

Inferencia Estadística I: Contrate de Hipótesis

Introducción a la Estadística

DVM

Contenidos

1 Introducción

2 Errores de Tipo I y de Tipo II, potencia

3 Contraste de hipótesis para μ

- Contraste de hipótesis para μ con σ conocida
- Contraste de hipótesis para μ con σ desconocida

4 Contraste de hipótesis para la proporción

5 Contraste de hipótesis para la diferencia de medias

- Diferencia de medias con varianzas conocidas
- Diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)
- Diferencia de medias con varianzas desconocidas y desiguales

Introducción

¿Qué es un contraste de hipótesis?



Un contraste de hipótesis es un procedimiento basado en datos muestrales que se usan como información para evaluar la validez de una conjetura sobre la naturaleza de una población (algún parámetro o su distribución).

A la conjetura a testar se le llama **hipótesis nula** (H_0) y la mantenemos a no ser que encontremos evidencia en los datos muestrales que no la apoyen si no que estén a favor de una **alternativa** (H_1)

Introducción

- A) Lays afirma que, en promedio, cada bolsa de patatas pesa al menos 250 g. Quieres contrastar esta información a partir de los pesos de las bolsas en una muestra aleatoria. Población: $X =$ 'peso de una bolsa de patatas (en g)

Hipótesis nula, $H_o : \mu \geq 250g$

Hipótesis alternativa, $H_1 : \mu < 250g$

¿Proporciona nuestra m.a.s. suficiente evidencia para rechazar H_o ?

- B) Un estudio afirma que las habilidades matemáticas no dependen del género. Elegimos una muestra representativa y nos planteamos:

Hipótesis nula, $H_o : \mu_{hombres} = \mu_{mujeres}$

Hipótesis alternativa, $H_1 : \mu_{hombres} \neq \mu_{mujeres}$

¿Proporciona nuestra m.a.s. suficiente evidencia para rechazar H_o ?

Contenidos

1 Introducción

2 Errores de Tipo I y de Tipo II, potencia

3 Contraste de hipótesis para μ

- Contraste de hipótesis para μ con σ conocida
- Contraste de hipótesis para μ con σ desconocida

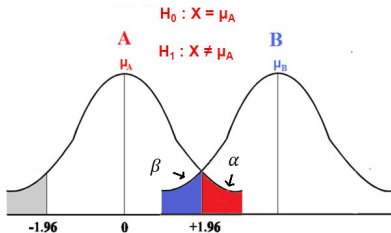
4 Contraste de hipótesis para la proporción

5 Contraste de hipótesis para la diferencia de medias

- Diferencia de medias con varianzas conocidas
- Diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)
- Diferencia de medias con varianzas desconocidas y desiguales

Error Tipo I, Tipo II y Potencia

		Muestra	
		No rechazar H_0	Rechazar H_0
Población	H_0 verdadera	Decisión correcta $1 - \alpha = P(\text{No rechazar } H_0 H_0 \text{ verdadera})$	Error tipo I $\alpha = P(\text{Rechazar } H_0 H_0 \text{ verdadera})$
	H_0 falsa	Error tipo II $\beta = P(\text{No rechazar } H_0 H_0 \text{ falsa})$	Decisión correcta $1 - \beta = P(\text{Rechazar } H_0 H_0 \text{ falsa})$
		SIGNIFICANCIA	POTENCIA




Error Tipo I, Tipo II y Potencia

$\hat{Y} = 0$
NEGATIVO

$\hat{Y} = 1$
POSITIVO


$Y = 0$
No embarazo

Verdadero negativo



Usted no está embarazado

Falso positivo




Usted está embarazado

ERROR TIPO 1

$Y = 1$
Embarazo


Falso negativo



Usted no está embarazada

ERROR TIPO 2

Verdadero positivo



Usted está embarazada

Error Tipo I, Tipo II y Potencia

- Los errores de Tipo I y de Tipo II no se pueden cometer simultáneamente
 - ▶ El error de Tipo I solo puede darse si H_0 es verdadera.
 - ▶ El error de Tipo II solo puede darse si H_0 es falsa.
- Si la probabilidad del error de Tipo I, α aumenta, entonces la probabilidad del error de Tipo II, β disminuye
- Si todo lo demás no cambia:
 - ▶ β es mayor cuando la diferencia entre el valor supuesto para el parámetro y su valor real disminuye
 - ▶ β es mayor cuando la varianza poblacional es mayor
 - ▶ β es mayor cuando el tamaño muestral es menor y cuando el nivel de significancia (α) es menor

Error Tipo I, Tipo II y Potencia

Contenidos

1 Introducción

2 Errores de Tipo I y de Tipo II, potencia

3 Contraste de hipótesis para μ

- Contraste de hipótesis para μ con σ conocida
- Contraste de hipótesis para μ con σ desconocida

