UNIDAD 4: Estimación de Parámetros y Pruebas de Hipótesis

Supongamos que trabaja para EIP SAC y le ha llegado a su oficina una nueva encuesta nacional sobre la intención de voto con miras a las elecciones de abril del 2016. Usted se encuentra sorprendido porque, en una semana, Keiko Fujimori bajó de 30% a 5% en las encuestas. Su experiencia le dice que una caída tan radical tal vez responda a un error en el recojo de información. Los medios de comunicación lo llaman insistentemente para preguntarle si es verídico el dato, pues tienen la intención de publicar este noticia desarrollando el argumento del reciente desplome fujimorista a nivel nacional. ¿Cómo comprobar que los datos son verídicos? O en su defecto, ¿dónde podría estar el error?

Antes de emitir cualquier conclusión a nivel poblacional es necesario distinguir las diferencias entre estadísticos y parámetros. Los primeros se refieren a estadísticas muestrales, y los segundos a las poblacionales. En esta unidad repasaremos algunos tópicos de estadística inferencial, la estadística que nos permite generalizar hallazgos de lo que pudimos recabar a través de cuestionarios en nuestro trabajo de campo.

Para esta sesión utilizaremos la base de datos de Género del Instituto de Opinión Pública.

A tener en cuenta: Tabla de Distribución Normal Estándar - Puntuaciones Típicas Z

Cuando no se conoce la varianza poblacional, se trabaja con la distribución normal estándar (puntuaciones Z). Los intervalos de confianza más usados para la investigación son 90%, 95 % y 99 % cuyas puntuaciones Z son 1.645, 0.025 y 0.005 respectivamente.

1-α	α/2	$Z_{\alpha/2}$	Intervalo de confianza
0,90	0,05	1,645	$(\bar{X} - 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 1{,}645 \cdot \frac{\sigma}{\sqrt{n}})$
0,95	0,025	1,96	$(\bar{X} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}})$
0,99	0,005	2,575	$(\bar{X} - 2{,}572 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{X} + 2{,}575 \cdot \frac{\sigma}{\sqrt{n}})$

Intervalo de Confianza y Error para una Muestra

Para el cálculo de intervalos de confianza de una media es necesario realizar la siguiente fórmula:

$$IC_{1-\alpha} = \overline{x} \pm Z\sigma_{\overline{x}}$$

$1-\alpha$ Es el nivel de confianza que vamos a calcular

Z	El valor de distribución normal estándar (puntuaciones Z) que nos brindará el nivel de confianza
$\sigma_{\bar{x}}$	Error estándar de la media y se calcula con la siguiente fórmula:

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

A continuación realizaremos el cálculo para el error muestral y el intervalo de confianza al 95% para la variable **P25**. Para esto, primero se debe realizar una exploración de la variable inicial.

```
install.packages ("foreign")
library (foreign)
data2<-read.spss("IOPGENERO.sav", to.data.frame=TRUE, use.value.labels = TRUE)
attach(data2)
install.packages("Hmisc")
library(Hmisc)
```

P25. ¿Cuál es la edad de su actual cónyuge o pareja?

```
describe(P25)
```

```
P25
      n missing  unique    Info    Mean
    688     515      62       1   42.85
    .05     .10     .25     .50     .75     .90     .95
     24      26      32      40      52      63      69
lowest : 17 19 20 21 22
highest: 75 77 78 80 83
```

Como se puede observar en los resultados, la variable tiene un total de 688 casos válidos y 515 valores perdidos. Esta disgregación se realiza por defecto, pero para estar más seguros de que no están ingresando datos perdidos al análisis realizaremos la siguiente función:

```
p25r<- P25                    #Creamos un nuevo objeto p25r
```

```
p25r[P25>83] <- NA                # Los datos que sean mayores a 83, valor máximo, serán considerados como NA.

describe(p25r)

p25r
     n missing  unique    Info    Mean  05    .10    .25    .50    .75    .90    .95
   688     515       6       1   42.85  24    26     32     40     52     63     69

   .

lowest : 17 19 20 21 22

highest: 75 77 78 80 83
```

