MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

Clase 2 - Parte 2: Métodos Estadísticos para IA

Test de Hipótesis de la media y la varianza.

Objetivos de esta presentación

- Comprender el test de hipótesis para la media.
- Comprender el test de hipótesis para la varianza.
- Aprender herramientas de software para estos tests.

Resumen

- ✓ Test de hipótesis para la media con varianza conocida
- ✓ Test de hipótesis para la media con varianza desconocida para muestras grandes
- ✓ Test de hipótesis para la media de una distribución normal con varianza desconocida para muestras pequeñas (*t*-Test)
- ✓ Test de hipótesis para la varianza

Resumen

- ✓ Test de hipótesis para la media con varianza conocida
- ✓ Test de hipótesis para la media con varianza desconocida para muestras grandes
- ✓ Test de hipótesis para la media de una distribución normal con varianza desconocida para muestras pequeñas (t-Test)
- ✓ Test de hipótesis para la varianza

Test de hipótesis para la media con varianza conocida

Aquí vamos a formalizar lo visto en la **Presentación 2.1**:

Supongamos que la variable aleatoria de interés X tiene una **media $\boldsymbol{\mu}$ y una varianza $\boldsymbol{\sigma^2}$ conocida**.

Asumimos que X tiene **distribución normal**, es decir $X \sim N(\mu, \sigma^2)$.

Podemos tomar como **estadístico de prueba** a la media muestral $\bar{X}$, la cual tendrá una distribución $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Si normalizamos a $\bar{X}$ de la forma $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ obtenemos el **estadístico de prueba Z con distribución normal estándar** $\mathrm{Z} \sim N(0,1)$.

Test de hipótesis para la media con varianza conocida

Caso 1: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa bilateral** siguientes:

$$H_0\colon \mu = \mu_0 \quad \text{contra} \quad H_1\colon \mu \neq \mu_0$$

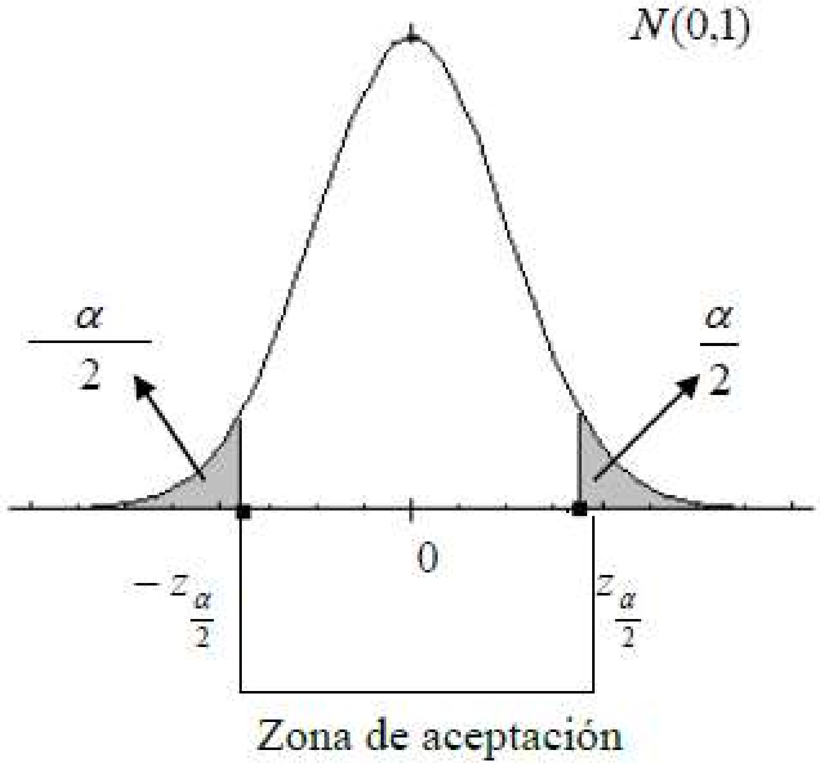
donde μ_0 es una constante específica. Entonces, el estadístico de prueba $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ tiene distribución $N(0,1)$ si $H_0\colon \mu = \mu_0$ es verdadera.

Si $H_0\colon \mu = \mu_0$ es verdadera entonces:

$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$

Por lo tanto, la **regla de decisión** es:

- Rechazar H_0 si $|Z_0| > z_{\alpha/2}$
- Aceptar H_0 si $|Z_0| \leq z_{\alpha/2}$



Test de hipótesis para la media con varianza conocida

Caso 2: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

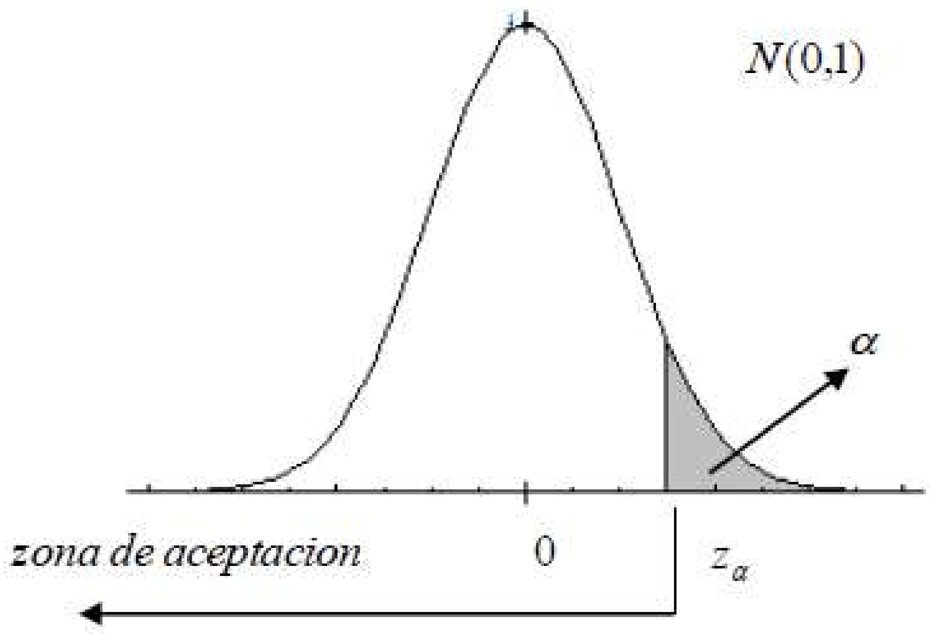
$$H_0\colon \mu = \mu_0 \quad \text{contra} \quad H_1\colon \mu > \mu_0$$

