Métodos estadísticos

Tarea 2

González Paz Edgar

Considere una muestra aleatoria $X_1, \dots, X_n$ con función de densidad que proviene de una $Normal(\mu, \sigma^2)$, ambos parámetros desconocidos, se sabe que $X \sim N(\mu, \frac{\sigma^2}{n})$ y $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ y además son independientes.

1. Construya una prueba de hipótesis para el contraste $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ de tamaño $\alpha = 0.05$.

La función de densidad de una $N(\mu, \sigma^2)$.

$$f(x|\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right)$$

La función de verosimiltud es

$$L(\mu, \sigma^2|\mathrm{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2} \right) \mathbb{1}_{\{x_i \in \mathbb{R}\}} =$$
$$\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \right)$$

Sea $\theta = (\mu, \sigma^2)$ y el estadístico del cociente de verosimilitud es de la siguiente manera:

$$\lambda(\mathrm{X}) = \frac{sup_{\theta \in \Theta_0} L(\theta|\mathrm{X})}{sup_{\theta \in \Theta} L(\theta|\mathrm{X})}$$

donde el espacio paramétrico de θ es $\Theta = \{\mu, \sigma^2| -\infty < \mu < \infty, \sigma^2 > 0\}$. El subconjunto de la hipótesis nula es $\Theta_0 = \{\mu, \sigma^2|\mu = 0, \sigma^2 > 0\}$ y para la hipótesis alternativa es $\Theta_1 = \{\mu, \sigma^2|\mu \neq 0, \sigma^2 > 0\}$.

Recordemos que

$$\sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i-\mu)^2$$

Para el numerador tomamos a $\mu = 0$, ahora bien, σ^2_0 se vería de la siguiente manera:

$$\sigma^2_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i-0)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Entonces

$$\sup_{\theta\in\Theta_0} L(\theta|\mathrm{X}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} x_i^2\right) =$$

$$\left(2\pi\sigma_0^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

$$\Rightarrow \sup_{\theta\in\Theta_0} L(\theta|\mathrm{X}) = \left(2\pi\sigma_0^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

Para $sup_{\theta\in\Theta} L(\theta|\mathrm{X})$, por el método de máxima verosimilitud, los estimadores de máxima verosimilitud $\hat{\mu}_{MLE}$ y $\hat{\sigma}_{MLE}$ son de la siguiente forma:

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$$

$$\hat{\sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \hat{\mu}_{MLE}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2$$

$$\sup_{\theta\in\Theta} L(\theta|\mathrm{X}) = \left(2\pi\sigma\hat{}_{MLE}\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma\hat{}_{MLE}}\sum_{i=1}^{n} \left(x_i - \hat{\mu}_{MLE}\right)^2\right) =$$

$$\left(2\pi\sigma\hat{}_{MLE}\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

$$\Rightarrow \sup_{\theta\in\Theta} L(\theta|\mathrm{X}) = \left(2\pi\sigma\hat{}_{MLE}\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)$$

El estadístico $\lambda(\mathrm{X})$

$$\lambda(\mathrm{X}) = \frac{\sup_{\theta\in\Theta_0} L(\theta|\mathrm{X})}{\sup_{\theta\in\Theta} L(\theta|\mathrm{X})} = \frac{\left(2\pi\sigma_0^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)}{\left(2\pi\sigma\hat{}_{MLE}\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)} =$$

$$\left(\frac{\sigma_0^2}{\sigma\hat{}_{MLE}}\right)^{-\frac{n}{2}} = \left(\frac{\sigma\hat{}_{MLE}}{\sigma_0^2}\right)^{\frac{n}{2}} = \left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right)^{\frac{n}{2}}$$

La región de rechazo es la siguiente:

$$RR = \left\{\mathrm{X} \,\middle|\, \left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{\sum_{i=1}^{n} x_i^2}\right)^{\frac{n}{2}} \geq c\right\}$$