Como podemos observar, en estos resultados no se han adicionado casos NA, pero este paso es necesario ya que nos evita trabajar con posibles valores NA. Ahora iniciaremos el cálculo del error estándar y el intervalo de confianza. En primer lugar, se pide un estadístico de centralidad (media), uno de dispersión (desviación estándar) y el tamaño de la muestra. Luego, se calcula el error estándar, según la fórmula y el valor Z. Por último, se establecen los límites inferior y superior del intervalo.

```
media <-mean(na.omit(p25r))

#Pedimos la media

desv  <-sd(na.omit(p25r))

#La desviación estándar

N<-length(na.omit(p25r))

#El  tamaño válido de  la muestra

error.est<-desv/sqrt(N)

# Calculamos  el error estándar. Recuerden la Fórmula:
```

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

```
[1] 0.532841

error   <- 1.96*error.est

#Establecemos la puntuación Z (1.96) para indicar un nivel de confianza de 95%

lim.inf<- media-error

# Límite inferior del intervalo

[1] 41.80447

lim.sup <- media+error

# Límite superior del intervalo
```

[1] 43.89321

Dato extra# *Para guardar los datos generados en un solo objeto señale "resultado1 <- data.frame (media, desv, N, error.est, error, lim.inf, lim.sup)"*

Podemos observar que el intervalo de confianza para la muestra con un intervalo de confianza de 95 % está entre {41.80447 – 43.89321}, lo que nos quiere decir que la edad de su actual cónyuge oscila entre estos límites en el 95% de la población.

Ahora calcularemos el error estándar para una media y el intervalo de confianza para la variable **P26** con un intervalo de confianza de 90 % y 99%.

Intervalo de Confianza y Error para un parámetro

Para el cálculo de intervalos de confianza para un parámetro será necesario realizar la siguiente fórmula:

$$IC_{1-\alpha} = p \pm Z\sigma_p$$

$1-\alpha$	Es el nivel de confianza que vamos a calcular; alfa es el nivel de significancia.
Z	El valor de distribución normal estándar (puntuaciones Z) que nos brindará el nivel de confianza
σ_p	Error estándar de la proporción y se calcula con la siguiente fórmula: $$\sigma_p = \sqrt{\frac{pq}{n}}$$

A continuación realizaremos el cálculo para el error de una proporción y el intervalo de confianza al 95% para la variable **P34.** Primero, se debe explorar la variable inicial.

```
install.packages ("foreign")

library ("foreign")

data2<-read.spss("IOP-GENERO.sav", to.data.frame=TRUE, use.value.labels = TRUE)
```

```
attach (data2)

P34

install.packages("Hmisc")   ##paquete necesario para utilizar el comando describe ()

library("Hmisc")

describe(P34)
```

P34

```
     n missing  unique

   690     513       6
```

Principalmente yo (114, 17%)

Principalmente mi cónyuge /pareja (73, 11%)

Algunas veces yo, y otras mi cÃ³nyuge/pareja (141, 20%)

Lo decidimos (o decidÃamos) juntos (356, 52%)

Alguna otra persona (4, 1%)

No contesta (2, 0%)

Como se puede observar en los resultados, la variable tiene un total de 690 casos válidos y 513 valores perdidos. A continuación, omitiremos los valores perdidos recodificando la variable:

Para ordenar los datos de frecuencia, porcentaje y porcentaje válido utilice el paquete "descr" y el comando "freq"

```
library(descr)
freq(P34, plot=FALSE)
```

A continuación, omitiremos los valores perdidos recodificando la variable:

```
p34r<-na.omit(P34)
```

Luego seleccionaremos la categoría de la variable con la que trabajaremos y crearemos un objeto que será llamado categoría y que tomará la categoría "Principalmente mi cónyuge/pareja" y su probabilidad será comparada con las otras cinco categorías de la variable ("Principalmente yo", "Algunas veces yo y otras mi cónyuge/pareja", "Lo decidimos (o decidíamos) juntos", "Alguna otra persona", "No contesta")

```
categoria<- ifelse (p34r=="Principalmente mi cónyuge /pareja", 1, 0)
```

Los valores 1 y 0 significan, la cantidad de personas que seleccionaron la categoría "Principalmente mi cónyuge/pareja" y las que no; por lo tanto, marcaron las otras categorías, respectivamente.

