Inferencia: Contraste de hipótesis

Grado en Ingeniería Biomédica

Clase 3

Jorge Calero Sanz

Departamento de Teoría de la Señal y Comunicaciones
Universidad Rey Juan Carlos

13 de marzo de 2023

Introducción

Dos poblaciones

- Hasta ahora hemos tratado problemas que involucraban una sola población.
- En un contraste de hipótesis de dos poblaciones, los parámetros de dichas poblaciones serán comparados, donde ninguno de éstos se asume que es conocido.
- Ejemplo: Queremos estudiar la relación entre el uso de un medicamento suministrado por vía oral, y el nivel de presión sanguínea.
- ¿Cómo podemos diseñar el estudio?

Introducción

Dos poblaciones

- En general, tendremos dos maneras distintas de proceder:
- **Estudio longitudinal** o de seguimiento: Elegimos un grupo de personas con una serie de características (por ejemplo, un rango de edad) que no estén tomando dicho medicamento.
  - A este grupo le medimos la presión, y a esta medida la llamamos **base**.
  - A continuación, suministramos a todo el grupo el medicamento durante un periodo de tiempo suficientemente grande (1 año, por ejemplo).
  - Pasado ese lapso de tiempo, se vuelve a medir la presión y se comparan.

Introducción

Dos poblaciones

- **Estudio transversal o de muestras independientes:** Volvemos a elegir un grupo de personas con una serie de características (por ejemplo, un rango de edad).
  - Separamos el grupo en 2: los que toman el medicamento y los que no.
  - Se toman las medidas y se comparan

Introducción

Dos poblaciones

- El primer estudio está hecho en base a unas **muestras pareadas**.
- Se dice que dos muestras son pareadas cuando cada dato de la primera muestra se relaciona con un único dato de la segunda muestra.
- En el primer ejemplo, está claro que cada persona de la muestra es un dato en la primera muestra (la **base**) y en la segunda (el **seguimiento**).
- El segundo estudio está hecho en base a unas **muestras independientes**.
- En el segundo ejemplo, se están comparando dos grupos totalmente diferentes.

Introducción

Dos poblaciones

- Dependiendo de la situación, puede ser más conveniente usar un tipo de estudio u otro.
- En nuestro ejemplo, es más recomendable utilizar el de muestras pareadas, ya que puede haber muchos otros factores externos que afecten a la presión sanguínea que pasen inadvertidos. Sin embargo, con una muestra pareada, esos factores influirán en menor medida, ya que los sujetos son los mismos.
- Puede ser conveniente también mantener un grupo de control en un diseño de seguimiento para descartar otras posibles causas del cambio de presión.
- Por último, un estudio de seguimiento en general será más caro a nivel económico.

Introducción

Test t para muestras pareadas

- Vamos a suponer que hemos tomado muestras pareadas para nuestra investigación. Denotamos por $x_{i,1}$ a las mediciones base, y $x_{i,2}$ a las mediciones de seguimiento.

i	SBP level while not using OCs (x_{i1})	SBP level while using OCs (x_{i2})	d_i^*
1	115	128	13
2	112	115	3
3	107	106	−1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	−2
10	115	117	2

*$d_i = x_{i2} - x_{i1}$

Introducción

Test t para muestras pareadas

- Tenemos que hacer la siguiente suposición:
  - Las variable aleatorias que siguen las mediciones base y las de seguimiento pueden ser distintas. Sin embargo, las diferencias $d_i = x_{i1} - x_{i2}$ suponemos que siguen una distribución normal $\Delta \sim N(\mu_d, \sigma_d)$.
- Entonces, la hipótesis nula será suponer que la media entre ambas medidas es la misma, es decir, $\mu_d = \mu_1 - \mu_2 = 0$.

Introducción

Test t para muestras pareadas

- $H_0 : \mu_d = 0$ vs $H_1 : \mu_d \neq 0$ con varianza desconocida.
- Calculamos $\overline{d} = \frac{d_1+d_2+...+d_n}{n}$ (media muestral de las diferencias).
- Y la cuasivarianza de las dif observadas

$$s_d^2 = \frac{\sum d_i^2 - \frac{1}{n}(\sum d_i)^2}{n-1}$$

- Con esto, tenemos el estadístico:

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

- Hemos convertido el problema de dos muestras, en un contraste para la media de las diferencias (una muestra).

