Inferencia Estadística

Test para muestras normales y test asintóticos

San Andres

June 11, 2022

Test para la media de una población

Caso normal con varianza conocida

Ejemplo (estudiantes)

Sea X el peso medido en kilograms de un estudiante hombre elegido al azar en una cierta universidad. Se quiere testear la hipótesis nula de que el peso medio de estudiantes hombres es 68 kilogramos, contra la alternativa de que es mayor a 68 kilogramos.

1. Plantear un test de nivel $\alpha = 0.05$ para una muestra de tamaño $n = 36$. Asumir que el peso de un estudiante hombre elegido al azar sigue una distribución normal con varianza conocida $\sigma^2 = 100$.
2. Calcule la potencia del test cuando el peso medio es en realidad $\mu = 70$.
3. Calcule la potencia del test cuando el peso medio es en realidad $\mu = 72$.

Caso normal con varianza conocida

Ejemplo (estudiantes)

Supongamos que la media muestral resultó $\overline{x}_{\text{obs}} = 71.23$.

1. ¿Que decisión se toma en el test?
2. ¿Cuál es el p-valor de esta muestra?
3. ¿Rechazaría el test a nivel $\alpha = 0.1$? ¿Y a nivel $\alpha = 0.02$?

Caso normal con varianza conocida

Ejemplo (estudiantes)

Repetir lo anterior pero si ahora se quiere testear

$$H_0 : \mu = 68 \quad \text{vs} \quad H_1 : \mu \neq 68$$

Caso normal con varianza conocida

Problema

Dada una muestra aleatoria

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

con σ conocida queremos testear

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \begin{cases} \mu & > \mu_0 \ \textit{(igual contra mayor)} \\ \mu & \neq \mu_0 \ \textit{(igual contra distinto)} \\ \mu & < \mu_0 \ \textit{(igual contra menor)} \end{cases}$$

Caso normal con varianza conocida

Estadístico

Consideramos el estadístico

$$Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Su distribución bajo H_0 resulta

$$Z \sim \mathcal{N}(0, 1)$$

Caso normal con varianza conocida

Regiones de rechazo

Por lo tanto obtenemos un test de nivel α con las siguientes regiones

Alternativa	Región de rechazo
$H_1 : \mu > \mu_0$	$Z \geq z_{1-\alpha}$
$H_1 : \mu \neq \mu_0$	$\lvert Z\rvert \geq z_{1-\frac{\alpha}{2}}$
$H_1 : \mu < \mu_0$	$Z \leq -z_{1-\alpha}$

Caso normal con varianza conocida

Distribución del estadístico bajo la alternativa

También conocemos la distribución del estadístico cuando H_0 no se cumple,

$$Z \sim \mathcal{N}\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$$

Dem: Como $X_i \sim \mathcal{N}(\mu, \sigma^2)$

$$
\begin{aligned}
Z &= \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \\
&= \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \\
&= \underbrace{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}_{\sim\mathcal{N}(0,1)} + \underbrace{\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}}_{\text{constante}}
\end{aligned}
$$

Caso normal con varianza conocida

La potencia del test es en realidad una función de cada posible valor de la alternativa. Se pueden deducir para el caso de normales con varianza conocida pero no las vamos a ver en este curso.
Se puede ver que la función de potencia

- aumenta cuando mayor es la diferencia entre μ y μ_0.
- aumenta cuando σ disminuye
- aumenta cuando n aumenta
- aumenta cuando el nivel α aumenta

Caso normal con varianza conocida

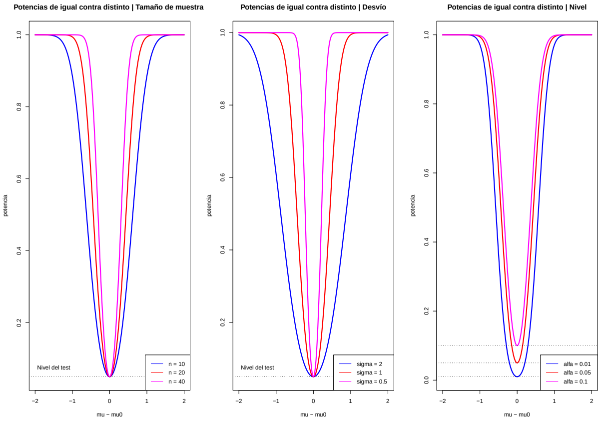


Figure: Funciones de potencia $\pi(\mu - \mu_0)$.