Por una parte, sabemos que $\sigma_0^2 = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 + n\bar{X}^2$

$$\left(\frac{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}{\sum_{i=1}^{n}x_i^2}\right)^{\frac{n}{2}}\leq c$$

$$\Rightarrow \frac{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2+n\bar{X}^2}\leq c^{\frac{2}{n}}$$

$$\Rightarrow \frac{1}{1+\frac{n\bar{X}^2}{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}}\leq c^{\frac{2}{n}}$$

$$Sea\ c_1=c^{\frac{2}{n}},$$

$$\Rightarrow \frac{1}{c_1}\leq 1+\frac{n\bar{X}^2}{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}$$

$$\Rightarrow \frac{1}{c_1}-1\leq \frac{n\bar{X}^2}{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}$$

$$Sea\ c_2=\frac{1}{c_1}-1$$

$$\Rightarrow c_2\leq \frac{(\frac{1}{n-1})n\bar{X}^2}{(\frac{1}{n-1})\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}=\frac{(\frac{1}{n-1}n\bar{X}^2)}{S^2}$$

$$\Rightarrow (n-1)c_2\leq \frac{n\bar{X}^2}{S^2}$$

$$\Rightarrow (n-1)c_2\leq \frac{\sqrt{n}\bar{X}}{S}^2$$

$$\Rightarrow \sqrt{((n-1)c_2)}\leq \sqrt{(\frac{\sqrt{n}\bar{X}}{S})^2}$$

$$Sea\ \sqrt{(n-1)c_2}=k$$

$$\Rightarrow k\leq \sqrt{(\frac{\sqrt{n}\bar{X}}{S})^2}$$

$$\Rightarrow k\leq \left|\frac{\sqrt{n}\bar{X}}{S}\right|$$

Y entonces,

$$RR=\left\{\mathrm{X}\ \middle|\ \left|\frac{\sqrt{n}\bar{X}}{S}\right|\geq k\right\}$$

Ahora bien, reescribimos:

$$\frac{\sqrt{n}\bar{X}}{S}=\frac{\bar{X}-0}{\sqrt{\frac{S^2}{n}}}=\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}}$$

Sea $T = \frac{\sqrt{n}\bar{X}}{S}$, entonces $T \sim t_{n-1}$ y la Región de Rechazo se ve de la manera:

$$RR = \{\mathrm{X} \mid |T| \geq k\} \; para \; algún \; k \in \mathbb{R}$$

Buscamos $\mathbb{P}\left(Error \; tipo \; I\right) = \mathbb{P}\left(\mathrm{X} \in RR | \theta \in \Theta_0\right) = \alpha$

$$\mathbb{P}\left(Error \; tipo \; I | \mu = 0\right) = \mathbb{P}\left(|T| \geq k | \mu = 0\right) =$$
$$|| \; Simétrica \; en \; el \; origen \; || = 2\mathbb{P}\left(T \geq k | \mu = 0\right) = \alpha$$
$$\Rightarrow \mathbb{P}\left(T \geq k | \mu = 0\right) = \frac{\alpha}{2}$$

Sabemos que $T \sim t_{n-1}$ entonces tomamos a $k = t_{n-1, 1-\frac{\alpha}{2}}$

$$\Rightarrow \mathbb{P}\left(|T| \geq k\right) = \mathbb{P}\left(|T| \geq t_{n-1,1-\frac{\alpha}{2}}\right) = \alpha$$

La región de rechazo es:

$$RR = \left\{\mathrm{X} \,\middle|\, t_{n-1,1-\frac{\alpha}{2}} < |T|\right\}$$

Encontrando el valor de $t_{n-1,1-\frac{\alpha}{2}}$ en valores de tablas o evaluando el valor en la inversa de la función de distribución. El valor es 2.0638.