donde μ_0 es una constante específica. Entonces, el estadístico de prueba $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ tiene distribución $N(0,1)$ si $H_0\colon \mu = \mu_0$ es verdadera.

Si $H_0\colon \mu = \mu_0$ es verdadera entonces:
$P(Z \leq z_\alpha) = 1 - \alpha$

Por lo tanto, la **regla de decisión** es:

- Rechazar H_0 si $Z_0 > z_\alpha$
- Aceptar H_0 si $Z_0 \leq z_\alpha$



Test de hipótesis para la media con varianza conocida

Caso 3: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0\colon \mu = \mu_0 \quad \text{contra} \quad H_1\colon \mu < \mu_0$$

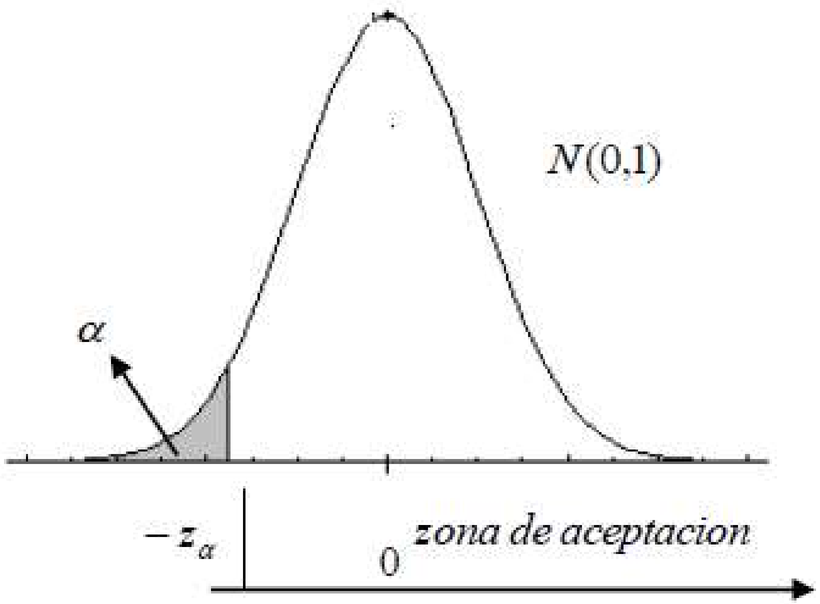
donde μ_0 es una constante específica. Entonces, el estadístico de prueba $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ tiene distribución $N(0{,}1)$ si $H_0\colon \mu = \mu_0$ es verdadera.

Si $H_0\colon \mu = \mu_0$ es verdadera entonces:

$P(Z \geq -z_\alpha) = 1 - \alpha$

Por lo tanto, la **regla de decisión** es:

- Rechazar H_0 si $Z_0 < -z_\alpha$
- Aceptar H_0 si $Z_0 \geq -z_\alpha$



Test de hipótesis para la media con varianza conocida

Cálculo del p-valor:

Caso 1: Si las hipótesis son H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$ entonces:

$$\boldsymbol{p-valor} = P(|Z| > |Z_0|) = 1 - P(|Z| < |Z_0|) = 1 - [\Phi(|Z_0|) - \Phi(-|Z_0|)] = 1 - [2\Phi(|Z_0|) - 1] = \mathbf{2[1 - \Phi(|Z_0|)]}$$

Caso 2: Si las hipótesis son H_0: $\mu = \mu_0$ contra H_1: $\mu > \mu_0$ entonces:

$$\boldsymbol{p-valor} = P(Z > Z_0) = 1 - P(Z \leq Z_0) = \mathbf{1 - \Phi(Z_0)}$$

Caso 3: Si las hipótesis son H_0: $\mu = \mu_0$ contra H_1: $\mu < \mu_0$ entonces:

$$\boldsymbol{p-valor} = P(Z < Z_0) = \mathbf{\Phi(Z_0)}$$

Test de hipótesis para la media con varianza conocida

Cálculo del error de tipo II (β) y del tamaño de la muestra (n)

Caso 1: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu \neq \mu_0$ entonces:

$$\beta(\mu) = \Phi\left(Z_{\alpha/2} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) - \Phi\left(-Z_{\frac{\alpha}{2}} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) \quad \text{y} \quad n > \frac{\left(Z_{\alpha/2}+Z_{\beta_0}\right)^2\sigma^2}{(\mu-\mu_0)^2}$$

Caso 2: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu > \mu_0$ entonces:

$$\beta(\mu) = \Phi\left(Z_{\alpha} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) \quad \text{y} \quad n > \frac{\left(Z_{\alpha}+Z_{\beta_0}\right)^2\sigma^2}{(\mu-\mu_0)^2}$$

Caso 3: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu < \mu_0$ entonces:

$$\beta(\mu) = 1 - \Phi\left(-Z_{\alpha} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) \quad \text{y} \quad n > \frac{\left(Z_{\alpha}+Z_{\beta_0}\right)^2\sigma^2}{(\mu-\mu_0)^2}$$

Resumen

- ✓ Test de hipótesis para la media con varianza conocida
- ✓ Test de hipótesis para la media con varianza desconocida para muestras grandes
- ✓ Test de hipótesis para la media de una distribución normal con varianza desconocida para muestras pequeñas (t-Test)
- ✓ Test de hipótesis para la varianza

Test de hipótesis para la media con varianza desconocida para muestras grandes

Supongamos que la variable aleatoria de interés X tiene una **media μ y una varianza σ^2 desconocida**.

