Práctica dirigida 5

Revisemos lo avanzado sobre estadística inferencial

En esta sesión repasaremos lo revisado hasta el momento sobre estadística inferencial: Intervalos de confianza, prueba T y revisamos comparación de proporciones.

La base a usar en esta sesión ("data-paises.xlsx") proviene de la unión de tres bases de datos distintas, estas son "Human Development Index" elaborada por UNDP (https://hdr.undp.org/data-center/documentation-and-downloads),); "Index of Economic Freedom" realizada por Heritage Foundation (https://indexedinvt.azurewebsites.net/index/explore) y "Fragile State Index" elaborada por The Fund For Peace (https://fragilestateindex.org/global-data/).) Asimismo, los datos con los que trabajaremos corresponden a información del 2021 y la unidad de análisis son países.

Sobre las bases de estudio:

- El Human Development Index mide el índice de desarrollo humano a través de 3 aspectos: esperanza de vida, educación e ingresos per cápita. Su escala es de 0 a 1.
- El Index of Economic Freedom mide el grado de libertad económica mediante 12 indicadores agrupados en 4 categorías: Estado de Derecho, tamaño de Gobierno, eficiencia regulatoria y apertura de mercados.
- El Fragile State Index mide la fragilidad de un Estado a través de 12 indicadores agrupados en 4 categorías: cohesión, economía, política y social. Su escala es de 0 a 120.

```
#Llamemos al paquete
library(rio)
data=import("data-paises.xlsx")
#Llamemos a nuestra base de datos
```

Prueba T

Recuerda que hay condiciones para poder realizar la prueba T:

- Independencia: Las muestras deben ser independientes. El muestreo debe ser aleatorio.
- Igualdad de varianza: La varianza de ambas poblaciones comparadas debe ser igual. *(Prueba Levene)*
- La variable numérica se distribuye de manera normal.

Pasos para realizar la Prueba T

1. Establecer hipótesis
2. Calcular el estadístico (parámetro estimado) que se va a emplear
3. Determinar el nivel de significancia (α (alpha)
4. Calcular el p-value y realizar la prueba t.test
5. Interpretar

Apliquemos lo revisado…

Primero, exploraremos un poco las variables de interés:

V27: Indice de desarrollo humano V1: Indice de libertad económica

Al ver la estructura de nuesta base de datos podemos observar que ambas variables de interés (V27 y V1) son categóricas. Por ello, haremos un pequeño cambio.

Segundo volveremos la variable V27 una variable numérica y crearemos 2 grupos.

Siendo los niveles: "Bajo / Medio" = si es menor o igual a 0.7350, y, "Alto/ Muy alto" = si es mayor a 0.7350

¿Cómo lo haremos? Con case when!

```
library(tidyverse)
data$V27 = as.numeric(data$V27)

data = data %>%
  mutate(
    grupo_IDH= case_when(V27<=0.7350 ~ "1. Bajo/Medio", V27>0.7350 ~ "2. Alto/Muy alto"))
```

Segundo,vamos a transformar la variable "V1" para que sea númerica

```
data$V1 = as.numeric(data$V1)
```

Ahora, analizaremos la varianza en los grupos, para ello usaremos la prueba Levene:

H0: La varianza del Indice de Desarrollo Humano es igual a la varianza del Indice de Libertad Económica. H1: La varianza del Indice de Desarrollo Humano NO es igual a la varianza del Indice de Libertad Económica.

```
library(DescTools)
LeveneTest(data$V1, data$V27)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 145  0.3759 0.9997
##       22
```

Como el p valor es mayor a 0.05 podemos afirmar que las varianzas son iguales, por lo que podemos realizar la Prueba T.

Realizamos la Prueba T

Primer paso: Establecer la hipótesis.

La hipotesis de la prueba T queda establecida de la siguiente forma:

- H0: No hay diferencia de promedio en los niveles de libertad económica entre aquellos paises que tienen un indice de desarrollo humano "Bajo / Medio" y los que tienen un indice de desarrollo humano "Alto / Muy alto" **(no diferencia de medias)**
- H1: Si hay diferencia de promedio en los niveles de libertad económica entre aquellos paises que tienen un indice de desarrollo humano "Bajo / Medio" y los que tienen un indice de desarrollo humano "Alto / Muy alto" **(si diferencia de medias)**

Ambas hipótesis son acerca de los parámetros de la población.