Caso normal con varianza conocida - Hipótesis nula compuesta

Problema

Dada una muestra aleatoria

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

con σ^2 conocida queremos testear

$$H_0 : \mu \leq \mu_0 \quad vs \quad \mu > \mu_0$$

Caso normal con varianza conocida

Proponemos un test de la forma

$$\text{rechazo } H_0 \Leftrightarrow Z = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq k$$

con k que verifica

$$\alpha = \max_{\mu \leq \mu_0} P_\mu(\text{rechazar } H_0) = \max_{\mu \leq \mu_0} P_\mu(Z \geq k)$$

No es difícil ver que

$$\max_{\mu \leq \mu_0} P_\mu(Z \geq k) = P_{\mu_0}(Z \geq k).$$

Por lo tanto es equivalente a realizar un test de la forma

$$H_0 : \mu_0 = \quad \text{vs} \quad \mu > \mu_0$$

Caso normal con varianza conocida - Hipótesis nula compuesta

Test menor o igual contra mayor

- Queremos testear las hipótesis:

$$H_0 : \mu \leq \mu_0 \qquad \text{vs} \qquad H_1 : \mu > \mu_0$$

- ¿Qué estadístico usamos?

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Región de rechazo $R = \{Z \geq z_{1-\alpha}\}$
- P-valor: $P_{\mu_0}(Z \geq Z_{obs})$.
- La función de potencia es la misma que la del test igual contra mayor.

Caso normal con varianza conocida - Hipótesis nula compuesta

Test mayor o igual contra menor

- Queremos testear las hipótesis:

$$H_0 : \mu \geq \mu_0 \qquad \text{vs} \qquad H_1 : \mu < \mu_0$$

- ¿Qué estadístico usamos?

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Región de rechazo: $R = \{Z \leq -z_{1-\alpha}\}$.
- P-valor: $= P_{\mu_0}(Z \leq Z_{obs})$.
- La función de potencia es la misma que la del test igual contra menor.

Caso normal con varianza conocida - Tamaño de muestra

Ejemplo (pintura)

El tiempo de secado de las pinturas tradicionales se distribuye normalmente condesvío estándar 9 min. Todas las pinturas del mercado tienen un tiempo medio de secado de 75 minutos o más

Se diseño una aditivo químico para disminuir el tiempo medio de secado de la pintura a menos de 75 minutos. Se cree que el tiempo de secado con este aditivo también se distribuye de manera normal con $\sigma = 9$. Se decide recolectar una muestra de $X_1, \ldots X_n$ tiempos de secados independientes.

Sea μ el tiempo de secado medio cuando el aditivo es usado. La hipótesis son

$$H_0: \quad \mu \geq 75$$
$$H_1: \quad \mu < 75$$

1. Plantear el estadístico del test y la región de rechazo para tener un nivel $\alpha = 0.05$.
2. Si se toma una muestra de $n = 25$ tiempos de secado. Calcular la potencia del test si el tiempo medio de secado es de 73 minutos.
3. ¿Que tamaño de muestra debe tomarse si se quiere que la probabilidad de que el test logre rechazar la hipótesis nula con probabilidad al menos 0.99 en el caso que el tiempo medio verdadero para el secado con el aditivo sea de 73 minutos?

Caso normal con varianza desconocida

Ejemplo (`BodyTemp50`)

¿Cuál es la temperatura promedio para personas sanas? La mayoría de las personas usando la escala Fahrenheit dirán que es 98.6°F (37°C).

Esto ha sido bien establecido durante muchos anõs, pero es posible que el promedio de la temperatura del cuerpo haya cambiado a lo largo del tiempo?

Tal vez la gente se está volviendo cada vez más "tibios" debido al calentamiento global del medio ambiente, o tiene temperaturas más bajas que en siglos pasados.

Allen Shoemaker presentó datos derivados de un estudio de adultos sanos los cuales contienen la temperatura del cuerpo, género y pulso por cada sujeto.

Los datos para $n = 50$ se encuentran en **BodyTemp50** dentro del paquete **Lock5Data**.

Caso normal con varianza desconocida

Ejemplo (`BodyTemp50`)

- Las hipótesis a testear son:

$$H_0 : \mu = 98.6(F) \qquad H_1 : \mu \neq 98.6(F).$$

- Las temperaturas registradas las asumimos independientes y distribuidas normalente. Sin embargo no conocemos la vairanza σ^2. Podemos estimarla con

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

- Si consideramos el estadístico

$$T = \frac{\overline{X} - \mu_0}{S_n/\sqrt{n}}$$

¿Será correcto utilizar una región de Rechazo $\mathcal{RR} = \{|T| > z_{1-\alpha/2}\}$ para obtener un test de nivel α?