Asumimos que X tiene **distribución normal**, es decir $X \sim N(\mu, \sigma^2)$.

Hasta ahora se ha desarrollado el procedimiento de test de hipótesis para la hipótesis nula $H_0\!: \mu = \mu_0$ suponiendo una varianza σ^2 conocida, pero en la mayoría de las situaciones prácticas **σ^2 es desconocida**.

En general, si $n \geq 30$, entonces la varianza muestral S^2 está próxima a σ^2 en la mayor parte de las muestras, de modo que es posible sustituir S^2 por σ^2.

Es decir el estadístico de prueba $Z_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \approx N(0,1)$,

aproximadamente, si $n \geq 30$ y si $H_0\!: \mu = \mu_0$

Test de hipótesis para la media con varianza desconocida para muestras grandes

Además, **si no podemos decir que la muestra aleatoria proviene de una población normal**, sea la varianza σ^2 conocida o no, por el **Teorema Central del Límite**, los estadísticos de prueba:

$Z_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \approx N(0,1)$, aproximadamente, si $n \geq 30$ y si H_0: $\mu = \mu_0$

$Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$, aproximadamente, si $n \geq 30$ y si H_0: $\mu = \mu_0$

Las pruebas de hipótesis tendrán entonces un nivel de significancia ***aproximadamente de*** α.

Resumen

- ✓ Test de hipótesis para la media con varianza conocida
- ✓ Test de hipótesis para la media con varianza desconocida para muestras grandes
- ✓ Test de hipótesis para la media de una distribución normal con varianza desconocida para muestras pequeñas (t-Test)
- ✓ Test de hipótesis para la varianza

Test de hipótesis para la media con varianza desconocida para muestras pequeñas y distribución normal (*t*-Test)

Supongamos que la variable aleatoria de interés X tiene una **media $\boldsymbol{\mu}$ y una varianza $\boldsymbol{\sigma^2}$ desconocida**. Y el tamaño de la muestra es pequeño.

Cuando se prueban hipótesis sobre la **media $\boldsymbol{\mu}$ de una** población donde la **varianza $\boldsymbol{\sigma^2}$ es desconocida,** es posible utilizar los procedimientos de prueba dados anteriormente siempre y cuando el tamaño de la muestra sea grande ($n \geq 30$).

Estos procedimientos son aproximadamente válidos sin importar si la población de interés es normal o no.

Pero si la muestra es pequeña y $\boldsymbol{\sigma^2}$ **es desconocida** debe suponerse que la **distribución de la variable de interés es normal**.

Ejemplo *t*-Test

- En la *presentación 2.1* vimos un test de hipótesis para la media conociendo la varianza. Para continuar, utilizaremos los datos del consumo de automóviles vistos en la unidad 1 (mtcars). Consideraremos sólo los datos de autos de 6 cilindros, en Python:

```
# Conversión
cte = 1.60934 / 3.78541  # km/l
datosConvertidos = mtcars
datosConvertidos['mpg'] = 100 / (datosConvertidos['mpg'] * cte)  # Columna convertida
datospre = datosConvertidos[datosConvertidos['cyl'] == 6]['mpg']
datos = datospre.values
print("n=", len(datos))
```

Código/ejemplo_11.ipynb

- Nota: pusimos el consumo en Litros/100 km, porque estamos más acostumbrados a esa unidad.