Entonces la región de rechazo es de la forma:

$$RR = \{\mathrm{X} \mid 2.0638 < |T|\}$$

Para el p-value, supongamos que se rechaza bajo la muestra observada, es decir,

$$\{\mathrm{X} \mid 3.2958 < |T|\}$$

Consideramos que bajo la m.a.o., $T = 3.2958$.

$$\Rightarrow \mathbb{P}\left(3.2958 < |T|\right) = ||Simetría|| = 2\mathbb{P}\left(3.2958 < T\right) =$$
$$2(1 - \mathbb{P}\left(T < 3.2958\right)) = p - value$$
$$T \sim t_{24}$$

2. Con ayuda de la muestra aleatoria observada, data2Ej4.txt , diga si se rechaza o no se rechaza la hipótesis nula y calcule el p-value.

In [1]:
```
#Librerías para las distribuciones de la t-student, ji-cuadrada y la t-student nc
from scipy.stats import t, chi2, nct
```

In [2]:
```
#Cargamos la m.a.o.
import pandas as pd
datos_4 = pd.read_csv('data2Ej4.txt')
```

In [3]:
```
datos_4.head()
```

Out[3]:

	x
0	1.708724
1	0.769615
2	-1.260394
3	1.742702
4	2.300945

In [4]:
```
datos_4.shape
```

Out[4]:
```
(25, 1)
```

In [5]:
```
import numpy as np
Muestra_Aleatoria_Observada = np.asarray(datos_4['x'].values.tolist())
```

In [6]:
```
#M.a.o.
Muestra_Aleatoria_Observada
```

Out[6]:
```
array([ 1.70872406,  0.76961521, -1.26039356,  1.74270239,  2.30094473,
        0.05077468,  1.35129325,  0.20134414,  0.26227203, -0.42621713,
        0.81476031,  0.72918117, -0.15321172,  0.91067085, -0.39005322,
        0.72655107,  0.15187769,  0.3779972 ,  2.7231732 ,  1.73695479,
        0.0895813 , -0.80422873, -0.39844238,  1.91218168,  1.58635116])
```

In [7]:
```
def MediaMuestral(x):
    n = len(x)
    y = (1/n)*sum(x)
    return y
```

In [8]:
```
def Cuasivarianza(x):
    n = len(x)
    y = (1/(n-1))*sum((x[i]- MediaMuestral(x))**2 for i in range(n))
    return y
```

In [9]:
```
def Sigma_MLE(x):
    n = len(x)
    y = (1/(n))*(sum((x[i] - MediaMuestral(x))**2 for i in range(n)))
    return y
```

In [10]:
```
def T(x, media_hipnul):
    n = len(x)
    y = (MediaMuestral(x) - media_hipnul)/(np.sqrt(Cuasivarianza(x)/n))
    return y
```

In [11]:
```python
T(Muestra_Aleatoria_Observada, media_hipnul=0)
```

Out[11]:
```
3.295825419112768
```

In [12]:
```python
def PruebaHipotesis(MA, MHP, alpha):
    Estadistico_T = T(x=MA, media_hipnul=MHP)
    n = len(MA)
    MM_MA = MediaMuestral(MA)
    Cuasivarianza_MA =Cuasivarianza(MA)
    #Es la cota de la región de rechazo
    cota_decision = t.ppf(1-(alpha/2), df= n-1)

    print('La muestra observada es: '+str(MA)+'\n')
    print('El tamaño de la muestra es: '+str(n))
    print('La media muestral es: '+str(MM_MA))
    print('La cuasivarianza es: '+str(Cuasivarianza_MA))
    print('El valor del estadístico T es: '+str(Estadistico_T))
    print('La cota de decisión es: '+str(cota_decision)+'\n')

    decision = Estadistico_T > cota_decision

    if decision == 1:
        print('El estadístico T='+str(Estadistico_T)+' es mayor al valor '+str(cc
    else:
        print('El estadístico T='+str(Estadistico_T)+' es menor al valor '+str(cc

    p_value = 2*(1-t.cdf(Estadistico_T, df=24))

    decision_2 = p_value > alpha

    if decision_2 == True:
        print('El valor del p-value: '+str(p_value)+' es mayor al valor '+str(cot
    if decision_2 == False:
        print('El valor del p-value: '+str(p_value)+' es menor al valor '+str(cot
```

