Texto

Descripción generada automáticamente con confianza baja

**Tecnológico de Costa Rica**

Estadística MA-3405

**Proyecto I**

**Profesor:** Esteban Ballestero Alfaro.

**Estudiantes:**

José Luis Ramírez Hernández

José Pablo Espinoza López

José Carlo Hidalgo Chacón

David González Agüero

Grupo 50

**Fecha de Entrega:** 30 agosto de 2022

Intervalos de Confianza

30/3/2022

**Dataset**: Elecciones.xlsx

# Descripción general.

Los datos utilizados en este proyecto provienen de un formulario que se realizó a distintas personas de Costa Rica, este formulario se realizó a través de la herramienta de Google Forms, se hizo la publicación del formulario en diferentes medios sociales como los mostrados en la tabla anterior, con el objetivo de recolectar la mayor cantidad de datos de las personas de Costa Rica. Se publicó y se obtuvieron las respuestas necesarias (300) con el objetivo de tener resultados importantes como las edades, sexo de la persona, escolaridad, entre otros.

Se hará un estudio para el análisis inferencias de Intervalos de Confianza utilizando distintos tipos de distribuciones, en este estudio se hace uso de la Base de datos llamada Elecciones la cuál a continuación se describirán algunas características de su composición.

Esta compuesta por 300 filas y 8 columnas, las cuales contienen las siguientes variables:

**marcaTemporal:** Hace referencia a la fecha y hora exacta de cuando se obtuvo la respuesta a la encuesta.

*Tipo de datos:* PSIXct. (Almacenamiento de fechas y horas).

**sexo:** Se refiere al tipo de sexo de los entrevistados ya sea hombre o mujer.

*Tipo de datos:* character.

**edad:** Es la edad de las personas que participaron en la encuesta.

*Tipo de datos:* numeric.

**estudios:** Se refiere al nivel de estudios que posee una persona.

*Tipo de datos:* character.

**salarioMes:** Salario aproximado por mes de las personas entrevistadas.

*Tipo de datos:* character.

**Nota:** Al ser **salarioMes** de tipo character, se creará una nueva variable llamada **salario** con los datos de la variable salarioMes de tipo: numeric**.**

**partidoPolítico:** Nombre del partido político por el cuál la persona desea votar en las próximas elecciones presidenciales. Las categorías son: PLN, PAC, PNR, PUSC, Otros.

*Tipo de datos:* character.

**interesPolítica:** se refiere al interés de la persona sobre los temas de política. Aquí se escoge entre un intervalo que va desde el 0 al 10. También se permiten valores decimales.

*Tipo de datos:* character.

**Nota:** Se creó una nueva variable llamada **interes** con los datos de la variable llamada **interesPolítica.** La variable **interes** es de tipo: numeric.

**mediosInfo:** Medios utilizados para informarse sobre las noticias de política. Las categorías son: Radio, Televisión, Redes sociales, Periódico, Otros.

*Tipo de datos:* character.

