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

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 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:

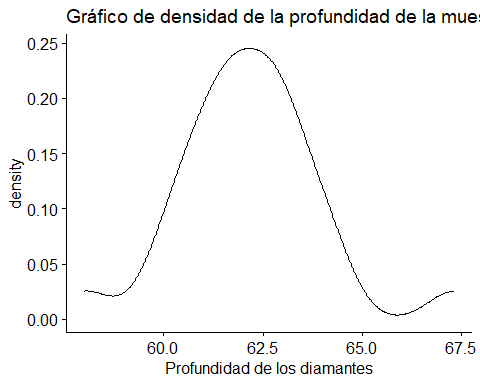
* El promedio de la muestra A es mayor que el de la muestra B
* Las muestras tengan el mismo promedio
* El promedio de la muestra B es mayor que el de la muestra A

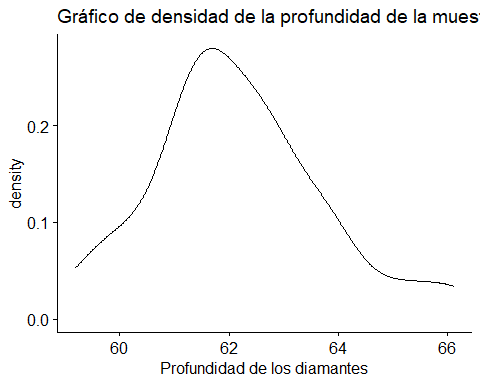
## IC de una diferencia de promedios usando distribución t

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:

IC <- t.test(x=muestra1$profundidad, y=muestra2$profundidad, conf.level = 0.95, var.equal = TRUE)$conf.int

Donde:

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