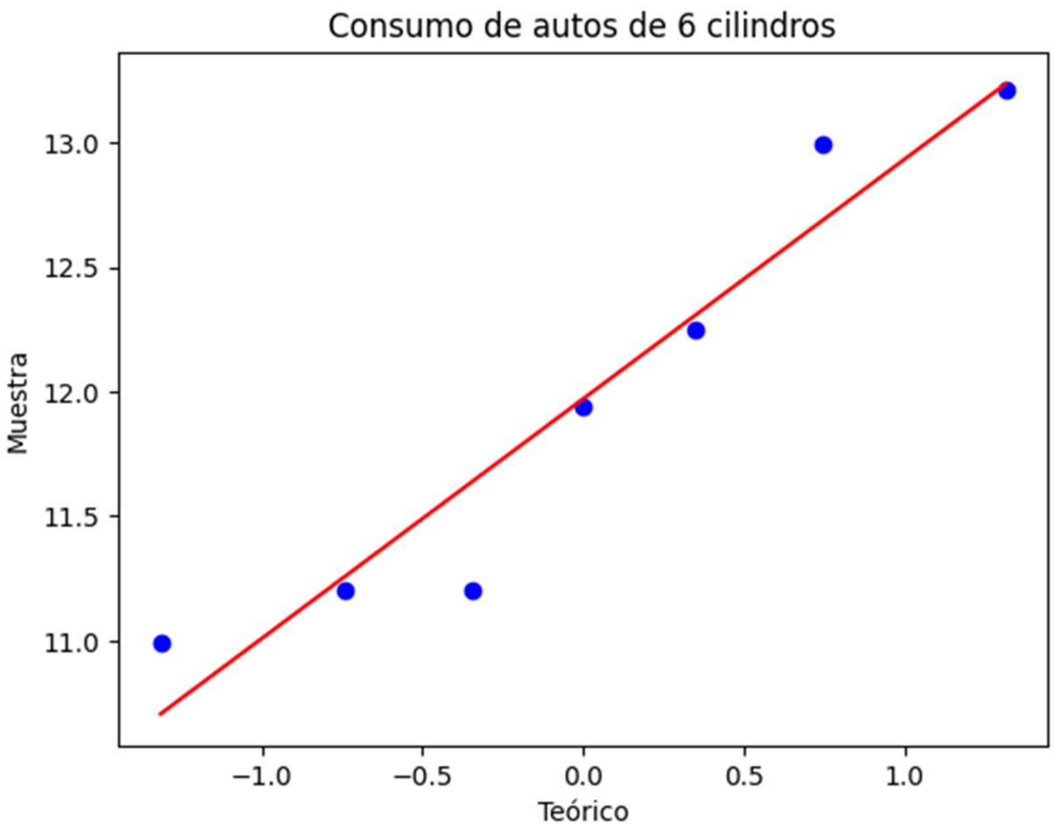
Ejemplo *t*-Test

► Se muestra el qqplot de los datos.

► Supongamos que no tenemos más información que ésta.

► Note que n es pequeño (n = 7) y $\bar{x}$ =11,97 Litros/100 km.

► Debemos entonces construir un Estadístico para desarrollar un test de hipótesis.



Construcción del Estadístico de prueba

- El único Estadístico que hemos visto hasta ahora es el Z_0[1] (Ec. (2) de la presentación 2.1). Conocíamos la distribución de ese Estadístico ($N(0,1)$) porque σ^2 era conocida.
- En este caso **se desconoce** σ^2 (la varianza poblacional). Entonces debemos estimarla con S^2 (la varianza muestral).

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad (1)$$

- Note que en la Ec. (1) se escribió todo con mayúscula, es decir, son todas variables aleatorias. Por lo tanto, si las X_i (con $i = 1, 2 \ldots, n$) son **independientes y están identicamente distribuídas** (iid) serán una muestra aleatoria y entonces S será un Estadístico también (lo usaremos más adelante).

[1] Como es muy utilizado se lo conoce como *Z*, y al test se le dice *Z* test.

Construcción del Estadístico de prueba

- Además note que n es **pequeño**.
- Continuando con el razonamiento, el Estadístico que se puede proponer es:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \qquad (2)$$

- Si las X_i tienen **distribución** $N\left(\mu_0, \sigma^2\right)$ la distribución del Estadístico se denomina t de Student (estandarizada, aunque se puede escalar)[2].

[2] Interesante historia sobre el descubrimiento de esta distribución, realizado por W. S. Gosset y profundizado por R. A. Fisher https://es.wikipedia.org/wiki/William_Sealy_Gosset

Construcción del Estadístico: revisión de t de Student

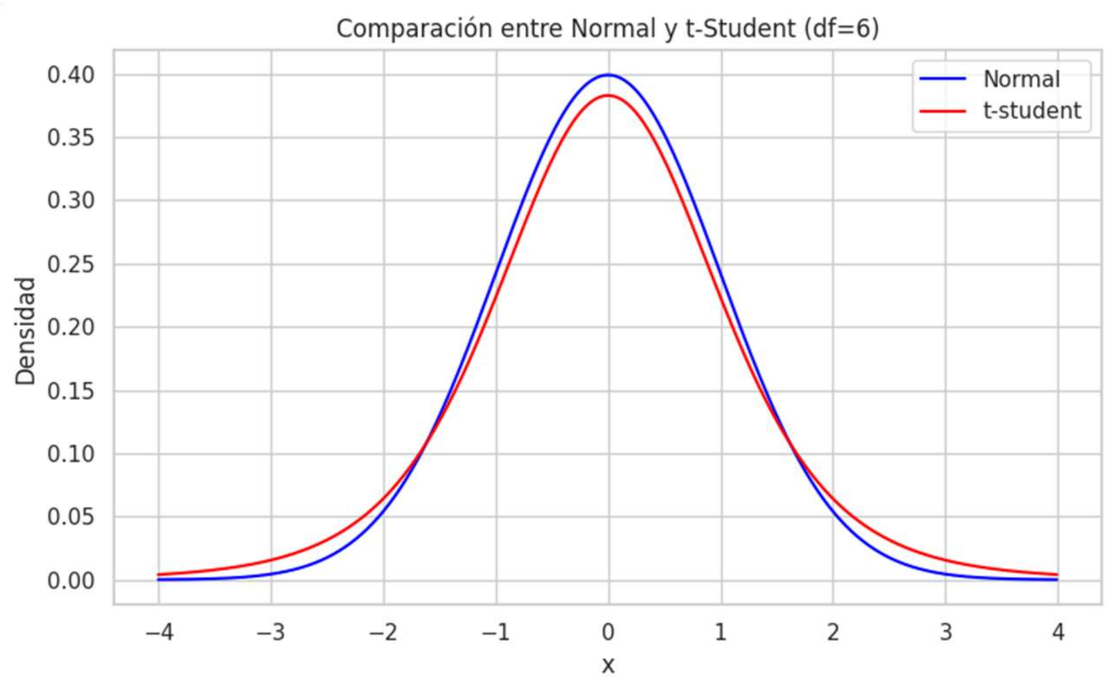
- En Python:

```
# Genero el vector x
x = np.arange(-4, 4, 0.01)
# fdp de la t de Student con 6 grados de libertad
y_fdp = t.pdf(x, df=6)
# Fda de la t de Student con 6 grados de libertad
y_fda = t.cdf(x, df=6)
```

Código/ejemplo_12.ipynb

- Note que sólo depende de los grados de libertad (df).