4 Contraste de hipótesis para la proporción

5 Contraste de hipótesis para la diferencia de medias

- Diferencia de medias con varianzas conocidas
- Diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)
- Diferencia de medias con varianzas desconocidas y desiguales

Contraste de hipótesis para μ con σ conocida

Si tenemos una población normal y conocemos su varianza, usamos el estimador puntual de la media poblacional (la media muestral) para el contraste de hipótesis

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Tendríamos como estadístico de contraste nuestra

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \rightarrow N(0, 1)$$

Podemos hacer tres tipos de contraste donde $H_o : \mu = \bar{X}$:

$$H_1 : \mu > \bar{X} \mid H_1 : \mu < \bar{X} \mid H_1 : \mu \neq \bar{X}$$

Contraste de hipótesis para μ con σ conocida

Se presupone que la hipótesis nula es cierta, y se rechaza cuando:

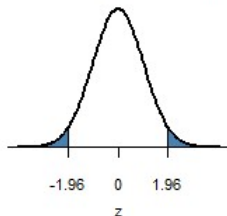
Con $H_o : \mu = \mu_0$

$$H_1 : \mu \neq \mu_0 \quad \text{Rechazo } H_o \text{ si } \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

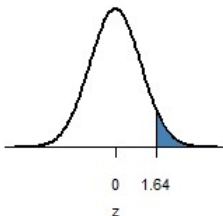
$$H_1 : \mu > \mu_0 \quad \text{Rechazo } H_o \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha}$$

$$H_1 : \mu < \mu_0 \quad \text{Rechazo } H_o \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha}$$

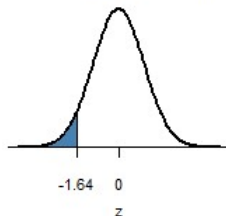
Two-tails (alpha=0.05)



Right tail (alpha=0.05)



Left tail (alpha=0.05)

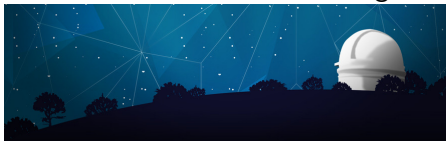


Contraste de hipótesis para μ con σ conocida

Ejemplo

[Ross] Suponga que si se emite una señal de intensidad μ desde una estrella en particular, entonces el valor recibido en un observatorio en la tierra es una v.a. $X \rightarrow N(\mu, 4)$. En otras palabras, el valor de la señal emitida está alterado por ruido aleatorio, que normalmente se distribuye con media 0 y desviación estándar 4. Se sospecha que la intensidad de la señal es igual a 10.

Pruebe si esta hipótesis es plausible si la misma señal se recibe 20 veces con un promedio de 11.6. Utilizar el nivel de significancia del 5 %.



Contraste de hipótesis para μ con σ conocida

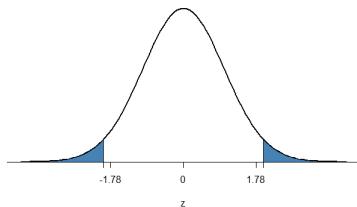
Ejemplo

$$H_o : \mu = 10$$

$$H_1 : \mu \neq 10$$

Como la distribución es normal con varianza conocida, usamos

$$Z = \frac{11.6-10}{4/\sqrt{20}}; Z = 1,78$$



Los datos no son inconsistentes con H_o . Sin embargo, ¿qué pasaría si el nivel de significación fuera de 0,10?

Contraste de hipótesis para μ con σ conocida

Ejemplo R

El registro histórico de los últimos tres años muestra que las tasas de accidentes de camiones en las carreteras de 2 carriles en una ciudad es 0.5 accidentes por millón de kilómetros por vehículo con una desviación estándar de 0.1.

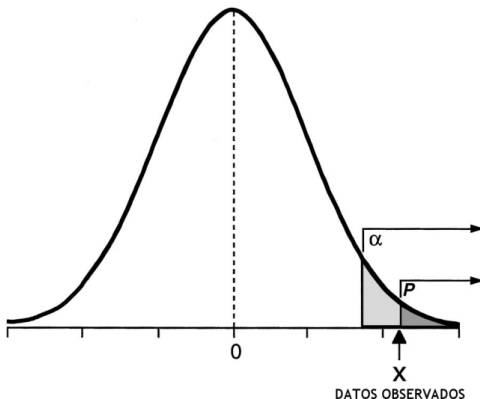
Después de completar el curso obligatorio de actualización para conductores de camiones, se registraron las tasas de accidentes en 50 sitios aleatorios de carreteras de 2 carriles para estimar las estadísticas actuales. Calcule si el curso de conducción hizo alguna diferencia en términos de las tasas de accidentes.