|  |  |  |
| --- | --- | --- |
| **Resumen de variables seleccionadas** | | |
| |  |  |  | | --- | --- | --- | | **Una población** | **Dos poblaciones** | **Poblacion1:** Mujeres y hombres entre 18 años a 55 años. | |  | | **Variables** | **Tipo** | **Parámetro por estimar (IC)** | | salario | Cuantitativa | Media, desviación estándar, varianza poblacional. | | **Una población** | **Dos poblaciones** | **Población1:** Personas menores a 40 años | | **Población2:** Personas mayores a 39 años | | **Variables** | **Tipo** | **Parámetro por estimar (IC)** | | interesPolíticaN | Cuantitativa | Diferencia de promedios. | | | |
| **Una población** | **Dos poblaciones** | **Poblacion1:** Edad de todas las personas quienes apoyan a los partidos tradicionales (PLN,PUSC) |
|  |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| Edad | Cuantitativa | Media, Desviación estándar, Grados de libertad, Nivel de confianza. |
| **Una población** | **Dos poblaciones** | **Población1:** Edad de todas las mujeres con carreras universitarias que participan en la política. |
| **Población2:** Edad de todos los hombres con carreras universitarias que participan en la política. |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| Edad | Cuantitativa | Media, Desviación estándar, Grados de libertad, Nivel de confianza. |
| **Una población** | **Dos poblaciones** | **Poblacion1:** Las personas que se inclinan por el partido político PLN que tienen una educación universitaria completa. |
|  |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| Partido político | Cualitativa | Número de éxitos, número de casos totales, Nivel de confianza. |
| **Una población** | **Dos poblaciones** | **Población1:** Personas que prefieren el PAC y que su salario es mayor a 400,000 colones. |
| **Población2:** Personas que prefieren el PLN y que su salario es mayor a 400,000 colones. |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| Partido político  Salario | Cualitativa  Cuantitativa | Número de éxitos, número de casos totales, Nivel de confianza. |
| **Una población** | **Dos poblaciones** | **Poblacion1:** Mujeres que tengan un salario mensual mayor o igual a 200000 |
|  |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| salarioMujer | Cuantitativa | Nivel de confianza. |
| **Una población** | **Dos poblaciones** | **Población1:** Salarios de las personas entre 20 y 70 años que siguen al PLN. |
| **Población2:** Salarios de las personas entre 20 y 70 años que siguen al PAC. |
| **Variables** | **Tipo** | **Parámetro por estimar (IC)** |
| personasPAC  personasPLN | Cuantitativa | Nivel de confianza. |

# Análisis Inferencial

## IC Distribución T para una Población

Se asume que las personas entre adulta y adultas mayores son las que suelen apoyar los partidos políticos tradicionales. Las personas jóvenes suelen apoyar a partidos nuevos, por lo que se analizara la edad de todas las personas quienes tiene preferencia hacia los partidos políticos tradicionalistas (PLN, PUSC).

**Población:**

## [1] 74 71 44 43 35 21 45 35 22 18 22 22 22 25 20 22 18 29 20 24 20 22 22 29  
## [26] 46 60 57 28 46 24 38 24 43 51 45 47 28 23 21 78 65 72 70 71 54 24 90 70  
## [51] 70 70 23 23 70 38 34 45 35 78 23 42 62 37 72 25 49 27 34 42 52 62 55 53 41  
## [76] 43 35 23 23 55 34 42 25 35 38 76 27 33 22 49 21 20 32 46 35 23 58 32 56 32  
## [101] 64 50 65 59 58 49 50 58 49 60 32 52

Para analizar la hipótesis se tomará un grupo de 30 personas quienes apoyan al Partido Liberación Nacional o al Partido Unidad Social Cristiana. Este muestreo se realiza de forma aleatoria.

Suponiendo que la muestra poblacional aleatoria sigue una distribución normal se requiere un intervalo de confianza del 95%.

**Datos de la población:**

Variable aleatoria: Normal

Media poblacional: 43.1

Desviación Estándar: 17.82163

Tamaño de muestra: 30

Grados de libertad: 29

Nivel de confianza: 95%

Para calcular el intervalo de confianza de la distribución T, se utiliza la siguiente función de R.

*t.test(Datos de la muestra (), Nivel de confianza)$conf.int*

El cálculo del intervalo de confianza de 95% en una distribución t es:

t.test(muestra.Votantes.tradicionales,conf.level = 0.95)$conf.int

## [1] 36.4453 49.7547  
## attr(,"conf.level")  
## [1] 0.95

El rango de edades dentro del intervalo de confianza muestra que las edades de las personas que apoyan los partidos de Liberación Nacional y el partido de la Unidad social cristiana suele ser personas adultas de 36 a 50 años, por lo que la se demuestra que los más tradicionalistas son las personas adultas y no tanto los jóvenes.

**Gráfico de intervalo de confianza para la distribución t** Gráfico

Descripción generada automáticamente

##   
## 5% of the random confidence intervals do not contain Mu = 43.1.

## IC Distribución T para dos Población

Se tiene dos poblaciones en la cual es busca analizar la edad de las mujeres y hombres con estudios universitarios que participan en la política para ver cuál de ambas tiene la mayor población joven que participa actualmente en la política del país.