Segundo paso: Calcular el estadístico a emplear

Para verificar la diferencia de medias se calcula el estadístico T, y uno de los primeros pasos es calcular las diferencias entre las medias muestrales, ya que es lo quiero extrapolar y por tanto saber si existe o no una diferencia significativa entre las medias poblacionales de ambos grupos:

```
library(lsr)
tabla=data%>%
  group_by(grupo_IDH) %>%
summarise(Desviacion = sd(V1, na.rm=T),
          Media = mean(V1, na.rm=T),
          min = ciMean(V1,conf = 0.95, na.rm=T)[1],
          max = ciMean(V1,conf = 0.95, na.rm=T)[2],
          n=length(V1))

tabla
```

```
## # A tibble: 2 × 6
##   grupo_IDH       Desviacion Media   min   max     n
##   <chr>                <dbl> <dbl> <dbl> <dbl> <int>
## 1 1. Bajo/Medio         7.03  55.4  54.1  57.2    82
## 2 2. Alto/Muy alto      9.32  68.0  66.0  70.0    86
```

Tercer paso: Determinar el nivel de significancia

De manera convencional establecemos la siguiente regla para nuestra prueba T:

- p-value<=0.05 Rechazo la H0 y acepto H1
- p-value>0.05 No rechazo la H0

Cuarto paso: Calcular el p-value y realizar la prueba t.test

Recuerda que el p-value mide la probabilidad de observar en una muestra una diferencia de medias como la observada, si la diferencia de medias poblacional fuera cero.

```
t.test(V1 ~ grupo_IDH, data = data,
  alternative = "two.sided",
  conf.level  = 0.95 #nivel de confianza (95%)
)
```

```
##
##  Welch Two Sample t-test
##
## data:  V1 by grupo_IDH
## t = -9.7594, df = 157.77, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1. Bajo/Medio and group 2. Alto/Muy alto is not equal to 0
## 95 percent confidence interval:
##  -14.895211  -9.889966
## sample estimates:
##      mean in group 1. Bajo/Medio mean in group 2. Alto/Muy alto
##                         55.69610                         67.99419
```

Quinto paso: Interpretar

¿Cómo interpreto?

Recordando nuestras hipotesis:

- H0: No hay diferencia de promedio en los niveles de libertad económica entre aquellos paises que tienen un indice de desarrollo humano "Bajo / Medio" y los que tienen un indice de desarrollo humano "Alto / Muy alto"
- H1: Si hay diferencia de promedio en los niveles de libertad económica entre aquellos paises que tienen un indice de desarrollo humano "Bajo / Medio" y los que tienen un indice de desarrollo humano "Alto / Muy alto"

Asimismo, en el paso 4, determinamos el nivel de significancia de la siguiente manera:

- Si el p-value del t test es <=0.05 Rechazo la H0 y se afirma H1.
- Si el p-value del t test es >0.05 No rechazo la H0

Entonces, vemos que el p-value es menor a 0.05, entonces rechazo la H0, por tanto, existe una diferencia estadisticamente significativa entre las medias del Indice de libertad económica entre los paises que tienen un Indice de Desarrollo Humano "Bajo / Medio" y los paises que tiene un Indice de Desarrollo Humano "Alto / Muy alto".

Paso FINAL: Graficar

Otro método para realizar la comparación entre grupos es realizar un gráfico de medias con intervalos de confianza de cada grupo.