Ir a: `contraste_hipo_una_media_var_con.R`

¿Cuál es el nivel de significancia "correcto"?

- No se puede dar una respuesta científica a esto.
- Si el hecho de rechazar H_0 cuando en realidad es cierta (Error tipo I) resultara en costos muy altos entonces probablemente se elija un nivel de significación pequeño (un α pequeño)
- Por ejemplo: Imaginemos que H_1 : el método X es superior al actual. Entonces, el rechazo de H_0 también implicaría muchos gastos que no queríamos incurrir equivocadamente. Queremos hacer que la probabilidad de rechazar la nula siendo verdadera sea muy baja.
- La metodología más razonable es obtener el p-valor y, si es posible, definir antes de la obtención de la muestra una diferencia mínima significativa que garantice la potencia deseada (definiremos a continuación estos dos conceptos). Sólo con estas tres cantidades el contraste queda satisfactoriamente planteado.

Concepto de p-valor



El p-valor representa la probabilidad de observar resultados iguales o más extremos que los proporcionados por la muestra. Un valor p pequeño (digamos, 0.05 o menos) es un indicador fuerte de que la hipótesis nula no es cierta.

Concepto de p-valor

Ejemplo

- En la práctica, el nivel de significación a menudo no se establece de antemano; más bien, los datos son usados para determinar el valor p .
- Este valor es a menudo tan grande que está claro que la hipótesis nula no debe ser rechazada o tan pequeña que queda claro que la hipótesis nula debe ser rechazada.
- En general si $p > \alpha$, no rechazamos H_0 y viceversa.

Supongamos que la media muestral en el [ejemplo](#) anterior es de 10,8. Calcula el p-valor y di si hay evidencias para rechazar o no la hipótesis nula.

Concepto de p-valor

Ejemplo

Tenemos: $Z = \frac{10.8-10}{4/\sqrt{20}}; Z = 0,89$

Contraste de hipótesis para μ con σ desconocida

Si tenemos una población normal y desconocemos su varianza, usamos el estadístico de la media muestral para el contraste de hipótesis

$$\bar{X} \rightarrow N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

Tendríamos como estadístico de contraste:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \rightarrow t_{n-1}$$

Podemos hacer tres tipos de contraste donde $H_o : \mu = \bar{X}$:

$$H_1 : \mu > \bar{X} \mid H_1 : \mu < \bar{X} \mid H_1 : \mu \neq \bar{X}$$

Contraste de hipótesis para μ con σ desconocida

Se presupone que la hipótesis nula es cierta, y se rechaza cuando:

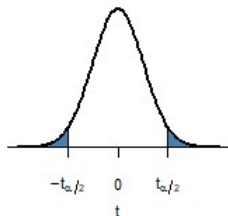
Con $H_o : \mu = \mu_0$

$$H_1 : \mu \neq \mu_0 \quad \text{Rechazo } H_o \text{ si } \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1; \alpha/2}$$

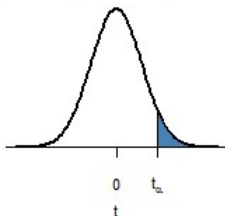
$$H_1 : \mu > \mu_0 \quad \text{Rechazo } H_o \text{ si } \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1; \alpha}$$

$$H_1 : \mu < \mu_0 \quad \text{Rechazo } H_o \text{ si } \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1; \alpha}$$

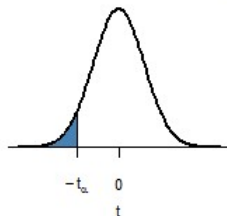
Two-tails (alpha=0.05, df>30)



Right tail (alpha=0.05, df>30)



Left tail (alpha=0.05, df>30)



Contraste de hipótesis para μ con σ desconocida

Ejemplo

[Ross] Los datos históricos indican que el nivel medio de acidez (pH) de la lluvia en una determinada región industrial es de 5.2. Para ver si ha habido algún cambio reciente en este valor, los niveles de acidez de 12 tormentas en el pasado año se han medido, con los siguientes resultados:



6.1, 5.4, 4.8, 5.8, 6.6, 5.3, 6.1, 4.4, 3.9, 6.8, 6.5, 6.3

¿Son estos datos lo suficientemente sólidos, con un nivel de significación del 5 %, para que podamos concluir que la acidez de la lluvia ha cambiado desde su valor histórico?