Se asume que los hombres con universidad completa que participan en la política son de mayor edad que las mujeres con universidad completa.

**Población de la edad de todas las mujeres:**

## [1] 47 37 23 23 38 38 57 38 45 32 28 61 48 33 39 23 45 25 29 71 27 24 45 70 27  
## [26] 24 42 42 40 25 34 38 51 35 32

**Población de la edad de todos los hombres:**

## [1] 74 36 36 35 23 23 40 24 22 27 43 22 29 59 48 24 42 51 44 78 56 69 69 28 54  
## [26] 30 23 34 29 37 25 29 55 53 41 58 33 55 29 42 45 27 38 32 32 56 32 35

Debido a que las poblaciones son mayores a 30 datos, se requiere hacer un muestreo de ambas poblaciones. Dicho muestreo se realiza de forma aleatoria.

**Muestra de la edad de las mujeres:**

## [1] 23 24 70 40 32 25 24 51 40 24 25 38 25 32 45 57 70 61 70 42 38 32 42 38 57  
## [26] 24 38 57 23 42

**Muestra de la edad de los hombres:**

## [1] 35 35 38 69 28 29 41 51 69 55 32 55 55 45 29 30 51 33 29 37 35 27 22 40 28  
## [26] 22 28 34 24 41

Suponiendo que la muestra poblacional de ambas poblaciones sigue una distribución normal se requiere un intervalo de confianza del 95%

**Datos de la población de edad de las mujeres:**

Variable aleatoria: Normal

Media poblacional: 40.3

Desviación Estándar: 15.14

Tamaño de muestra: 30

Grados de libertad: 29

Nivel de confianza: 95%

**Datos de la población de edad de los hombres:**

Variable aleatoria: Normal

Media poblacional: 38.23

Desviación Estándar: 12.78

Tamaño de muestra: 30

Grados de libertad: 29

Nivel de confianza: 95%

Para calcular el intervalo de confianza de la distribución T, se utiliza la siguiente función de R.

*t.test(Datos de la muestra uno (),Datos de la muestra dos (), Nivel de confianza)$conf.int*

t.test(muestra.HombresEdad ,muestra.MujeresEdad ,conf.level = 0.95)$conf.int

## [1] -9.312410 5.179076  
## attr(,"conf.level")  
## [1] 0.95

t.test(muestra.MujeresEdad,muestra.HombresEdad ,conf.level = 0.95)$conf.int

## [1] -5.179076 9.312410  
## attr(,"conf.level")  
## [1] 0.95

Por medio del intervalo de confianza se ha demostrado que no hay diferencias significativas de edades entre las 2 poblaciones debido a que el 0 se encuentra incluido dentro del intervalo, por lo que podemos asegurar que ambas poblaciones son iguales.

## Intervalos de Confianza para una proporción

Para el intervalo de confianza de un promedio, se analizan los datos recolectados, específicamente la población que prefiere al partido político PLN y se estudia la proporción de las personas que tienen un grado académico de universidad completa.

Se espera que el grado académico las personas que se inclinan por este partido tienen preferencia por el PLN es alto.

Para realizar este análisis se usa la siguiente fórmula:

*prop.test(x, n, conf.level)$conf.int*

Donde:

**x:**  es el número de casos exitosos o los que se quieren estudiar.

**n:**  el número total de casos.

**conf.level:** es el nivel de confiabilidad.

Ahora, se procede a obtener la cantidad de personas que se inclinan por el PLN

## [1] 86

Después se obtiene el total de personas se inclinan por el PLN y a la vez, tienen un estudio universitario completo.