In [13]:
```python
PruebaHipotesis(MA=Muestra_Aleatoria_Observada, MHP=0, alpha=0.05)
```

```
La muestra observada es: [ 1.70872406  0.76961521 -1.26039356  1.74270239  2.30
094473  0.05077468
  1.35129325  0.20134414  0.26227203 -0.42621713  0.81476031  0.72918117
 -0.15321172  0.91067085 -0.39005322  0.72655107  0.15187769  0.3779972
  2.7231732   1.73695479  0.0895813  -0.80422873 -0.39844238  1.91218168
  1.58635116]

El tamaño de la muestra es: 25
La media muestral es: 0.6685761668633025
La cuasivarianza es: 1.0287583963320184
El valor del estadístico T es: 3.295825419112768
La cota de decisión es: 2.0638985616280205

El estadístico T=3.295825419112768 es mayor al valor 2.0638985616280205, por lo
tanto se rechaza la hipótesis nula.
El valor del p-value: 0.0030427364777887433 es menor al valor 2.063898561628020
5, por lo tanto se rechaza la hipótesis nula.
```

4. Encontrar un intervalo de confianza

Para el intervalo de confianza, definimos la región de no rechazo:
$A(\theta_0) = \left\{ \mathrm{X} \,\middle|\, t_{n-1;\frac{\alpha}{2}} < T < t_{n-1;1-\frac{\alpha}{2}} \right\}$

Sea el conjunto aleatorio $C(\mathrm{X}) = \{\theta_0 \in \Theta \mid \mathrm{X} \in A(\theta_0)\}$ Entonces

$$C(\mathrm{X}) = \left\{ \mu_0 \,\middle|\, t_{n-1;\frac{\alpha}{2}} < T < t_{n-1;1-\frac{\alpha}{2}} \right\} =$$

$$\left\{ \mu_0 \,\middle|\, t_{n-1;\frac{\alpha}{2}} < \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < t_{n-1;1-\frac{\alpha}{2}} \right\} =$$

$$\|\ t_{n-1;\frac{\alpha}{2}} = -t_{n-1;1-\frac{\alpha}{2}}\ \| = \left\{ \mu_0 \,\middle|\, -t_{n-1;1-\frac{\alpha}{2}} < \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < t_{n-1;1-\frac{\alpha}{2}} \right\} =$$

$$\left\{ \mu_0 \,\middle|\, -t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \bar{X} - \mu_0 < t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right\} =$$

$$\left\{ \mu_0 \,\middle|\, -t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} - \bar{X} < -\mu_0 < t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} - \bar{X} \right\} =$$

$$\left\{ \mu_0 \,\middle|\, \bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu_0 < \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right\}$$

$\Rightarrow \left(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$ es un intervalo de confianza para μ_0.

In [14]:
```
def Intervalo_de_Confianza(Muestra_Aleatoria_Observada, alpha):
    n = len(Muestra_Aleatoria_Observada)
    MM_MA = MediaMuestral(Muestra_Aleatoria_Observada)
    Cuasivarianza_MA =Cuasivarianza(Muestra_Aleatoria_Observada)

    Intervalo_confianza_media = []
    Intervalo_confianza_media.append(MM_MA - (t.ppf(1-(alpha/2), df= 24)*(np.sqrt
    Intervalo_confianza_media.append(MM_MA + (t.ppf(1-(alpha/2), df= 24)*(np.sqrt

    print('El intervalo de confianza para la media es: '+str(Intervalo_confianza_
```

In [15]:
```
Intervalo_de_Confianza(Muestra_Aleatoria_Observada, alpha=0.05)
```

```
El intervalo de confianza para la media es: [0.24990308390024746, 1.08724924982
63575]
```