Para calcular la diferencia de medias

```
library(Rmisc)
ic_grupo = group.CI(V27~grupo_IDH,data)
ic_grupo
```

```
##                 grupo_IDH V27.upper  V27.mean V27.lower
## 1   1. Bajo/Medio 0.4123113 0.5916618 0.5701106
## 2 2. Alto/Muy alto 0.8622095 0.8473837 0.8325579
```

Barras de Error

```
library(ggplot2)

ggplot(ic_grupo, aes(x= grupo_IDH, y =V27.mean)) +
  geom_errorbar(aes(ymin=V27.lower, ymax=V27.upper), width = 0.2)+
  geom_text(aes(label=paste(round(V27.mean,3)), size=3))+
  xlab("Indice de Desarrollo Humano") + ylab("Indice de Libertad Económica")
```


Interpretación: Tal como se observa ambos intervalos de confianza no se traslapan, por lo que se puede concluir gráficamente que existe una diferencia estadisticamente significativa entre los grupos. El grupo que tiene un IDH "Alto/Muy alto" tiene mayor libertad económica que el grupo que tiene un IDH "Bajo/Medio" con un 95% de confianza.

Comparación de proporciones

Para este ejercicio trabajaremos con dos variables:

- *V20: Legitimidad del Estado*
- *V21: Servicios públicos*

Revisemos a nuestras variables

Variable V20:

```
class(data$V20) #Revisamos como está catalogada nuestra variable
```

```
## [1] "numeric"
```

```
summary(data$V20)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   3.600   4.250   5.593   7.725  10.000
```

Recordemos que para comparar proporciones necesitamos que nuestra variable sea categórica. La recodificaremos para que tengamos dos grupos: Baja/Media (de 7.73 a menos) y Alta (más de 7.73).

```
library(tidyverse)#Llamemos al paquete
data = data %>%
  mutate(V20_2 = case_when(V20 <= 7.73 ~ "Baja/Media",
                           TRUE ~ "Alta"))
```

Realizamos el mismo ejercicio con nuestra variable V21:

```
class(data$V21) #Revisamos como está catalogada nuestra variable
```

```
## [1] "numeric"
```

```
summary(data$V21)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   3.700   5.300   5.585   7.600  10.000
```

La recodificaremos para que tengamos dos grupos: Baja/Media (de 7.6 a menos) y Alta (más de 7.6).

```
data = data %>%
  mutate(V21_2 = case_when(V21 <= 7.6 ~ "Baja/Media",
                           TRUE ~ "Alta"))
```

Necesitamos calcular la diferencia entre aquellos paises que cuenta con un indicador de servicios públicos alto y alta legitimidad, y aquellos que tienen una alta legitimidad y un indicador de servicios público bajo o medio.

```
#Realizamos una tabla de frecuencias
table(data$V20_2,data$V21_2)
```

```
##
##              Alta Baja/Media
##   Alta         23         19
##   Baja/Media   18        108
```

Identificamos lo que nos interesa: La frecuencia de los que tienen un indicador alta en legitimidad y servicios públicos es 23; mientras que, los que tienen un indicador alto de legitimidad y bajo o medio de servicios públicos es de 19.

```
#Hallamos la proporción
prop.test(x=c(23,19),n=c(23+18,19+108))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(23, 19) out of c(23 + 18, 19 + 108)
## X-squared = 25.822, df = 1, p-value = 3.744e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2231536 0.5915850
## sample estimates:
##    prop 1    prop 2
## 0.5609756 0.1496063
```

Interpretación: la diferencia entre aquellos paises que cuenta con un indicador de servicios públicos alto y alta legitimidad, y aquellos que tienen una alta legitimidad y un indicador de servicios público bajo o medio se encuentra entre 23.1% y 59.2%, a un 95% de confianza.

Ejercicios

- Analizaremos la variable V17- Economia.
  - Halla el intervalo de confianza para la media.
  - Halla el intervalo de confianza para la media segun **gasto de gobierno (V6)**. Toma en consideración que la variable gasto de gobierno está como numérica, necesitamos que esté como categórica. Para ello usamos case_when y recodificamos segun gasto bajo, medio y alto.
- Analizaremos la variable *V30- Promedio de años de escolaridad*
  - Halla el intervalo de confianza para la proporción de paises que tienen un promedio de años de escolaridad alto. Para ello recodifica de la siguiente manera:
    - De 12 años a menos: "Doce años a menos"
    - Más de 12: "Más de 12 años"
- *¿Existe diferencia de medias de gasto del gobierno (V6) según tiempo de escolaridad (V30_2; creada en ejercicio anterior)?* Recuerda realizar la prueba Levene.