## [1] 24

Con los datos obtenidos se realiza el intervalo de la siguiente manera:

**x:** 24

**n:** 86

**conf.level:** 0.95

#Calculando el intervalo  
  
prop.test(x=CantidadUniversidadPLN, n=CantidadTotalPLN, conf.level=0.95)$conf.int

## [1] 0.1902945 0.3878661  
## attr(,"conf.level")  
## [1] 0.95

A partir del intervalo obtenido se puede concluir, con un nivel de confianza del  
95%, que la proporción de las personas que se inclinan por el PLN y que tienen un nivel académico de Universidad completa está entre 0.1902945 0.3878661

Gráfico, Histograma

Descripción generada automáticamente

## 

## Intervalos de Confianza para diferencia de dos proporciones

Al estudiar la información obtenida en la base de datos, se sabe que, del total de 300 votantes, 50 se inclinan por el PAC y 86 por el PLN, se infiere que no hay mucha diferencia entre el salario de las dos poblaciones por lo que se procede a analizar cuantas personas ganan más de 400 000 colones al mes y con esta información se realiza el intervalo de confianza de una diferencia de poblaciones.

Para realizar este cálculo se usa la siguiente fórmula:

*prop.test(x = c(x1, x2), n = c(y1), y2), conf.level)$conf.int*

Donde:

**x:** vector con el conteo de éxitos de las dos muestras

**x1:** datos de la población 1 (de éxitos)

**x2:** datos de la población 2 (de éxitos)

**n:** vector con el número de ensayos de ambas muestras

**y1:** datos de la población 1 (de la muestra)

**y2:** datos de la población 2 (de la muestra)

**conf.level:** nivel de confianza.

Primero se obtienen los datos de los votantes del PLN y del PAC que ganan más de 400 000 colones al mes.

## [1] 500000 500000 7800000 750000 600000 550000 850000 600000 800000  
## [10] 700000 780000

## [1] 500000 750000 915000 500000 500000 500000 650000 6000000 50000  
## [10] 500000 5000000 65000 540000 700000 830000 600000 553000 650000  
## [19] 900000 600000 850000 500000 980000 550000 650000 650000 950000  
## [28] 750000 50000

Seguidamente se obtiene la cantidad de las personas que se inclinan por cada partido y que ganan más de 400 000 colones.

Luego, obtiene la cantidad total de votantes en cada partido.

## [1] 86

## [1] 50

Ahora se procede a calcular el intervalo.

#Calculando el intervalo  
  
prop.test(x = c(CantidadPLN, CantidadPAC), n = c((CantidadTotalPLN), (CantidadTotalPAC)), conf.level = 0.95)$conf.int

## [1] -0.05081233 0.28523093  
## attr(,"conf.level")  
## [1] 0.95

Con una confianza del 95% podemos concluir que la diferencia de proporciones entre las personas del PAC y del PLN y que ganan más de 400 mil colones mensuales se encuentra en el intervalo de: ]-0.05081233 , 0.28523093[. Lo que indica que el intervalo contiene al 0 en él por lo que se pude concluir que la diferencia entre las personas de las dos poblaciones no es tan significativa.

## IC de una varianza utilizando una población

Se quiere analizar el salario de las mujeres que tengan que este sea igual o supere los 200000 en el período que la encuesta fue realizada.

**Población:** Mujeres que tengan un salario mensual mayor o igual a 200000. **Variable cuantitativa:** salarioMujer.

Se asume que las mujeres que tengan un salario mensual mayor o igual a 200000 es normal debido a su tamaño de muestra > 30.

Para calcular el intervalo de confianza de una variancia, se utiliza la siguiente función de R.

*varTest(salarioMujer, conf.level = 0.95)$conf.int*

varTest(salarioMujer, conf.level = 0.95)$conf.int

## LCL UCL   
## 8.343435e+11 1.659251e+12   
## attr(,"conf.level")  
## [1] 0.95

De acuerdo con el intervalo de confianza con un nivel de 95% que se obtiene se concluye, que la proporción de Mujeres votantes que reciben un salario mayor o igual a 200000 colones está entre [8.343435e+11, 1.659251e+12]

## IC de un cociente de variancias utilizando dos poblaciones

Se quiere comparar ambos grupos, con respecto a la afinidad hacia dos partidos políticos de las personas entre 30 y 50 años.

