del coeficiente de determinación mod0  
cd0 <- summary(lm(mod0))$r.squared  
  
## Cálculo del coeficiente de correlación mod0  
cc0 <- sqrt(summary(mod0)$r.squared)  
  
## Cálculo del coeficiente de determinación mod1  
cd1 <- summary(lm(mod1))$r.squared  
  
## Cálculo del coeficiente de correlación mod1  
cc1 <- sqrt(summary(mod1)$r.squared)  
  
## Cálculo del coeficiente de determinación mod2  
cd2 <- summary(lm(mod2))$r.squared  
cd2

## [1] 0.8971943

## Cálculo del coeficiente de correlación mod2  
cc2 <- sqrt(summary(mod2)$r.squared)  
cc2

## [1] 0.9472034

## Cálculo del coeficiente de determinación mod3  
cd3 <- summary(lm(mod3))$r.squared  
  
## Cálculo del coeficiente de correlación mod3  
cc3 <- sqrt(summary(mod3)$r.squared)

### 1. Comparación de coeficiente de relación y determinación de los modelos.

A continuación, se muestra una gráfica con los datos resumidos de los coeficiente de correlación de todos los modelos así como los coeficientes de determinación.

|  |  |  |
| --- | --- | --- |
| Ecuación del modelo | Coeficiente de correlación | Coeficiente de determinación |
| ventas = 7.03259 + 0.04754 \* tv | 0.7822244 | 0.6118751 |
| ventas = 2.938889 + 0.045765 \* tv + 0.188530 \* radio + -0.001037 \* periodico | 0.947212 | 0.8972106 |
| ventas = 2.92110 + 0.04575 \* tv + 0.18799 \* radio | 0.9472034 | 0.8971943 |
| ventas = 6.750220 + 0.019101 \* tv + 0.028860 \* radio + 0.001086(tv\*radio) | 0.9837635 | 0.9677905 |

### 2. Interpretación de los coeficientes del modelo 2.

Se tiene que para el modelo mod2 0.9472034, esto significa que la correlación lineal entre las ventas y tv, radio es positiva y moderada. Por otra parte, se tiene que 0.8971943, de esto se interpreta que el 94.72% de variación en las ventas se debe a la tv, radio y el restante 7.28% es por otros factores.

### 3. Comparación mediante valores AIC.

Se utilizará la función AICctab con el fin organizar los modelos de mejor a peor.

AICctab(mod0, mod1, mod2, mod3, base = T, delta = T, sort = T, weights = T, nobs = ncol(ventasdb))

## AICc dAICc df weight  
## mod3 490.3 0.0 5 1   
## mod1 722.4 232.1 5 <0.001  
## mod0 1068.1 577.8 3 <0.001  
## mod2 Inf Inf 4 <0.001

Como podemos ver tras ejecutar la función se pueden clasificar los modelos de mejor a peor de la siguiente forma:

1. mod3 con un puntaje de 490.3.
2. mod1 con un puntaje de 722.4
3. mod0 con un puntaje de 1068.1
4. mod2 con un puntaje de Inf.

### 4. Análisis de los residuales.

Se tomarán solo los modelos mod0, mod2 y mod3 debido a que el modelo 1 y 2 resultaron ser similares. Luego como primer paso se crean los Explain Model.

exp\_lm0 <- explain(mod0, data = ventasdb, label = "lm1", y = ventasdb$ventas)

## Preparation of a new explainer is initiated  
## -> model label : lm1   
## -> data : 200 rows 5 cols   
## -> target variable : 200 values   
## -> predict function : yhat.lm will be used ( [33m default [39m )  
## -> predicted values : No value for predict function target column. ( [33m default [39m )  
## -> model\_info : package stats , ver. 4.1.0 , task regression ( [33m default [39m )   
## -> predicted values : numerical, min = 7.065869 , mean = 14.0225 , max = 21.12245   
## -> residual function : difference between y and yhat ( [33m default [39m )  
## -> residuals : numerical, min = -8.385982 , mean = -5.146515e-15 , max = 7.212369   
## [32m A new explainer has been created! [39m

exp\_lm2 <- explain(mod2, data = ventasdb, label = "lm2", y = ventasdb$ventas)

## Preparation of a new explainer is initiated  
## -> model label : lm2   
## -> data : 200 rows 5 cols   
## -> target variable : 200 values   
## -> predict function : yhat.lm will be used ( [33m default [39m )  
## -> predicted values : No value for predict function target column. ( [33m default [39m )

## Warning: 'newdata' had 2 rows but variables found have 200 rows

## -> model\_info : package stats , ver. 4.1.0 , task regression ( [33m default [39m )   
## -> predicted values : numerical, min = 3.595686 , mean = 14.0225 , max = 24.78353   
## -> residual function : difference between y and yhat ( [33m default [39m )  
## -> residuals : numerical, min = -8.7977 , mean = -7.476123e-15 , max = 2.832837   
## [32m A new explainer has been created! [39m

exp\_lm3 <- explain(mod3, data = ventasdb, label = "lm3", y = ventasdb$ventas)