Después de observar los datos mostrados en las tablas, podemos observar que la categoría que analizaremos tiene un total de 73 personas que optaron por esta respuesta, mientras que el número total de respuestas válidas de la muestra es 690. Estos resultados se pueden observar mejor en la tabla siguiente:

```
prop.table(table(categoria))
```

```
categoría
      0       1
0.8942029 0.1057971
```

Realizado todo este procedimiento, se iniciará el cálculo del error estándar y el intervalo de confianza

```
p<- mean(cat)
p
#Esta es la proporción de personas que respondieron esta opción
media <-mean(na.omit(p34r))
n <- length(categoria)
# Tamaño de la muestra
error.est.p<- sqrt((p*(1-p))/n)
# Calculamos  el error estándar. Recuerden la Fórmula:   σ_p = √(pq/n)
# pq = la posibilidad de éxito por la posibilidad de fracaso
# q = 1 – p
# p = f1 / n
```

```
[1] 0.01170928
```

```
error.p<-1.96*error.est.p
#Establecemos la puntuación Z (1.96) para indicar un nivel de confianza de 95%
lim.inf.p<-p-error.p
# Límite inferior del intervalo
```

```
[1] 0.0828469
```

```
lim.sup.p<-p+error.p
```

```
# Límite superior del intervalo
```

```
[1] 0.1287473
```

Podemos concluir con la siguiente interpretación: estamos un 95% seguros que para entre el 0,08% al 0,12 % de la población, las decisiones para el fin de semana son tomadas principalmente por el cónyuge/pareja.

Prueba de hipótesis de una y dos muestras para medias y proporciones

Para la realización de estas pruebas volveremos a utilizar la variable **RESEMAVAL, índice que mide Emacipative Values** de la base de datos **WordVS2015** y con el análisis del caso peruano.

Si queremos realizar la **prueba de hipótesis de una muestra única** es necesario que nuestra variable numérica pase las pruebas de normalidad correspondientes, tal como se observó en la Unidad 3.

```
install.packages("foreign") #necesario cargar el paquete foreign para utilizar archivos SPSS

library(foreign) #llamemos al paquete

data<-read.spss("WordVS2015.sav", to.data.frame=TRUE, use.value.labels = TRUE)

#importemos la data

data1<-data[data$V2=="Peru",]

# creemos un nuevo objeto (data1) para filtrar sólo el caso peruano.

install.packages("nortest") #paquete necesario para realizar pruebas de normalidad

library(nortest) # llamemos al paquete

attach(data1)

lillie.test(RESEMAVAL)$p.value

#pedimos el comando que nos dará el resultado de la prueba de Kolgomorov Smirnov
```

```
[1] 0.1502064
```

El valor es mayor a 0,05 entonces aceptamos que la variable es normal.

A continuación, realizaremos una “Prueba T para una muestra única”, para ello es necesario establecer una media que consideramos que puede ser significativamente diferente a la media del nuestro índice de Emancipative Values para el caso peruano**.** Para este caso estableceremos la media 0.4098 que es la media de Emacipative Values para todos los países.

```
t.test(RESEMAVAL, mu=0,4098)
```

One Sample t-test

data: data1$RESEMAVAL

t = 6.6683, df = 1193, p-value = 3.948e-11

alternative hypothesis: true mean is not equal to 0.4098

95 percent confidence interval:

0.4280812 0.4433231

sample estimates:

mean of x

0.4357022

RESULTADO

En los resultados podemos observar que al 95% , el intervalo de confianza esta entre 0.4280812 y 0.4433231; además, observamos que se rechaza la hipótesis 0. Por lo que podemos decir que existen diferencias significativas entre el estadístico muestra y un parámetro igual a 0,4098.