Poblaciones:

**X:** Personas entre 20 y 70 años que siguen al PLN.

**Y:** Personas entre 20 y 70 años que siguen al PAC

Se asume que ambas poblaciones son normales ya que tienen el tamaño de muestra >30

Para calcular el intervalo de confianza de un cociente de variancias, se utiliza la siguiente función de R.

stests :: var.test(Población 1, Población 2,conf.level = 0.95)$conf.int

IC de un cociente de variancias utilizando dos poblaciones:

stests :: var.test(personasPLN, personasPAC,conf.level = 0.95)$conf.int

## [1] 0.4019532 1.1592313  
## attr(,"conf.level")  
## [1] 0.95

A través del intervalo de confianza se ha demostrado que la varianza que se da de los ingresos son significativo entre las 2 poblaciones debido al ser este [0.4019532, 1.1592313], por lo que podemos asegurar que ambas poblaciones son similares.

## 

## IC de un promedio usando distribución Z.

Se quiere analizar el salario de las personas entre 18 años y 60 años.

**Población:** personas entre 18 años y 60 años.

**Variable cuantitativa:** salario.

Por medio del Teorema del límite Central se asume que la población de personas entre 18 años y 60 años es normal debido a su tamaño de muestra >=30.

#### 

#### Calculo:

Para poder hacer la implementación de un promedio utilizando la distribución z, se utilizarán 3 paquetes: *PASWR2*, *BSDA* y *STESTS*.

Al automatizar el proceso se puede usar la función de R llamada *z.test* del paquete *PASWR2* cuya estructura seria:

z.test(x, sigma.x, conf.level = 0.95)

donde:

* **x:** vector de valores numéricos.
* **sigma.x:** desviación estándar poblacional.
* **conf.level:** nivel de confianza del intervalo, está restringido por valores entre 0 y 1.

Para solamente mostrar el Intervalo de Confianza se utiliza el valor de *conf.int*

z.test(x, sigma.x, conf.level = 0.95)$conf.int

donde:

* **conf.int:** Intervalo de confianza.

A continuación, la implementación de z.tests del paquete PASWR2:

PASWR2::z.test(x = salario, sigma.x = salarioDE, conf.level = 0.90)$conf.int

## [1] 657029.2 857396.5  
## attr(,"conf.level")  
## [1] 0.9

Seguidamente se detalla la función *z.test* para el paquete *BSDA* cuya estructura seria:

z.test(x, sigma.x, conf.level = 0.95)

donde:

* **x:** vector de valores numéricos.
* **sigma.x:** desviación estándar poblacional.
* **conf.level:** nivel de confianza del intervalo, está restringido a valores entre 0 y 1.

Para mostrar el IC se utiliza el valor de *conf.int*

z.test(x, sigma.x, conf.level = 0.95)$conf.int

donde: - **conf.int:** Intervalo de confianza.

Implementación de z.test utilizando el paquete BSDA:

BSDA::z.test(x = salario, sigma.x = salarioDE, conf.level = 0.90)$conf.int#comentario

## [1] 657029.2 857396.5  
## attr(,"conf.level")  
## [1] 0.9

Finalmente, dentro del paquete *stests* existe la función *z.test* cuya estructura es:

z.test(x, sigma2, conf.level = 0.95)

donde:

* **x:** vector de valores numéricos.
* **sigma2:** varianza poblacional conocida.
* **conf.level:** nivel de confianza dle intervalo, por defecto viene en 0.95

*conf.int* mostrará únicamente el Intervalo de Confianza:

z.test(x, sigma2, conf.level = 0.95)$conf.int

donde:

* **conf.int:** Intervalo de confianza.

A continuación, se muestra la implementación utilizando z.test del paquete stests:

stests::z.test(x = salario, sigma2 = varianza, conf.level = 0.90)$conf.int

## [1] 657029.2 857396.5  
## attr(,"conf.level")  
## [1] 0.9

Al observar los resultados obtenidos podemos concluir que el salario de las personas entre 18 y 60 años ronda entre los 65 7029 colones y 85 7366 colones.

En el caso de realizar análisis con ***dos poblaciones***, el paquete *stests* no trae una función que permita hacer el análisis con dos poblaciones. Por lo tanto, se hará el estudio con los paquetes *PASWR2* y *BSDA*.