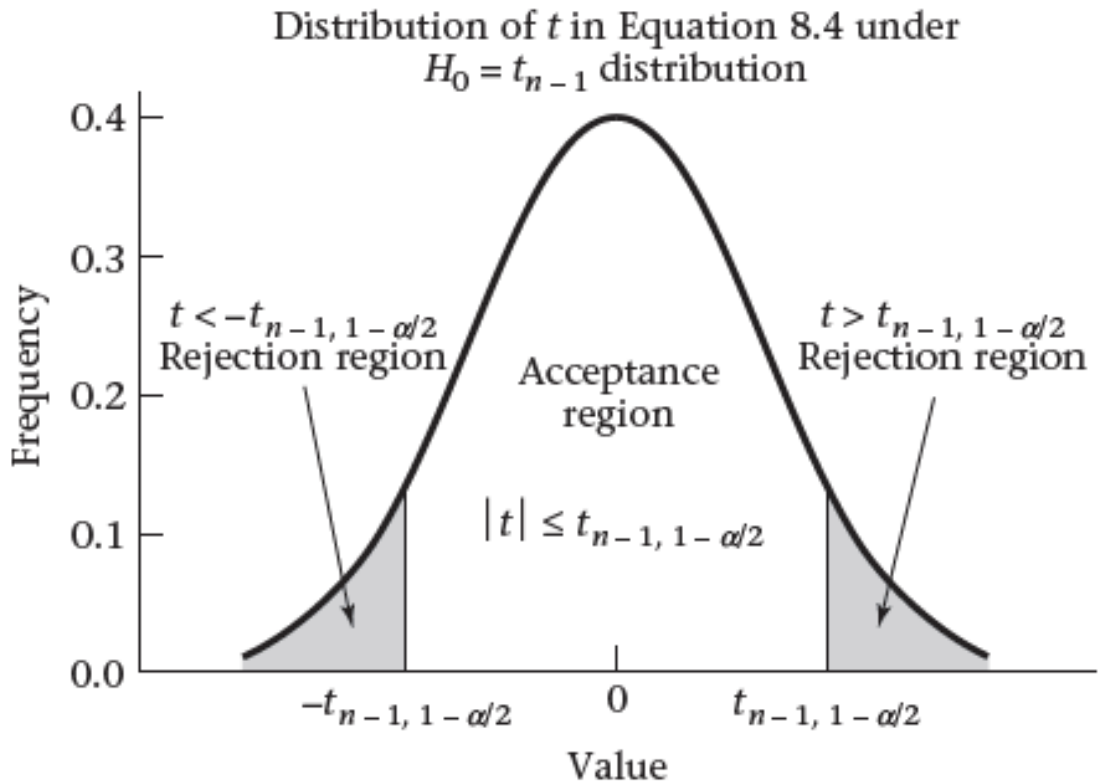
Introducción

Test t para muestras pareadas

- Si pedimos que el test tenga un nivel de significación α, entonces tendremos:
  - Si $t > t_{n-1,1-\alpha/2}$ o $t < -t_{n-1,1-\alpha/2}$, rechazamos H_0
  - Si $-t_{n-1,1-\alpha/2} \leq t \leq t_{n-1,1-\alpha/2}$ aceptamos H_0.

Introducción

Test t para muestras pareadas



Introducción

Test t para muestras pareadas

- El p-valor se calcula teniendo en cuenta que el contraste es de doble alternativa:
- Si $t < 0$, entonces:

$$p = 2 \cdot P(t_{n-1} < t = \frac{\overline{d}}{(s_d/\sqrt{n})})$$

- Si $t > 0$. entonces:

$$p = 2 \cdot P(t_{n-1} > t = \frac{\overline{d}}{(s_d/\sqrt{n})})$$

Introducción

Test t para muestras pareadas

- Decidir para el caso del ejemplo con $\alpha = 0.05$.