3. Gráficar la función potencia de la prueba.

Para la función potencia:

$$\pi(\theta) = \mathbb{P}\,(\mathrm{X} \in RR) =$$
$$\mathbb{P}\left(T \in (-\infty, -t_{n-1,1-\frac{\alpha}{2}}) \cup (t_{n-1,1-\frac{\alpha}{2}}, \infty) | \mu = 0\right) =$$
$$1 - \mathbb{P}\left(T \notin (-\infty, -t_{n-1,1-\frac{\alpha}{2}}) \cup (t_{n-1,1-\frac{\alpha}{2}}, \infty) | \mu = 0\right) =$$
$$1 - \mathbb{P}\left(T \in (-t_{n-1,1-\frac{\alpha}{2}}, t_{n-1,1-\frac{\alpha}{2}}) | \mu = 0\right) =$$
$$1 - \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} < \frac{\bar{X}}{\sqrt{\frac{S^2}{n}}} < t_{n-1,1-\frac{\alpha}{2}} | \mu = 0\right) =$$
$$1 - \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} < \frac{(\bar{X} + \mu) - (\mu - 0)}{\sqrt{\frac{S^2}{n}}} < t_{n-1,1-\frac{\alpha}{2}} | \mu = 0\right) =$$
$$1 - \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} < \frac{\frac{(\bar{X}+\mu)}{\sqrt{\frac{(\sigma^2}{n}}} - \frac{(\mu-0)}{\sqrt{\frac{\sigma^2}{n}}}}{\frac{\sqrt{\frac{S^2}{n}}}{\sqrt{\frac{\sigma^2}{n}}}} < t_{n-1,1-\frac{\alpha}{2}} | \mu = 0\right)$$

$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$, $S^2 \sim Gamma(n-1, \sigma^2)$ y $\frac{S^2}{\sigma^2} \sim Gamma(n-1, 1)$.

Entonces $W_{n-1} = \frac{\frac{(\bar{X}+\mu)}{\sqrt{\frac{\sigma^2}{n}}} - \frac{\mu-0)}{\sqrt{\frac{\sigma^2}{n}}}}{\frac{\sqrt{\frac{S^2}{n}}}{\sqrt{\frac{\sigma^2}{n}}}} \sim t_{n-1, \frac{\mu-0}{\sqrt{\frac{\sigma^2}{n}}}}$, una T-student con n-1 grados de libertad y un parámetro de descentralización de $\frac{(\mu-0)}{\sqrt{\frac{\sigma^2}{n}}}$

$$\Rightarrow \pi(\theta) = 1 - \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} < W_{n-1} < t_{n-1,1-\frac{\alpha}{2}}\right) =$$
$$1 - (F_{W_n-1}(t_{n-1,1-\frac{\alpha}{2}}) - F_{W_n-1}(-t_{n-1,1-\frac{\alpha}{2}}))$$

In [16]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

In [17]:
```
def FuncionPotencia(Muestra_Aleatoria_Observada, med_hip_nul, alpha):
    n = len(Muestra_Aleatoria_Observada)
    def Sigma_MLE(x):
        n = len(x)
        y = (1/(n))*(sum((x[i] - MediaMuestral(x))**2 for i in range(n)))
        return y

    z = t.ppf(1-(alpha/2), df= n -1)

    Dominio_mu=np.arange(med_hip_nul-2.5, med_hip_nul+2.5 ,0.01)
    Rango_Funcion_Potencia = []
    for i in range(len(Dominio_mu)):
        Rango_Funcion_Potencia.append(1 - (nct.cdf(z, n-1, (Dominio_mu[i]/(np.sqr
    plt.plot(Dominio_mu, Rango_Funcion_Potencia)
    plt.title('Función potencia')
```

In [18]:
```
Sigma_MA = Sigma_MLE(Muestra_Aleatoria_Observada)
```

In [19]:
```
FuncionPotencia(Muestra_Aleatoria_Observada, med_hip_nul=0, alpha=0.05)
```