Proyecto de estadístisca: Etapa 1

Sebastián Rojas Vargas

Francisco Soto Quesada

Jairo Pacheco Campos

Jason Barrantes Rodríguez

17/4/2021

# I Parte: explicación de los datos

A continuación se presenta una tabla con los principales aspectos de el dataset utilizado

**Dataset:** Diamantes

**Descripción general:** el dataset resulta del paquete de datos que el profesor puso a disposición para la elección de la base de datos. Se eligió el dataset diamantes porque su estructura nos parece muy apropiada para el proyecto ya que cuenta con variables tanto cuantitativas como cualitativas, además que cuenta con gran cantidad de datos para la toma de muestras.

**Filas:** 53 940

**Columnas:** 10

## 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 milimetros | 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

## IC de una diferencia de proporciones

## IC de una diferencia de promedios usando distribución z

Para este ejemplo se utilizaran las siguientes muestras de la variable profundidad del dataset de diamantes:

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Varianza |
| A | 100 | 61.691 | 1.3369889 |
| B | 100 | 61.8 | 1.9587879 |

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.

Ya sabiendo estos datos se puede aproximar el IC mediante el caso 1 de diferencia de promedios, para esto se implementa una función manual:

z.test2 <- function(mu1, mu2, n1, n2, var1, var2, alpha) {  
 c((mu1-mu2)+qnorm(alpha/2)\*sqrt((var1/n1)+(var2/n2)),  
 (mu1-mu2)-qnorm(alpha/2)\*sqrt((var1/n1)+(var2/n2)))  
}

En donde:

* **mu1:** media de la primera muestra
* **mu2:** media de la segunda muestra
* **n1:** tamaño de la primera muestra
* **n2:** tamaño de la segunda muestra
* **var1:** varianza de la primera muestra
* **var2:** varianza de la segunda muestra
* **alpha:** valor de alpha

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

]-0.4648172, 0.2468172[

Del cual se podría concluir que existe la posibilidad de que:

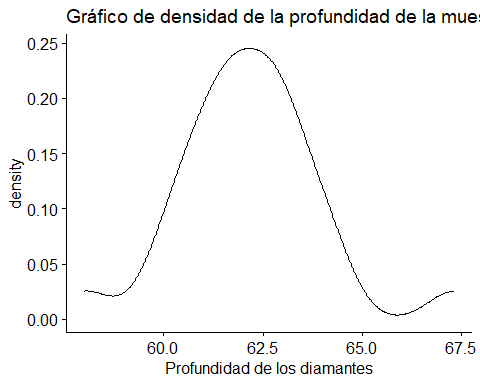
* El promedio de la muestra A es mayor que el de la muestra B
* Las muestras tengan el mismo promedio
* El promedio de la muestra B es mayor que el de la muestra A

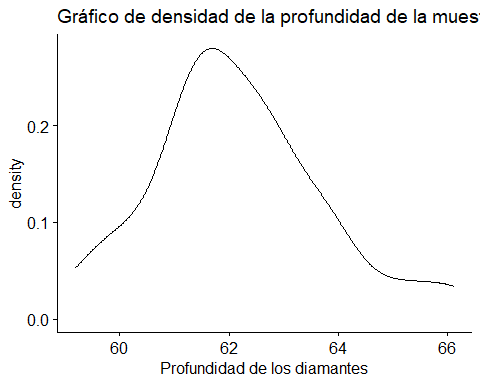
## IC de una diferencia de promedios usando distribución t

Para este ejemplo se utilizaran las siguientes muestras de la variable profundidad del dataset de diamantes:

|  |  |  |  |
| --- | --- | --- | --- |
| Muestra | Tamaño | Promedio | Varianza |
| A | 25 | 62.156 | 3.2134 |
| B | 25 | 62.168 | 2.5256 |

En este caso dado que los tamaños de la 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





Como se puede observar ambas muestras tiene forma de campana, por lo que se puede asumir que siguen una distribución normal. Sin embargo, como las muestras son pequeñas no se pueden utilizar s₁ y s₂ para realizar una aproximación σ₁ y σ₂, por lo cual el caso 3 sería el óptimo en esta situación.

Conociendo esto se puede aproximar el IC mediante la funcion t.test, la cual tiene la siguiente estructura:

t.test(x=muestra1profundidad, conf.level = 0.95, var.equal = TRUE)

Donde:

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

Y ahora se calcula la diferencia de promedios para un IC del 95%

]-0.9753438, 0.9513438[

Del cual se podría concluir que existe la posibilidad de que:

* El promedio de la muestra A es mayor que el de la muestra B
* Las muestras tengan el mismo promedio
* El promedio de la muestra B es mayor que el de la muestra A

## 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 especificamente su metodo vartTest que nos dará el intervalo que estamos buscando.

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

Debemos tener en cuenta que:

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

El codigo utilizado para este calculo 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[

## 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 calculo ademas de la variable de quilates para realizar el cálculo.

Debemos tener en cuenta lo siguiente:

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

El codigo utilizado para este calculo 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 deberian ser iguales ya que el numero 1 no pertenece al intervalo, sin embargo como no es 100% seguro y el intervalo esta muy cerca de 1, puede que en algun punto sus varianzas sean iguales, tambien 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.