i	SBP level while not using OCs (x_{i1})	SBP level while using OCs (x_{i2})	d_i^*
1	115	128	13
2	112	115	3
3	107	106	−1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	−2
10	115	117	2

*$d_i = x_{i2} - x_{i1}$

Introducción

Test t para muestras pareadas

$$
\begin{aligned}
\overline{d} &= (13 + \ldots + 2)/10 = 4.8 \\
s_d^2 &= [(13 - 4.8)^2 + \ldots + (2 - 4.8)^2]/9 = 20.84 \\
s_d &= \sqrt{20.84} = 4.56 \\
t &= 4.8/(4.56/\sqrt{10}) = 3.32 \\
t_{9,0.975} &= 2.262
\end{aligned}
$$

Introducción

Intervalo de confianza para la Δ

- El intervalo de confianza $1-\alpha$ para la diferencia de medias (μ_d) de dos muestras pareadas es

$$\left[\overline{d} - t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}, \overline{d} + t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right]$$

- Calcular el intervalo de confianza del ejemplo anterior.

Introducción

Test para dos poblaciones independientes

- Cuando las dos poblaciones son distintas, cada una de ellas tendrá una distribución.
- Tendremos entonces la gente que no toma el medicamento distribuido como $N(\mu_1, \sigma_1^2)$ y los que lo toman $N(\mu_2, \sigma_2^2)$
- El contraste será: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.
- Vamos a asumir (de momento) $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Introducción

Test para dos poblaciones independientes

- Pero nosotros no conocemos ninguno de estos parámetros (μ_1, μ_2), solo tenemos 2 muestras.
- La idea será rechazar H_0 si la diferencia de medias muestrales $\overline{x_1} - \overline{x_2}$ es muy grande, y no hacerlo si es muy cercano a cero.
- Sabemos que nuestro estimador para la media de la primera población (μ_1) es $\overline{X_1} \sim N(\mu_1, \sigma^2/n_1)$ y $\overline{X_2} \sim N(\mu_2, \sigma^2/n_2)$.
- Como son independientes, tenemos entonces que

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})\right)$$

Introducción

Test para dos poblaciones independientes

- Tendremos 2 opciones: Si σ es conocido, podemos tipificar $\overline{X_1} - \overline{X_2}$ como una normal estándar, y utilizar un contraste de los que hemos visto.
- Si σ es desconocido, tendremos que estimarlo.
- Pero ahora tenemos dos muestras, cada una con su cuasivarianza muestral s_1^2 y s_2^2. Ambas estiman σ, por lo que podríamos coger cualquiera de ellas.
- Pero para hacerlo mejor, podemos intentar buscar un promedio de ambas cuasivarianzas a través de la siguiente fórmula:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Introducción

Test para dos poblaciones independientes

- $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ con significación α y la misma varianza para ambas poblaciones, pero desconocida.
- Se calcula:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- Si $t > t_{n_1+n_2-2,1-\alpha/2}$ o $t < -t_{n_1+n_2-2,1-\alpha/2}$ rechazamos H_0
- Si $-t_{n_1+n_2-2,1-\alpha/2} \leq t \leq t_{n_1+n_2-2,1-\alpha/2}$ aceptamos H_0.

Introducción

Test para dos poblaciones independientes

- $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ con significación α y la misma varianza para ambas poblaciones, pero desconocida.
- Se calcula:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- Si $t > t_{n_1+n_2-2,1-\alpha/2}$ o $t < -t_{n_1+n_2-2,1-\alpha/2}$ rechazamos H_0
- Si $-t_{n_1+n_2-2,1-\alpha/2} \leq t \leq t_{n_1+n_2-2,1-\alpha/2}$ aceptamos H_0.

Introducción

Test para la igualdad de dos varianzas en dos poblaciones indep.