Caso normal con varianza desconocida

Distribución del estadístico T

Si $X_1, \ldots, X_n$ es una muestra aleatoria con distribución $\mathcal{N}(\mu, \sigma^2)$ entonces el estadístico

$$T = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}$$

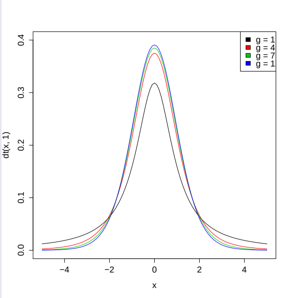
Sigue una distribución *t de Student* con $n-1$ grados de libertad que notamos t_{n-1}

Caso normal con varianza desconocida

Definición

Decimos que distribución t de student con ν grados de libertad tiene función de densidad

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} .$$



Caso normal con varianza desconocida

Estadístico

Consideramos el estadístico

$$T = \frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}$$

Por lo tanto obtenemos un test de nivel α con las siguientes regiones

Nula	Alternativa	Región de rechazo
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$Z \geq t_{n-1,\alpha}$
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$\lvert Z \rvert \geq t_{n-1,\frac{\alpha}{2}}$
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$Z \leq -t_{n-1,\alpha}$

Caso normal con varianza desconocida

Ejemplo (BodyTemp50)

La media muestral de los datos fue $\bar{x}_{\text{obs}} = 98.26$ y el $s_{\text{obs}} = 0.7653$.

- Determine la región de rechazo para el test de nivel $\alpha = 0.01$.
- ¿Qué decisión toma con los datos observados?
- Calcular el p-valor para la muestra observada.
- ¿Es posible determinar la potencia para algún valor de $\mu \neq 98.6$?

Caso normal con varianza desconocida

Ejemplo (BodyTemp50)

- Como $2.68 = qt(0.995, df = 49)$ resulta

$$\mathcal{RR} = \{|T| > 2.68\}$$

- Tenemos que

$$|T_{\text{obs}}| = \left| \frac{98.26 - 98.6}{\frac{0.7653}{\sqrt{50}}} \right| = 3.141.$$

  Como $T \in \mathcal{RR}$, rechazamos la hipótesis nula a favor de la alternativa.

- p-valor: $P_{H_0}(|T| > 3.141) = 0.0028.$
- Para determinar la potencia debemos calcular $P_\mu(|T| > 2.68)$ el problema es que no conocemos la distribución de T bajo la alternativa.

Ejemplo: temperatura corporal normal

```
> t.test(BodyTemp, mu = 98.6, alternative = "two.sided")

One Sample t-test

data:  BodyTemp
t = -3.1414, df = 49, p-value = 0.002851
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
98.0425 98.4775
sample estimates:
mean of x
98.26
```

Test de hipótesis para muestras grandes
(niveles de significación asintóticos)

Test asintóticos

¿Y si no tenemos normalidad de los datos?

Cuando no tenemos la normalidad de los datos pero tenemos un estimador asintóticamente normal podemos utilizar la distribución asintótica para construir un test de hipótesis.

Nivel de significación asintótico

Definition

Una sucesión de tests de nivel α_n tiene **nivel de significación asintótico** α si $\alpha_n \to \alpha$ cuando $n \to \infty$.

Vamos a ver como construir tests de nivel asíntotico para una media, sin asumir normalidad ni varianzas conocidas, o para una proporción, basandonos en el Teorema Central del Límite.

Test asintótico para la media

Test asintótico para la media

Sea $X_1, \ldots, X_n$ una muestra aleatoria con media μ y varianza σ^2 ambas desconocidas

- Queremos testear las hipótesis:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0; \quad (\mu > \mu_0 \quad \text{o} \quad \mu < \mu_0).$$

- ¿Qué estadístico usamos?

$$Z = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}.$$

- ¿Cuál es la distribución asintótica del estadístico cuando H_0 es cierta? Por el TCL y Slutzky,

$$Z \overset{a}{\sim} \mathcal{N}(0, 1)$$

- Rechazamos H_0

Alternativa	Región de rechazo
$H_1 : \mu > \mu_0$	$Z \geq z_{1-\alpha}$
$H_1 : \mu \neq \mu_0$	$\lvert Z \rvert \geq z_{1-\frac{\alpha}{2}}$
$H_1 : \mu < \mu_0$	$Z \leq -z_{1-\alpha}$

Test asintótico para la media

Ejemplo (máquina de café)

El volumen de café producido con cada uso de cierta máquina de café es una variable aleatoria con media $\mu \in \mathbb{R}$. La máquina se considera bien calibrada si el volumen esperado es de 260 mL, es decir si $\mu = 260$.