## Preparation of a new explainer is initiated  
## -> model label : lm3   
## -> data : 200 rows 5 cols   
## -> target variable : 200 values   
## -> predict function : yhat.lm will be used ( [33m default [39m )  
## -> predicted values : No value for predict function target column. ( [33m default [39m )

## Warning: 'newdata' had 2 rows but variables found have 200 rows

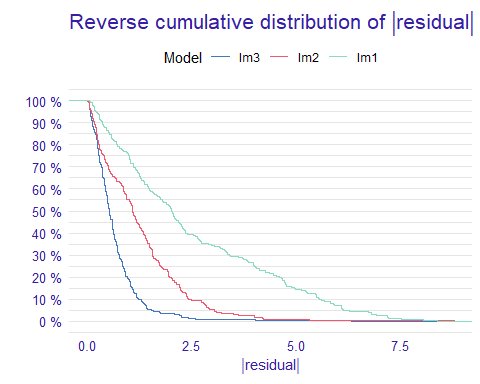
## -> model\_info : package stats , ver. 4.1.0 , task regression ( [33m default [39m )   
## -> predicted values : numerical, min = 6.994718 , mean = 14.0225 , max = 28.16216   
## -> residual function : difference between y and yhat ( [33m default [39m )  
## -> residuals : numerical, min = -6.336578 , mean = -5.43117e-15 , max = 1.52462   
## [32m A new explainer has been created! [39m

Como segundo paso se crean los Performance Model

lm0 <- model\_performance(exp\_lm0)  
lm2 <- model\_performance(exp\_lm2)  
lm3 <- model\_performance(exp\_lm3)

Y por último se grafican los modelos

plot(lm0,lm2, lm3)



Como podemos ver en el gráfico el mejor modelo es el mod3 ya que presenta la curva más baja de la gráfica, como segundo mejor modelo está mod2 y como peor modelo está mod1.

### 5. Análisis de varianza

Para este punto se realizará un análisis de los RSS de la función anova de R.

anova(mod1, mod2, mod3)

## Analysis of Variance Table  
##   
## Model 1: ventasdb$ventas ~ ventasdb$tv + ventasdb$radio + ventasdb$periodico  
## Model 2: ventasdb$ventas ~ ventasdb$tv + ventasdb$radio  
## Model 3: ventasdb$ventas ~ ventasdb$tv + ventasdb$radio + (ventasdb$tv \*   
## ventasdb$radio)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 196 556.83   
## 2 197 556.91 -1 -0.09 0.0312 0.8599   
## 3 196 174.48 1 382.43 134.6139 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Como conclusión, se puede observar que tras realizar la prueba de anova el resultado nos muestra nuevamente que el modelo mod3 es el mejor para utilizar en nuestras predicciones.

# VI Parte: Modelos de regresión no lineal

## Caso 7:

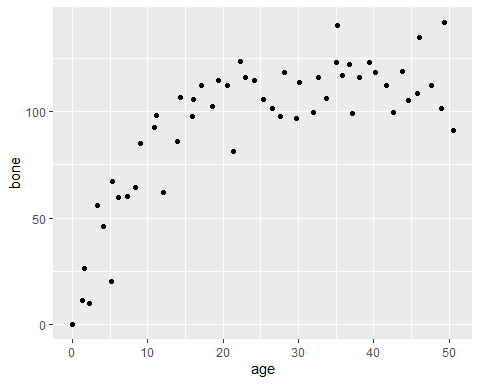
Para esta sección se van a generar y analizar dos modelos de RNLS usando los datos de bones de la base de datos jaws, la cual contiene información sobre la longitud de la mandíbula de los venados, según la edad.

### 1. Cargar los datos del archivo de texto.

jaws <- read.table("jaws.txt", header=T)

### 2. Gráfico de dispersión de los datos.

ggplot(jaws, aes(x = age, y = bone)) + geom\_point()



### 3. Generar modelo RNLS

Para esta parte se generará un modelo RNLS de la forma , para ello se hará uso de la función nlm() del paquete stats y se usarán como valores iniciales a = 120 y c = 0.064.

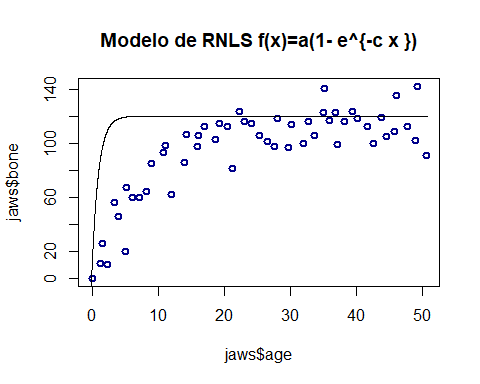
modelo1 <- nls(bone~a\*(1-exp(-c\*age)), data = jaws, start = list(a=120, c=0.064))  
modelo1