- $H_0 : \sigma_1 = \sigma_2$ vs $H_1 : \sigma_1 \neq \sigma_2$ con significación α y medias desconocidas y distintas.
- Se calcula el estadístico:

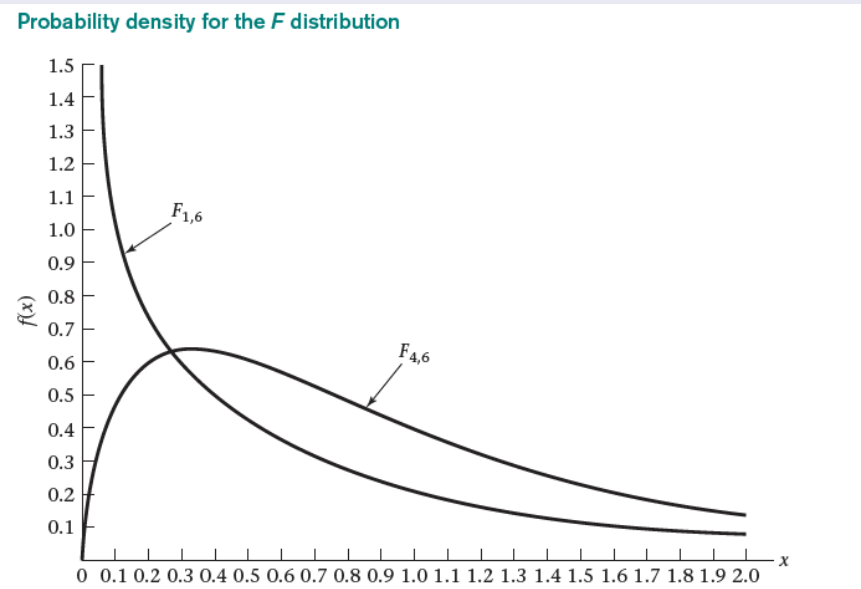
$$F = \frac{s_1^2}{s_2^2}$$

- Si $F > F_{n_1-1,n_2-1,1-\alpha/2}$ o $F < F_{n_1-1,n_2-1,\alpha/2}$ rechazamos H_0
- Si $F_{n_1-1,n_2-1,\alpha/2} \leq F \leq F_{n_1-1,n_2-1,1-\alpha/2}$ aceptamos H_0.

Introducción

Test para la igualdad de dos varianzas en dos poblaciones indep.

- F es una función de distribución conocida como F de Snedecor (estudiada por Fisher y G. Snedecor).
- Esta función de distribución ahora depende de 2 parámetros denominados numerador y denominador de los grados de libertad (n_1 y n_2).



Introducción

Test para la igualdad de dos varianzas en dos poblaciones indep.

- Denotamos por $F_{n_1,n_2,p}$ al p-ésimo percentil como el valor que hace: $P(F_{n_1,n_2} \leq F_{n_1,n_2,p}) = p$

Introducción

Test para la igualdad de dos varianzas en dos poblaciones indep.

- Este test de las varianzas debemos hacerlo antes del test de la media, pues para éste habíamos supuesto que las varianzas son iguales. El test de la F de Snedecor nos arroja dos resultados posibles:
  - Podemos suponer que las varianzas son iguales, y aplicamos el test que hemos visto.
  - No podemos suponer que las varianzas sean iguales... ¿Qué hacemos ahora?

Introducción

Test para la igualdad de dos varianzas en dos poblaciones indep.

- Recordando el caso anterior, tenemeos ahora que nuestro estimador para la media de la primera población (μ_1) es $\overline{X_1} \sim N(\mu_1, \sigma_1^2/n_1)$ y $\overline{X_2} \sim N(\mu_2, \sigma_2^2/n_2)$.
- Siguen siendo independientes, así que podemos sumar las variables, y tenemos

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \sigma_1^2 \cdot \frac{1}{n_1} + \sigma_2^2 \cdot \frac{1}{n_2}\right)$$

- El test que tenemos para varianzas distintas es:

Introducción

Test para la media con varianzas distintas y desconocidas

- $H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$.
- Calculamos el estadístico

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Calculamos los grados de libertad de la t de Student con la fórmula (redondeando al entero más próximo):

$$d' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

Introducción

Test para la media con varianzas distintas y desconocidas

- Si $t > t_{d',1-\alpha/2}$ o $t < -t_{d',1-\alpha/2}$, rechazamos H_0
- Si $-t_{d',1-\alpha/2} \leq t \leq t_{d',1-\alpha/2}$, aceptamos H_0