Para determinar si la máquina necesita recalibrarse, se plantea el contraste de hipótesis

$$H_0 : \ \mu = 260$$
$$H_1 : \ \mu \neq 260$$

Los resultados de una muestra aleatoria de los volúmenes (mL) de café producido en 1200 usos de la máquina son $\overline{x} = 257.19, s^2 = 442.36$.

Realizar un test con un nivel de significación asintótico de $\alpha = 0.05$.

Test asintótico para la media

Ejemplo (máquina de café

Para contruir el test asintótico, nos basamos en el estadístico

$$Z = \frac{\overline{X}_n - 260}{S/\sqrt{n}}$$

Definimos la región de rechazo como

$$\mathcal{R} = \left\{Z \leq -z_{1-\alpha/2}\right\} \cup \left\{Z \geq z_{1-\alpha/2}\right\}$$

Como vimos, este test tiene el nivel de significación asintótico deseado porque, suponiendo que H_0 es verdadera, $\mu = 260$. Luego, por TCL y Slutzky

$$Z \xrightarrow{d} \mathcal{N}(0,1)$$

Lo que implica que $P(Z \in \mathcal{R}) \to \alpha$ si H_0 es cierta.

Test asintótico para la media

Ejemplo (máquina de café)

A partir de los datos de la muestra:

$$z = \frac{257.19 - 260}{\sqrt{442.36/1200}} \approx -4.63$$

$-z_{1-\alpha/2} = -z_{0.975} \approx -1.96$. Como $-4.63 \leq -1.96$, el estadístico está en la región de rechazo: rechazamos H_0. Es decir, hay suficiente evidencia de que la máquina de café necesita recalibrarse.

Test asintótico para una proporción

Proporción

Sea $X_1, \ldots, X_n$ una muestra aleatoria $X_i \sim \mathcal{B}\rceil(p)$

- Queremos testear las hipótesis:

$$\begin{aligned} H_0&: \; p = p_0 \\ H_1&: \; p \neq p_0 \; (p < p_0, \; p > p_0) \end{aligned}$$

- ¿Qué estadístico usamos? $Z = \left(\overline{X} - p_0\right) / \sqrt{p_0(1-p_0)/n}$.
- ¿Qué distribución tiene el estadístico cuando H_0 es cierta? Por el TCL,

$$Z = (\overline{X} - p_0)/\sqrt{p_0(1-p_0)/n} \stackrel{a}{\sim} \mathcal{N}(0,1)$$

- Rechazamos H_0 si

Alternativa	Región de rechazo
$H_1 : p > p_0$	$z \geq z_{1-\alpha}$
$H_1 : p < p_0$	$z \leq -z_{1-\alpha}$
$H_1 : p \neq p_0$	$\lvert z \rvert \geq z_{1-\alpha/2}$

Test asintótico para una proporción

Ejemplo (hamburguesas veganas)

En un cierto mercado, estudios mostraron que la proporción de consumidores de hamburguesas que consideran a la hamburguesa vegana como parte de su dieta se mantuvo estable en los últimos años, en 17%. Para intentar aumentar esta poporción, una firma llevó a cabo una campaña publicitaria por unos meses.

Tras la campaña, se realizó un nuevo estudio con una muestra aleatoria representativa del mercado de consumidores de hamburguesas y 843 encuestados. De estos, 153 respondieron que consideran a la hamburguesa vegana como parte de su dieta.

¿Sugieren los resultados que la proporción de consumidores de este mercado que incorporan a la hamburguesa vegana en su dieta es ahora mayor a 17%? Para responder, realizar un test de hipótesis asintótico de nivel $\alpha = 0.1$.