- Es una distribución simétrica y los valores superiores a 3 (en módulo) tienen todavía una probabilidad alta, porque se agrega aleatoriedad con la estimación de σ.



Planteo de hipótesis y cálculo de p-valor

- Supongamos que queremos plantear el siguiente test de hipótesis unilateral:

$$H_0 : \mu = \mu_0 = 11{,}7 \text{ Litros/100 km}$$

$$H_1 : \mu > \mu_0$$

Para tomar una decisión podríamos calcular el p-valor. Como es un test de una cola entonces debemos computar la siguiente probabilidad:

$$\text{p-valor} = P\left(\overline{X} > \overline{x} \mid H_0 \text{ verdadera}\right) = P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} > \frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right) = P\left(T > t_0\right)$$

donde $t_0 = \left(\overline{x} - \mu_0\right) / \left(s/\sqrt{n}\right)$, es el valor muestral del Estadístico, $\overline{x}$ y s, son los valores muestrales de la media y la desviación típica.

Planteo de hipótesis y cálculo de p-valor

► En Python se puede implementar de la siguiente forma:

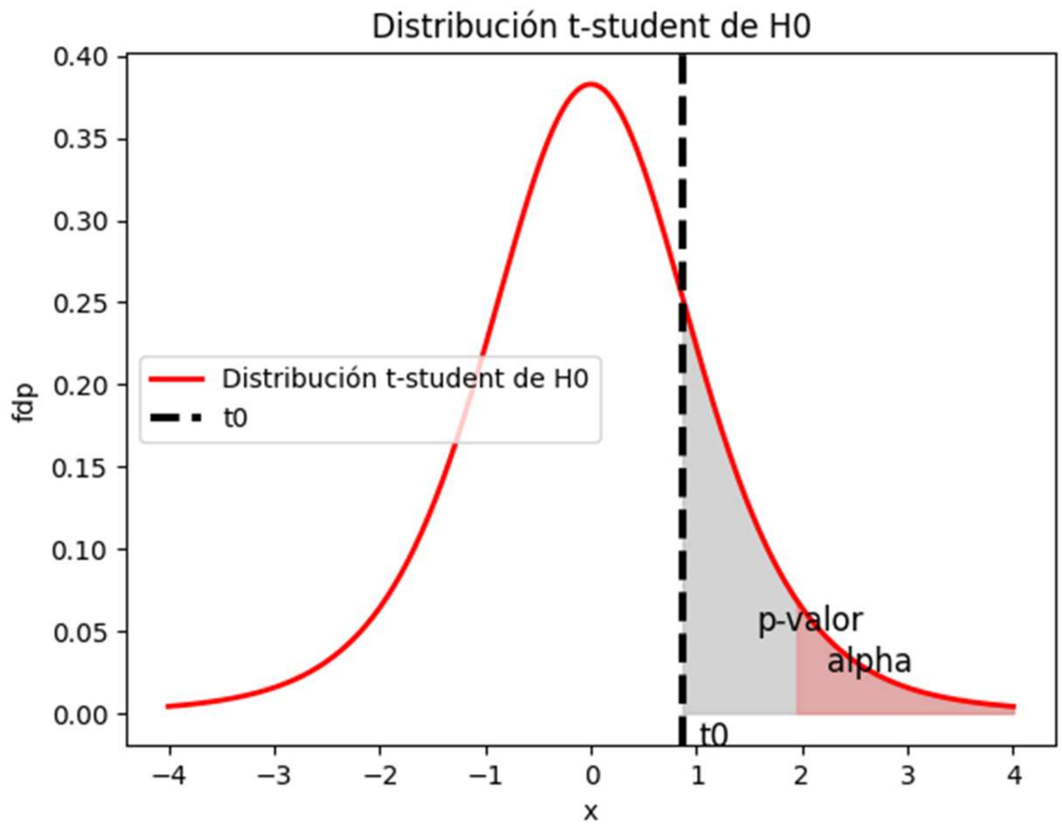
```
mu0 = 11.7  # Media de la H0

# Valor muestral del Estadístico
t0 = (np.mean(datos) - mu0) / (np.std(datos) / np.sqrt(len(datos)))
pvalor = 1 - stats.t.cdf(t0, df=len(datos) - 1)

print("Valor muestral del Estadístico:", t0)
print("p-valor:", pvalor)
print("Valor medio muestral:", np.mean(datos))
```

Código/ejemplo_11.ipynb

► El valor muestral del Estadístico es t_0= 0,862, cuyo p-valor = 0,211. Como el p-valor es muy grande (superior a una significancia de, por ejemplo, α = 0,05), esto significa que no existe evidencia para rechazar la Hipótesis nula.

Planteo de hipótesis y cálculo de p-valor

► Si se ejecuta el ejemplo 11 se obtendrá el siguiente gráfico.



► En rojo se muestra una significancia de α = 0,05.