Prueba t para Muestras Independientes:

Esta técnica utiliza una variable nominal dicotómica y una variable escalar. El objetivos es comparar ambas categorías de la variable nominal y determinar si existe igualdad o diferencia de medias entre ambas. De la misma manera que para los demás pruebas t, la Hipótesis Nula es la existencia de igualdad de medias y la Hipótesis Alternativa es la inexistencia de igualdad de medias. Esta técnica incluye como paso previo a la prueba, la prueba de Levene que calcula la homogeneidad de varianzas. La cual se realiza con el paquete *car* y el comando *LeveneTest*. Ejecutemos esta herramienta con la data que trabajamos la prueba anterior "Wordvs2015.sav" para el caso peruano y usaremos la variable nominal **V240 (SEXO)** y la variable escalar **RESEMAVAL (Índice de emancipación).**

```
install.packages("foreign") #necesario cargar el paquete foreign para utilizar archivos SPSS

library(foreign) #llamemos al paquete

data<-read.spss("WordVS2015.sav", to.data.frame=TRUE, use.value.labels = TRUE)

data1<-data[data$V2=="Peru",]

# creamos un nuevo objeto (data1) para filtrar sólo el caso peruano.

install.packages("Hmisc")

library("Hmisc")

describe(V240)
```

V240

n missing unique

```
  1210      0      2

Male (607, 50%)  Female (603, 50%)
```

```
describe(RESEMAVAL)

RESEMAVAL

     n   missing  unique    Info    Mean

  1194     16       940       1        0.4357

   .05       .10       .25       .50       .75

 0.2253  0.2680  0.3406  0.4319  0.5196

   .90       .95

 0.6132  0.6748

lowest : 0.05528 0.08250 0.09667 0.11524 0.11650

highest: 0.81944 0.83333 0.84218 0.86133 0.87417
```

```
install.packages("Hmisc") #Pedimos el paquete para pedir pruebas de normalidad.

library(nortest)

lillie.test(RESEMAVAL)$p.value #Pedimos la normalidad de las variables

[1] 0.1502064
```

Se puede apreciar que la variable escalar pasa la **Prueba de Normalidad**. El siguiente paso es la segmentación del archivo con el comando *transform*.

```
data2<-transform(data1, Sexo=factor(V240, labels=c("Male","Female")))

attach(data2)

library(car)  #Librería para la prueba de Levene

leveneTest(RESEMAVAL, Sexo, center=mean)   #Prueba de Levene

Levene's Test for Homogeneity of Variance (center = mean)

      Df F value Pr(>F)

group    1  0.7897 0.3744

     1192
```

La Prueba de Levene indica que se acepta la Hipótesis Nula: existencia de igualdad de varianzas y que se debe trabajar con varianzas homogéneas. Por lo tanto al pedir el comando *t.test* en el algoritmo *var.eqal* se digitará la opción **TRUE.**

```
t.test(RESEMAVAL~Sexo, alternative="two.sided", conf.level=.95,var.eqal=TRUE, data=data2)
```

```
        Welch Two Sample t-test


data:  RESEMAVAL by Sexo
t = 0.053244, df = 1190.9, p-value = 0.9575
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01482910  0.01565642
sample estimates:
 mean in group Male mean in group Female
          0.4359073           0.4354936
```

La prueba de significancia tiene un valor menor a 0.05; por lo tanto, se rechaza la Hipótesis Nula y se acepta la Hipótesis Alternativa: la existencia de diferencia de medida entre los grupos de la variable nominal.