Test asintótico para una proporción

Ejemplo (hamburguesas veganas)

Si llamamos p a la proporción poblacional de consumidores de hamburguesas veganas, las hipótesis a contrastar son:

$$H_0: \; p = 0.17$$

$$H_1: \; p > 0.17$$

Sea $\overline{X}_n$ la proporción de consumidores de hamburguesas veganas en la muestra. Armamos el test a partir del estadístico

$$Z = \frac{\overline{X}_n - 0.17}{\sqrt{\frac{0.17(1-0.17)}{n}}}$$

Bajo H_0 es verdadera, como el tamaño de muestra es grande $Z \approx \mathcal{N}(0,1)$.
Región de rechazo para nivel α

$$\mathcal{R} = \{Z \geq z_{1-\alpha}\}$$

Test asintótico para una proporción

Ejemplo (hamburguesas veganas)

A partir de los datos de la muestra:

$$z = \frac{153/843 - 0.17}{\sqrt{0.17(1-0.17)/843}} \approx 0.89$$

$z_{1-\alpha} = z_{0.9} \approx 1.28$. Como $0.89 < 1.28$, el estadístico está fuera de la región de rechazo: no rechazamos H_0. Es decir, no hay suficiente evidencia para afirmar que la proporción de consumidores de hamburguesas veganas aumentó.

Test asintóticos

Test de Wald

Sea $X_1, \ldots, X_n$ una muestra con distribución F_θ.
Se quiere testear

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Si $\widehat{\theta}$ es asimptóticamente normal entonces bajo H_0

$$W = \frac{(\widehat{\theta} - \theta_0)}{\widehat{se}(\hat{\theta})} \approx \mathcal{N}(0, 1).$$

Region de rechazo: $R = \{|W| \geq z_{1-\alpha/2}\}$.

Test asintóticos

Comparación de algorítmos de predicción

Se testean dos algoritmos de clasificación en datasets de tamaño n y m independientes. Sea X la cantidad de predicciones incorrectas del algoritmo 1 e Y la cantidad de predicciones del algoritmo 2.

$$X \sim \mathcal{B}(n, p_1), \quad Y \sim \mathcal{B}(m, p_2).$$

Queremos testear $p_1 = p_2$. Llamamos $\delta = p_1 - p_2$

$$H_0 : \delta = 0, \quad \text{vs} \quad H_1 : \delta \neq 0.$$

Comparación de algoritmos de predicción

- Hallar el estadístico y la región de rechazo para tener un nivel aproximado $\alpha = 0.05$.
- Calcular la potencia si $n = 200$, $p_1 = 0.3$, $m = 250$ y $p_2 = 0.35$.
- Si para los mismos valores de n y m se obtuvieron $X_{0bs} = 62$ e $Y_{obs} = 89$. ¿Qué decisión se toma? Calcular el p-valor de la muestra. ¿Se rechaza a nivel $\alpha = 0.1$?

Test asintóticos

Comparación de medias

Sean $X_1, \ldots, X_n$ e $Y_1, \ldots Y_m$ muestras independientes de pobalciones con medias μ_x y μ_y. Sea la hipótesis nula $\mu_x = \mu_y$ es decir $\delta = \mu_x - \mu_y$

$$H_0 : \delta = 0 \quad \text{vs} \quad \delta \neq 0$$

Podemos estimar $\widehat{\delta} = \overline{X}_n - \overline{Y}_m$ y el error estandard estimado.

$$\widehat{se}(\hat{\delta}) = \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}.$$

Luego rechazamos si $|W| \geq z_{1-\alpha/2}$ con

$$W = \frac{\widehat{\delta} - 0}{\widehat{se}(\hat{\delta})} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}.$$

Test asintóticos

Temperatura corporal

Para los datos de temperatura corporal realizar un test para decidir a nivel aproximado $\alpha = 0.1$ si la temperatura media de los hombres es distínta de la temperatura media de las mujeres.
Obtener el p-valor de la muestra.

Test asintóticos - Muestras apareadas

Dieta

Se quiere ver si una dieta particular permite disminuir la presión arterial en pacientes hipertensos. Para esto se junta un grupo de n pacientes a los que se les mide la presión antes de comenzar la dieta y otra vez al concluirla. Tenemos paraes de observaciones

$$(A_1, D_1), (A_2, D_2), \ldots, (A_n, D_n).$$

Donde A_i y D_i son la presión del paciente i antes y despues de realizar la dieta. Si llamamos δ al efecto de la dieta sobre la presión de los pacientes, i.e $\delta = E(A_i - D_i)$. Plantear un test de hipótesis de nivel $\alpha = 0.05$ para decidir si la dieta disminuye la presión arterial. Escribir las hipótesis, el estadístico a utilizar, la distribución del mismo bajo H_0 y encontrar la región de rechazo.