## Nonlinear regression model  
## model: bone ~ a \* (1 - exp(-c \* age))  
## data: jaws  
## a c   
## 115.5806 0.1188   
## residual sum-of-squares: 8929  
##   
## Number of iterations to convergence: 5   
## Achieved convergence tolerance: 1.369e-06

### Gráfico del modelo.

Ahora se procederá a realizar un gráfico del modelo anterior junto al de dispersión:

xo <- seq(from = min(jaws$age) - 0.2, to = max(jaws$age) + 0.2, by = 0.01)  
yo <- 120 \* (1 - exp(-0.064 - xo))  
  
# Graficamos dispersión y el modelo de ajuste en un mismo plot  
plot(jaws$age, jaws$bone, main = "Modelo de RNLS f(x)=a(1- e^{-c x })", col = "darkblue", lwd = 2)  
lines(xo, yo, type = "l")



### 5. Segundo modelo (selección del modelo).

Para este punto se hará un segundo modelo. Considerando que la gráfica tiene una asintota alrededor de 120 en el eje X y que parece tener también una menor que 0 en el eje Y (ya que ya hay un punto que toca 0) se decidirá por utilizar el modelo el modelo hiperbólico ya que este contiene asintotas en ambos ejes y permite más libertad en el posicionamiento de estas.

### 6. Linealización del segundo modelo.

En esta parte se hará una estimación de los parámetros y a través de un proceso de linealización a partir de un modelo RLS, para esto se hará uso una vez más de la función lm haciendo la sustitución de variables respectiva, que es la siguiente:

y

Teniendo en cuenta esto último se procede a realizar el RLS:

X2 <- 1/jaws$age  
Y2 <- 1/jaws$bone  
datoslm <- cbind(X2, Y2)  
datoslm.df <- as.data.frame(datoslm)  
  
## Se eliminan los infinitos del caso 0/0  
datoslm.df <- datoslm.df[!is.infinite(rowSums(datoslm.df)),]  
modelolm <- lm(Y2 ~ X2, data = datoslm.df)  
modelolm

##   
## Call:  
## lm(formula = Y2 ~ X2, data = datoslm.df)  
##   
## Coefficients:  
## (Intercept) X2   
## 0.00523 0.10339

Se tiene que:

y

### 7. Creación del segundo RNLS.

A partir de la estimación anterior de y se procederá a utilizar la función nls para generar un segundo modelo RNLS utilizando y como los valores iniciales

modelo2 <- nls(bone~(age)/(a\*age+b), data = jaws, start = list(a=0.00523, b=0.10339))  
modelo2

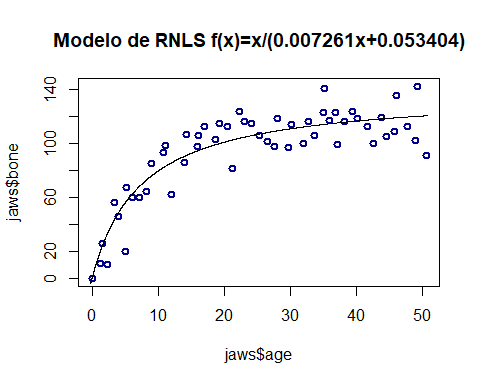
## Nonlinear regression model  
## model: bone ~ (age)/(a \* age + b)  
## data: jaws  
## a b   
## 0.007261 0.053403   
## residual sum-of-squares: 9854  
##   
## Number of iterations to convergence: 6   
## Achieved convergence tolerance: 1.898e-06

Finalmente se tiene que la ecuación de RNLS es:

### 8. Gráfico del segundo modelo RNLS.

En este punto se procede a realizar un gráfico del modelo generado en el paso anterior de la siguiente manera:

xo <- seq(from = min(jaws$age) - 0.2, to = max(jaws$age) + 0.2, by = 0.01)  
yo <- xo / (0.007261 \* xo + 0.053404)  
  
# Graficamos dispersión y el modelo de ajuste en un mismo plot  
plot(jaws$age, jaws$bone, main = "Modelo de RNLS f(x)=x/(0.007261x+0.053404)", col = "darkblue", lwd = 2)  
lines(xo, yo, type = "l")



### 9. Comparación de ambos modelos.

En este punto se compararán ambos modelos con el fin de cuantificar cual es mejor, para esto se hará uso de la función anova del paquete stats de R.

anova(modelo1, modelo2)

## Analysis of Variance Table  
##   
## Model 1: bone ~ a \* (1 - exp(-c \* age))  
## Model 2: bone ~ (age)/(a \* age + b)  
## Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)  
## 1 52 8929.1   
## 2 52 9854.4 0 0

En este caso bajo el criterio de análisis de varianzas el modelo 1 resulto ser el mejor ya que tiene un RSS menor. Sin embargo, esto parece no coincidir con las gráficas ya que parece ser que la gráfica del modelo 2 se ajusta mejor a los datos en comparación a la del modelo 1.