Prueba t para muestras relacionadas:

Esta prueba sigue la misma lógica de las anteriores pruebas, la diferencia radica en que esta vez se busca comparar la media de una misma variable en dos momentos distintos. La Hipótesis Nula es la existencia de igualdad de medias en ambos contextos y la Hipótesis Alternativa es la inexistencia de igualdad de medias en ambos contextos. Para ejemplificar esta situación, se utilizarán las variables **RESEMAVAL y SECVALWGT**

```
describe(RESEMAVAL)
```

```
RESEMAVAL
      n  missing  unique    Info    Mean
   1194      16      940       1  0.4357
    .05     .10     .25     .50     .75
 0.2253  0.2680  0.3406  0.4319  0.5196
    .90     .95
 0.6132  0.6748
lowest : 0.05528 0.08250 0.09667 0.11524 0.11650
highest: 0.81944 0.83333 0.84218 0.86133 0.87417
```

describe(SECVALWGT)

```
SECVALWGT
      n missing  unique    Info    Mean
   1210       0       5    0.46  0.9794

          0.66 0.745 0.83 0.915   1
Frequency    1     8   50   165 986
%            0     1    4    14  81
```

install.packages("Hmisc") #Pedimos el paquete para pedir pruebas de normalidad.

library(nortest)

lillie.test(RESEMAVAL)$p.value #Pedimos la normalidad de las variables

[1] 0.1502064

lillie.test(SECVALWGT)$p.value

[1] 0

La Prueba de Normalidad, nos indica que se debe aceptar la Hipótesis Nula de normalidad; por lo tanto, se podría realizar la Prueba t para muestras relacionadas.

> t.test(RESEMAVAL, SECVALWGT, alternative='two.sided', conf.level=.95, paired=TRUE)

Paired t-test

data: RESEMAVAL and SECVALWGT

t = -135.15, df = 1193, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.5537115 -0.5378657

sample estimates:

mean of the differences

-0.5457886

Vemos que la significancia es menor a 0.05; por lo tanto se rechaza la Hipótesis Nula: existencia de igualdad de medias entre las variables, y se acepta la Hipótesis Alternativa: No existe igualdad de medias entre las variables. Esto nos permite asegurar que en dos contextos distintos esta variable se ha comportado de manera distinta.

Ahora es tu turno

1. El Índice de Desarrollo Humano, es un indicador social desarrollado por Programa de las Naciones Unidas para el Desarrollo que tiene como fin determinar el nivel de desarrollo que tiene los países en el mundo. Este indicador trata de observar un poco más allá del desempeño económico, por lo que utiliza la esperanza de vida de los ciudadanos; así como su progreso en educación. En ese sentido, podemos inferir lo heterogéneo que es esta variable, dada las grandes diferencias del IDH entre todos los países, por lo que es necesario observar que una región como América, cuyas diferencias entre países son relativamente grandes, puede ser un indicador generalizador respecto a lo que pasa en el mundo

   En la base de datos "economist" podrás encontrar el IDH de todos los países del mundo. Después de filtrar los países de la región "América", tienes que observar cómo ha sido la media del continente y, asumiendo, que la variable tiene un desempeño normal, observar con un intervalo de confianza del 95%, cuáles son los límites inferior y superior de la media; así como cuál es el error. Finalmente, realiza la media del IDH de todos los países y compara mediante una prueba T, si la media de América puede ser generalizadora para el resto del mundo.

2. El Global Study on Homicide 2013 de la Oficina de Naciones Unidades contra la Droga y el Delito, señala que, del total de personas que perdieron la vida en el mundo a causa de homicidios, el continente americano concentra el 36% de esas cifras; mientras que África, el 31%; Asia, el 28%; Europa concentra un 5%; y Oceanía un 0,3%. Asimismo, se observa que América del Sur posee una tasa promedio de homicidios de 23,4.

   En ese sentido, ahora te toca observa la base de datos sobre comisarias del Perú realizada por INEI para el año 2013, recolecta información respecto a algunos de los principales casos que la policía atiende, cotidianamente. Una de las variables titula "Delitos contra la vida, el cuerpo y la salud", en esta lo que se observa son todos los casos de homicidios, tentativa de homicidios, lesiones graves, aborto entre otras. Por lo que te toca comparar mediante una prueba T, como ha si la media de Perú es comparable con la media de la región.