Proyecto de estadística: Etapa 2

Sebastián Rojas Vargas

Francisco Soto Quesada

Jairo Pacheco Campos

Jason Barrantes Rodríguez

17/4/2021

# 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)

# 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

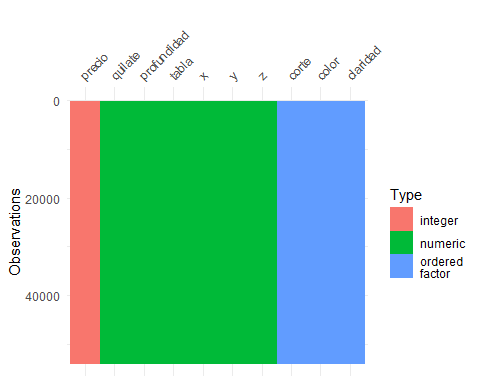
**Resumen del dataset:**

str(diamantes)

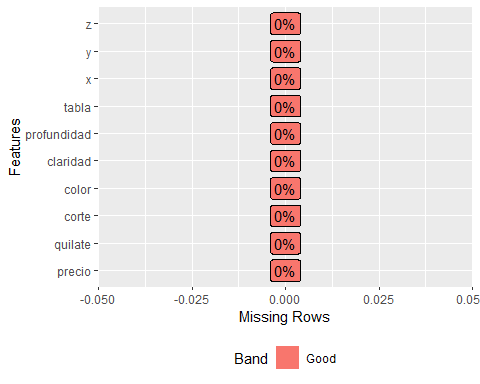
## tibble [53,940 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ precio : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...  
## $ quilate : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...  
## $ corte : Ord.factor w/ 5 levels "Regular"<"Bueno"<..: 5 4 2 4 2 3 3 3 1 3 ...  
## $ 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 población 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 como la media de precios del dataset diamantes 3932.7997219 se encuentra en el intervalo ]3899.1327134, 3966.4667305[ obtenido de tres librerías distintas, todo está correcto.

## 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 estandar |
| 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 como la media de precios del dataset de diamantes 3932.7997219 se encuentra en el intervalo ]2250.6105535, 7008.5894465[, todo está correcto, aunque comparándolo con el IC de la distribución Z, 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 ambas siguen una distribución normal y que 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 parametro:** primera muestra
* **conf.level:** nivel de confidencia

**Conclusión**

Dando Como resultado el IC de varianza: ]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 parametro:** primera muestra
* **Segundo parametro:** 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 mismo 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 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 primer 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 primer 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 |  |
| Estadístico de prueba |  |
| Región de aceptación | ]3904.545, +[ |
| Región de rechazo | ]-, 3904.545[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

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 estandar |
| 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 siguient 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 | 19 |
| Estadístico de prueba |  |
| Región de aceptación | ]2664.219, +[ |
| Región de rechazo | ]-, 2664.219[ |
| Nivel de confianza | 0.95 |

**Conclusiones**

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

## Prueba de hipótesis para una diferencia de proporciones

## 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 |  |
| Estadístico de prueba | y |
| 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 la 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 | y |
| 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 de el dataset diamantes

quilates <- diamantes$quilate

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

**Hipótesis asumidas**

Antes de realizar el calculo 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 estandar 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 | 53939 |
| Grados de libertad | 53939 |
| Estadístico de prueba | $\s^{2}$ |
| 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[