► Python tiene una función que realiza el *t*-test:

```
# T-test
t_statistic, p_value = stats.ttest_1samp(datos, mu0)
print("Valor muestral del Estadístico:", t_statistic)
print("p-valor:", p_value)
```

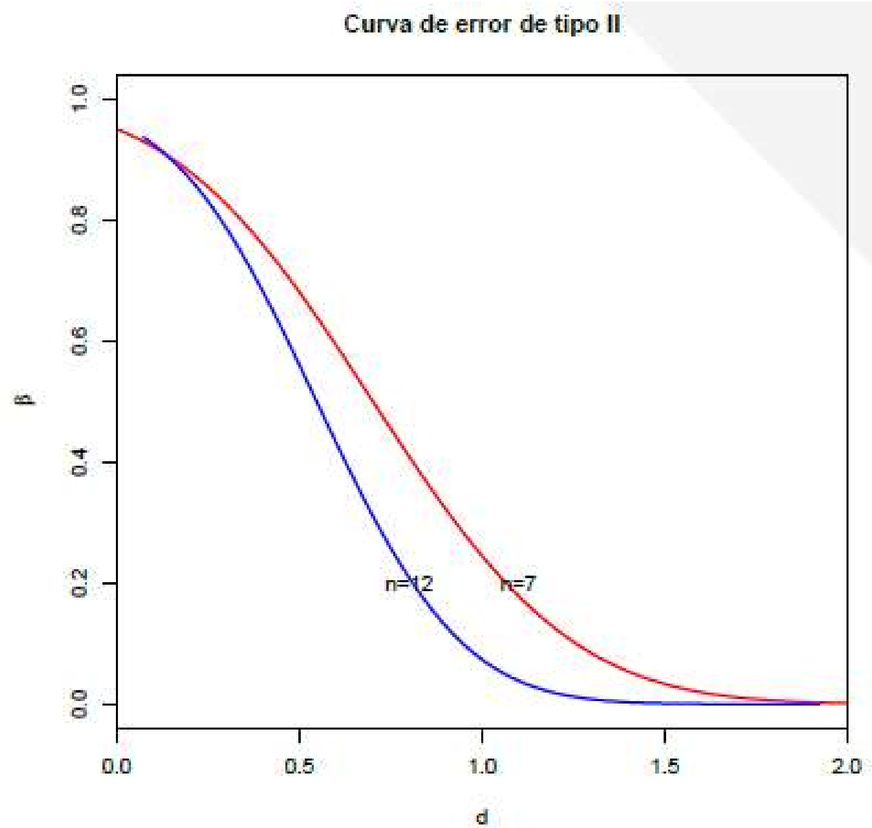
Código/ejemplo_11.ipynb

Comentarios

- ¿Cómo se calcula el error de tipo II? Por definición:

$$\beta = P\{\text{Acepto } H_0 |\ H_0 \text{ falsa}\} = P\{T_0 \leq t_{\alpha,n-1} \mid H_1 \text{ verdadera}\}. \qquad (3)$$

- El problema es que T_0 no tiene distribución t de student estandarizada, será una distribución t de student no central[a]. Si hacemos algo similar que la Ec. (4) de la presentación 2.1, obtendremos un parámetro de "no centralidad" $ncp = \frac{\delta\sqrt{n}}{\sigma}$.
- Lo podemos escribir en función del cociente $d = \delta/\sigma$. La figura muestra β en función de d para $n = 7$ y para $n = 12$ (ejemplo 13).
- Si uno quisiera definir un β y calcular el tamaño de muestra necesario, entonces tendría que utilizar este tipo de gráficos.
- Note que ncp depende de σ, deberíamos estimarlo con s.



[a] https://en.wikipedia.org/wiki/Noncentral_t-distribution

Comentarios

- Es importante hacer un gráfico qqplot con los datos para verificar que la distribución sea simétrica y que no difiera mucho de una distribución Normal (los puntos caen sobre la recta).
- Aunque la distribución no sea Normal, si es simétrica el t-test es robusto, y se logra tomar decisiones correctas.
- Si proviene de una distribución que no es Normal (y n es pequeño[3]) estrictamente no se puede aplicar el t-test. Se utilizan tests no paramétricos (más adelante trabajaremos con esto).

Test de hipótesis para la media con varianza desconocida para muestras pequeñas y distribución normal (*t*-Test)

Aquí vamos a formalizar lo visto en esta sección:

Supongamos que la variable aleatoria de interés X tiene **distribución normal** $X \sim N(\mu, \sigma^2)$ donde $\boldsymbol{\mu}$ **y** $\boldsymbol{\sigma^2}$ **son desconocidas**. Y el tamaño de la muestra es pequeño.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de la variable aleatoria X y sean $\boldsymbol{\bar{X}}$ **y** $\boldsymbol{S^2}$ **la media y la varianza muestrales**, respectivamente.

Podemos tomar como **estadístico de prueba** a $T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$, el cual, si la hipótesis nula es verdadera, tiene distribución **Student con n-1 grados de libertad**.

t-Test

Caso 1: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa bilateral** siguientes:

$$H_0: \mu = \mu_0 \quad \text{contra} \quad H_1: \mu \neq \mu_0$$

donde μ_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $|T_0| > t_{\frac{\alpha}{2},n-1}$
- Aceptar H_0 si $|T_0| \leq t_{\frac{\alpha}{2},n-1}$

La lógica sigue siendo la misma, si el estadístico de prueba toma un valor inusual, entonces se considera que hay evidencia en contra de H_0 y se rechaza la hipótesis nula. Como ahora la distribución del estadístico es Student, nos fijamos si T toma un valor t_0 en las colas de la distribución Student con $n-1$ grados de libertad.

t-Test

Caso 2: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0\colon \mu = \mu_0 \quad \text{contra} \quad H_1\colon \mu > \mu_0$$

entonces la **regla de decisión** es:

- Rechazar H_0 si $T_0 > t_{\alpha,n-1}$
- Aceptar H_0 si $T_0 \leq t_{\alpha,n-1}$

t-Test

Caso 3 Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0\colon \mu = \mu_0 \quad \text{contra} \quad H_1\colon \mu < \mu_0$$

entonces la **regla de decisión** es:

