Proyecto de estadística: Etapa 1

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)

# 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

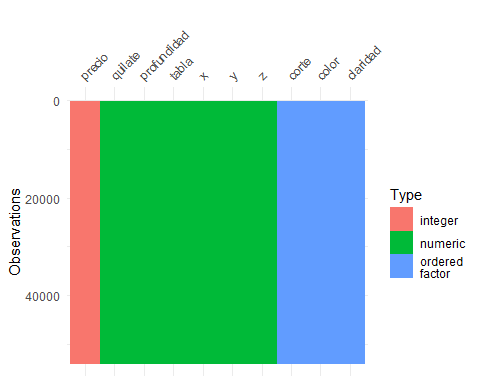
**Resumen del dataset:**

str(diamantes)

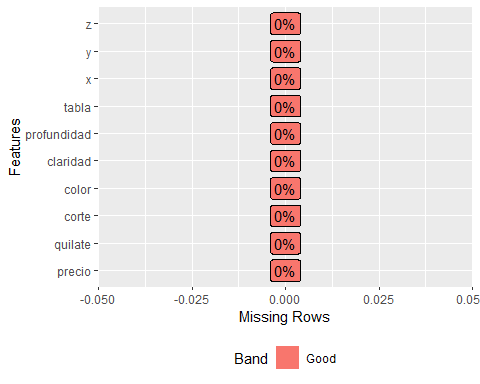
## tibble [53,940 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ precio : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...  
## $ quilate : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...  
## $ corte : Ord.factor w/ 5 levels "Regular"<"Bueno"<..: 5 4 2 4 2 3 3 3 1 3 ...  
## $ 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 continuacion se presentan los promedios usando distribución normal estandar de una población de datos sacados del dataset diamantes, utilizando la siguiente muestra de la variable precio del dataset de diamantes:

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Desviación estantar |
| A | 100 | 3486.47 | 3605.9501261 |

Dado que la muestra es mayor a 30 se puede asumir que la población siguen una distribución normal y que se puede aproximar σ₁ mediante s₁. Conociento los datos se puede aproximar el IC utilizando estimación por intervalo, para esto se implementa la siguente función manual:

ICZ <- function(x, sigma, alpha){  
 c(mean(x) + qnorm(alpha/2)\*sigma/sqrt(length(x)),   
 mean(x) - qnorm(alpha/2)\*sigma/sqrt(length(x)))}

En donde:

* **x:** la muestra
* **sigma:** desviación estandar de la muestra
* **alpha:** nivel de confidencia

Y ahora se calcula el intervalo para un IC del 95%:

]2779.7167623, 4193.2232377[

Del cual se puede concluir que como la media de precios de todo el dataset de diamantes 3932.7997219 se encuentra en el intervalo, todo está correcto.

## IC de un promedio usando distribución t

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

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Desviación estantar |
| 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.



Como se puede observar, la muestra tiene forma de campana, por lo que se puede asumir que sigue una distribución normal. Conociendo esto se puede aproximar el IC mediante la funcion t.test, la cual tiene la siguiente estructura:

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

Donde:

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

Y ahora se calcula el intervalo para un IC del 95%:

]2250.6105535, 7008.5894465[

Del cual se puede concluir que como la media de precios de todo el dataset de diamantes 3932.7997219 se encuentra en el intervalo, todo está correcto.

## IC de una proporción

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

|  |  |  |
| --- | --- | --- |
| 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 cual 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:

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

Dado que las muestras son mayores a 30 se puede asumir que la población sigue una distribución normal. Conociendo los datos se puede calcular la proporción, para esto se implementa la función prop.test para una población de la siguiente manera

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

Y se calcula la proporción para un 95% de confianza:

]0.1221170, 0.1290541[

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 proporciones

A continuación se presenta la diferencia de proporciones para dos poblaciones de datos tomados del dataset diamantes, para ello se han utilizado dos muestra de la variable “corte” donde en la primera se inculyen 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:

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

|  |  |  |
| --- | --- | --- |
| Gama | Corte | Tamaño de muestra |
| Alta | Regular y Bueno | 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 cual 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 | 825 |

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

Dado que las muestras son mayores a 30 se puede asumir que la población sigue una distribución normal. Conociendo los datos se puede calcular 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

Y se calcula la proporción para un 95% de confianza:

]0.6460786, 0.6691132[

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

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

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 hay una mayor cantidad de diamantes con colores de baja calidad que 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 presentan las varianzas de dos poblaciones de datos sacados del dataset de diamantes.

Primero se crean dos poblaciones distintas y random a partir de el dataset padre las cuales son de tamaño 100 como lo muestra la siguiente tabla.

|  |  |
| --- | --- |
| Muestra | Tamaño |
| A | 100 |
| B | 100 |

Una vez generadas las dos poblaciones, utilizaremos la variable de quilates para obtener el IC de varianza de una población, para esto utilizaremos la librería EnvStats y específicamente su método vartTest que nos dará el intervalo que estamos buscando.

En cuanto a las muestras o las poblaciones es importante mencionar que se están usando las mismas que para la sección anterior

Debemos tener en cuenta que:

* El nivel de confianza utilizado será de 95%
* La población 1 y la población 2 siguen una distribución normal

El código utilizado para este cálculo fue:

##Población 1  
P1 <- varTest(poblacion1$quilate, conf.level = 0.95)$conf.int  
  
##Poblacion 2  
P2 <- varTest(poblacion2$quilate, conf.level = 0.95)$conf.int

Donde:

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

Dando Como resultado para la primera población: ]0.214497, 0.375487[

Y dando como resultado para la segunda población: ]0.1413782, 0.2474892[

VP <- var(diamantes$quilate)

Tras realizar estos cambios sacamos la varianza de todo el dataset para comprobar que se encuentra dentro del IC de las poblaciones dando como resultado 0.2246867 y como podemos ver si se encuentra dentro de los dos intervalos lo que muestra que el cálculo fue hecho de manera correcta.

## IC de un cociente de Varianza

Gracias a la librería stests podemos realizar el cociente de varianza de dos poblaciones, para esto utilizaremos a la población 1 y población 2 para realizar este cálculo además de la variable de quilates para 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.

El código utilizado para este cálculo fue:

R <- stests :: var.test(poblacion1$quilate, poblacion2$quilate,conf.level = 0.95)$conf.int

Donde:

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

Dando como resultado: ]1.0208255, 2.2548921[

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, sin embargo como no es 100% seguro y el intervalo está muy cerca de 1, puede que en algún punto sus varianzas sean iguales, también podemos decir que en este caso la varianza de la población 1 es mayor a la de la población 2 ya que el intervalo es positivo.