## 

## IC de una diferencia de promedios usando distribución Z.

Sean X e Y variables aleatorias, donde:

**X:** Población de personas menores a 40 años.

**Y:** Población de personas mayores a 39 años.

Supongamos que la población de personas menores a 40 años y la población de personas mayores a 39 años, siguen una distribución normal.

Se quiere comparar ambos grupos, con respecto al interés de la política.

Automatizando el proceso utilizamos la función de R llamada *z.test* del paquete *PASWR2* cuya estructura para dos poblaciones seria:

z.test(x, y, sigma.x, sigma.y, conf.level = 0.95)

donde:

* **x:** vector de valores numéricos.
* **y:** vector de valores numéricos opcional.
* **sigma.x:** desviación estándar poblacional de x.
* **sigma.y:** desviación estándar poblacional de y.
* **conf.level:** nivel de confianza del intervalo, está restringido por valores entre 0 y 1.

Para mostrar el IC se utiliza el valor de *conf.int* :

z.test(x, y, sigma.x, sigma.y, conf.level = 0.95)$conf.int

donde:

-**conf.int:** Intervalo de confianza.

A continuación, la implementación de z.tests del paquete PASWR2:

PASWR2::z.test(x = interesPol1, y = interesPol2, sigma.x = desvE1, sigma.y = desvE2 ,conf.level = 0.90)$conf.int

## [1] -0.6320728 0.4230857  
## attr(,"conf.level")  
## [1] 0.9

En el caso del paquete *BSDA* la estructura de *z.test* es:

z.test(x, y, sigma.x, sigma.y, conf.level = 0.95)

donde:

* **x:** vector de valores numéricos de x.
* **y:** vector de valores numéricos de y.
* **sigma.x:** desviación estándar poblacional de x.
* **sigma.y:** desviación estándar poblacional de y.
* **conf.level:** nivel de confianza del intervalo, está restringido por valores entre 0 y 1.

Para mostrar el IC se utiliza el valor de conf.int

z.test(x, y, sigma.x, sigma.y, conf.level = 0.95)$conf.int

donde:

* **conf.int:** Intervalo de confianza.

A continuación, la implementación de z.tests del paquete BSDA:

Implementación de z.test utilizando el paquete BSDA:

BSDA::z.test(x = interesPol1, y = interesPol2, sigma.x = desvE1, sigma.y = desvE2 ,conf.level = 0.90)$conf.int

## [1] -0.6320728 0.4230857  
## attr(,"conf.level")  
## [1] 0.9

Podemos concluir que al tener una probabilidad del 90% la diferencia de las medias poblacionales está en algún lugar de este intervalo.

También notamos que ambos grupos tienen el mismo promedio debido a que el valor 0 está contenido dentro del intervalo *]-0.6320728, 0.4230857[*

Por lo tanto, el grupo de personas menores a 40 años como el grupo de personas mayores a 39 años tienen el mismo nivel de interés en la política.

# Gráfico de Partidos Políticos

Para finalizar, se mostrará un gráfico el cual contiene en el eje x, los partidos políticos y en el eje y, las edades de todos los participantes.

En el gráfico los partidos políticos se encuentran enumerados del 1 al 5 donde:

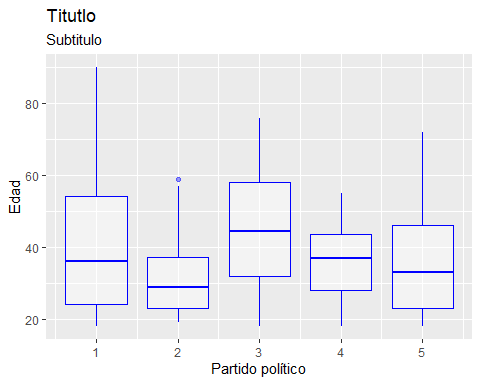
**1** es el Partido Liberación Nacional.

**2** es el Partido Acción Ciudadana.

**3** es el Partido Unidad Social Cristiana.

**4** es el Partido Nueva República.

**5** se refiere a otros partidos políticos.



Se puede observar que el partido que presenta más variabilidad es el de PLN, ya que posee edades desde antes de 20 años y mayor a 80 años.

Como dato atípico tenemos a una persona de 60 años que sigue el PAC.