- Rechazar H_0 si $T_0 < -t_{\alpha,n-1}$
- Aceptar H_0 si $T_0 \geq -t_{\alpha,n-1}$

t-Test

Cálculo del p-valor:

Caso 1: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu \neq \mu_0$ entonces:

$$\boldsymbol{p{-}valor} = P(|T| > |T_0|) = 1 - P(|T| \leq |T_0|) = \mathbf{2}(\mathbf{1} - \boldsymbol{P}(\boldsymbol{T} \leq \boldsymbol{T_0}))$$

Caso 2: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu > \mu_0$ entonces:

$$\boldsymbol{p{-}valor} = P(T > T_0) = 1 - P(T \leq T_0)$$

Caso 3: Si las hipótesis son $H_0\colon \mu = \mu_0$ contra $H_1\colon \mu < \mu_0$ entonces:

$$\boldsymbol{p{-}valor} = P(T \leq T_0)$$

Resumen

- ✓ Test de hipótesis para la media con varianza conocida
- ✓ Test de hipótesis para la media con varianza desconocida para muestras grandes
- ✓ Test de hipótesis para la media de una distribución normal con varianza desconocida para muestras pequeñas (t-Test)
- ✓ Test de hipótesis para la varianza

Construcción del Estadístico de prueba

- Intentaremos desarrollar un test de hipótesis para tomar una decisión sobre la varianza.
- Un punto de partida puede ser la Ec. (1):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{4}$$

- Supondremos que las X_i provienen de una **muestra aleatoria** (iid) y que la **distrución es Normal**.

Construcción del Estadístico de prueba

- Si conocemos el valor de σ_0 entonces el siguiente Estadístico:

$$X_0^2 = \frac{(n-1)\,S^2}{\sigma_0^2} \tag{5}$$

posee distribución χ^2_{n-1}, se lee Chi-cuadrado (o Ji-cuadrado[4]) con $n-1$ grados de libertad.

- Ya vimos en la unidad 1 las funciones en R que se pueden utilizar para cálculos con esta distribución.

[4] El nombre y el descubrimiento se le adjudican a Karl Pearson aunque fue Friedrich R. Helmert quien la descubrió primero.

Planteo de hipótesis

- Supongamos que queremos plantear el siguiente test de hipótesis bilateral:

$$H_0 : \sigma = \sigma_0$$

$$H_1 : \sigma \neq \sigma_0$$

- Elegimos una significancia $\alpha = 0,05$ y calculamos los valores críticos:

$$\alpha = P\{\text{Rechazo } H_0 |\ H_0 \text{ verdadera}\} =$$
$$= P\{\frac{(n-1)\,S^2}{\sigma_0^2} < x^2_{\alpha/2,n-1}\} + P\{\frac{(n-1)\,S^2}{\sigma_0^2} > x^2_{1-\alpha/2,n-1}\} \qquad (6)$$

- En el ejemplo que veníamos siguiendo, teníamos $n = 7$ datos, por lo tanto los valores críticos serán:

$$x^2_{0,05/2,\ 7-1} = 1,237 \text{ y } x^2_{0,95/2,\ 7-1} = 14,449.$$

- Entonces si el valor muestral cae fuera del intervalo $[1,237,\ 14,449]$ existe evidencia para rechazar la Hipótesis nula.

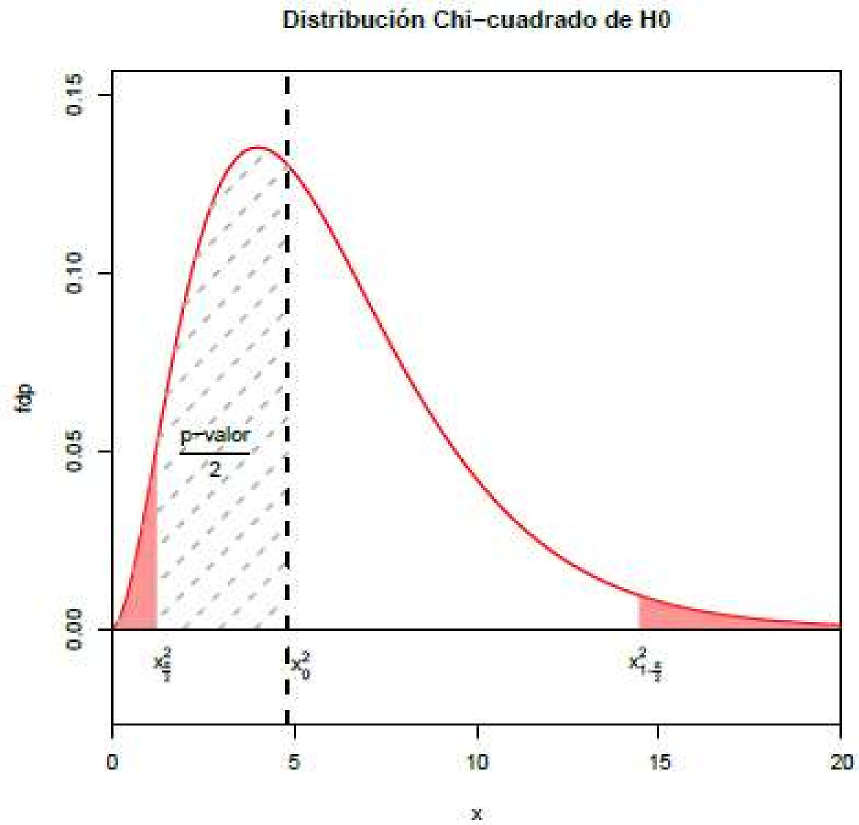
Planteo de hipótesis y toma de decisión

- Si planteamos $\sigma_0^2 = 1{,}0$, para nuestro ejemplo el valor muestral del Estadístico es:

$$x_0^2 = \frac{(7-1)\,s^2}{\sigma_0^2} = 4,821$$

  y cae dentro del intervalo calculado anteriormente. Entonces no existe evidencia para descartar la H_0.

- Para el p-valor se deben calcular dos probabilidades:
  $P\{X_0^2 < x_0^2\} = 0,433$
  $P\{X_0^2 > x_0^2\} = 0{,}866.$



El p-valor será: $= 2 \min\left(P\{X_0^2 < x_0^2\}, P\{X_0^2 > x_0^2\}\right) = 0,865.$
Se concluye lo mismo, el p-valor es muy alto, no hay evidencia para rechazar H_0[5].

[5]Todo está implementado en el ejemplo 14.

Comentarios

- En los cálculos anteriores debe notarse que no es la única manera de generar regiones de aceptación y rechazo en tests bilaterales. Se podría por ejemplo, dejar una región a derecha de $\alpha/3$ y a izquierda $2\alpha/3$, sería también un test de significancia α. Se puede demostrar que dejando $\alpha/2$ a ambos lados se obtiene el intervalo más pequeño, en cierta forma el mejor.
- Podemos hacer el mismo análisis para el cálculo del tamaño de la muestra n para obtener un determinado β, aunque ahora debemos parametrizarlo con σ_0^2/σ^2, donde σ^2 es la varianza verdadera.
- Utilizaremos la distribución χ^2_{n-1} para otros tests más adelante.

Test de hipótesis para la varianza

Aquí vamos a formalizar lo visto en esta sección:

Supongamos que se desea probar la hipótesis de que la **varianza de una población normal** es igual a un valor específico, por ejemplo $\boldsymbol{\sigma_0^2}$.

Sea $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria de tamaño n de una v.a. X, donde $X \sim N(\mu, \sigma^2)$.

Tomamos como **estimador puntual** de σ^2 a $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Luego a partir de este estimador puntual construimos el **estadístico de prueba** $X = \frac{(n-1)S^2}{\sigma^2}$

Este estadístico de prueba contiene al parámetro desconocido a estimar σ^2 y ya sabemos que tiene una distribución llamada ***chi-cuadrado con n-1 grados de libertad***

Test de hipótesis para la varianza

Caso 1: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa bilateral** siguientes:

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{contra} \quad H_1: \sigma^2 \neq \sigma_0^2$$

donde σ_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $X_0 > \chi^2_{\frac{\alpha}{2},n-1}$ o $X_0 < \chi^2_{1-\frac{\alpha}{2},n-1}$
- Aceptar H_0 si $\chi^2_{1-\frac{\alpha}{2},n-1} \leq X_0 \leq \chi^2_{\frac{\alpha}{2},n-1}$

donde $X_0 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

Nuevamente, el razonamiento es, si el estadístico de prueba toma un valor inusual, entonces se considera que hay evidencia en contra de H_0 y se rechaza la hipótesis nula. Recordar que la distribución χ^2_{n-1} es asimétrica.

Test de hipótesis para la varianza

Caso 2: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{contra} \quad H_1: \sigma^2 > \sigma_0^2$$

donde σ_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $X_0 > \chi^2_{\alpha,n-1}$
- Aceptar H_0 si $X_0 \leq \chi^2_{\alpha,n-1}$

donde $X_0 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

Test de hipótesis para la varianza

Caso 3: Supongamos que planteamos la **hipótesis nula** y la **hipótesis alternativa unilateral** siguientes:

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{contra} \quad H_1: \sigma^2 < \sigma_0^2$$

donde σ_0 es una constante específica.

La **regla de decisión** es:

- Rechazar H_0 si $X_0 < \chi^2_{1-\alpha,n-1}$
- Aceptar H_0 si $X_0 \geq \chi^2_{1-\alpha,n-1}$

donde $X_0 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

Test de hipótesis para la varianza

Cálculo del p-valor:

Caso 1: Si las hipótesis son $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 \neq \sigma_0^2$ entonces:

$$\boldsymbol{p-valor} = 2\min(P(X < X_0), P(X > X_0))$$

Caso 2: Si las hipótesis son $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 > \sigma_0^2$ entonces:

$$\boldsymbol{p-valor} = P(X > X_0)$$

Caso 3: Si las hipótesis son $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 < \sigma_0^2$ entonces:

$$\boldsymbol{p-valor} = P(X < X_0)$$

Conclusiones

- Hemos aprendido algunos comandos y herramientas para estadística descriptiva y test de hipótesis[6].
- Hemos repasado algunos tests de hipótesis para una muestra, es decir, siempre utilizamos **una muestra aleatoria** iid. Más adelante aplicaremos tests similares para dos muestras.
- Vimos tests para la media y la varianza.
- Para la media, podemos concluir algunas cosas:
  - No importa cual sea n, si X_i tiene distribución Normal y conocemos la varianza σ^2 entonces utilizamos el Estadístico Z.
  - Si n es grande $(n > 30)$ podríamos utilizar el Teorema Central del Límite y se puede utilizar también Z, estimando σ^2 con s^2.
  - Cuando n es pequeño X_i tiene distribución Normal entonces el Estadístico que utilizamos es T. Si no queremos imponer normalidad, entones debemos aplicar tests no paramétricos, es decir, que no se suponga ninguna distribución (tema que veremos más adelante).

Bibliografía

- D. C. Montgomery and G. C. Runger, “Applied Statistics and Probability for Engineers Third Edition”, John Wiley & Sons, Inc.

- Peter Dalgaard, “Introductory Statistics with R”, Second Edition, Springer.