Contenidos

- 1 Introducción
- 2 Errores de Tipo I y de Tipo II, potencia
- 3 Contraste de hipótesis para μ
 - Contraste de hipótesis para μ con σ conocida
 - Contraste de hipótesis para μ con σ desconocida
- 4 Contraste de hipótesis para la proporción
- 5 Contraste de hipótesis para la diferencia de medias
 - Diferencia de medias con varianzas conocidas
 - Diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)
 - Diferencia de medias con varianzas desconocidas y desiguales

Contraste de hipótesis para la proporción

Si tenemos una población normal, usamos el estadístico de la proporción muestral para el contraste de hipótesis que sigue una distribución:

$$\hat{p} \rightarrow N(p, \frac{\sqrt{pq}}{\sqrt{n}})$$

Tendríamos como estadístico de contraste:

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \rightarrow N(0, 1)$$

Podemos hacer tres tipos de contraste donde $H_o : p = p_0$:

$$H_1 : p > p_0 \mid H_1 : p < p_0 \mid H_1 : p \neq p_0$$

Contraste de hipótesis para la proporción

Se presupone que la hipótesis nula es cierta, y se rechaza cuando:

Con $H_o : p = p_0$

$$H_1 : p \neq p_0 \quad \text{Rechazo } H_o \text{ si } \left| \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right| > z_{\alpha/2}$$

$$H_1 : p > p_0 \quad \text{Rechazo } H_o \text{ si } \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_{\alpha}$$

$$H_1 : p < p_0 \quad \text{Rechazo } H_o \text{ si } \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < z_{\alpha}$$

Contraste de hipótesis para la proporción

Ejemplo



Una muestra de 800 artículos producidos en una nueva máquina mostró que 48 de ellos son defectuosos. La fábrica dice que se deshará de la máquina si los datos indican que la proporción de artículos defectuosos es significativamente mayor que 5%.

Con un nivel de significación del 10%, ¿qué aconsejas?

Contraste de hipótesis para la proporción

Ejemplo

$$H_0 : p \leq 0,05$$

$$H_1 : p > 0,05$$

Nuestro estadístico es : $Z = \frac{0,06 - 0,05}{\sqrt{(0,05 * 0,95) / 800}} \approx 1.3$ con un p-valor:

$P(Z \geq 1.3) = 0,0968$ por lo que rechazamos la hipótesis al 0.10 nivel de sig.

En R, usamos la función `z.test`:

```
1 z.test(48,800,p=0.05, alternative="greater")
```

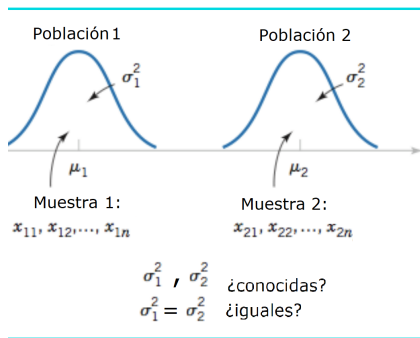
Contenidos

- 1 Introducción
- 2 Errores de Tipo I y de Tipo II, potencia
- 3 Contraste de hipótesis para μ
 - Contraste de hipótesis para μ con σ conocida
 - Contraste de hipótesis para μ con σ desconocida
- 4 Contraste de hipótesis para la proporción
- 5 Contraste de hipótesis para la diferencia de medias
 - Diferencia de medias con varianzas conocidas
 - Diferencia de medias con varianzas desconocidas e iguales(Pooled t-test)
 - Diferencia de medias con varianzas desconocidas y desiguales

Contraste de hipótesis para la diferencia de medias

- Hasta ahora, solo hemos trabajado con una población.
- Si tuviéramos x_{11}, \dots, x_{1n} un m.a.s. de n observaciones procedente de una población con media μ_1 y varianza σ_1^2 ; y x_{21}, \dots, x_{2n} , otra m.a.s. de n observaciones tomada de una segunda población con valor esperado μ_2 y varianza σ_2^2
- Asumiendo la independencia de ambas muestras, si \bar{x}_1 y \bar{x}_2 son las medias muestrales, entonces la estadística $\bar{x}_1 - \bar{x}_2$ es un estimador puntual de $\mu_1 - \mu_2$, y tiene una distribución normal si las dos poblaciones son normales, o aproximadamente normal si cumple con las condiciones del TCL.

Contraste de hipótesis para la diferencia de medias



- 1 Diferencia de medias con varianzas conocidas
- 2 Diferencia de medias con varianzas desconocidas e iguales.
- 3 Diferencia de medias con varianzas desconocidas y desiguales.

Contraste de hipótesis para la diferencia de medias con varianzas conocidas

Si tenemos dos muestras m.a.s. procedentes de poblaciones normales e independientes y queremos evaluar la diferencia de sus medias, usamos el estimador $\bar{x}_1 - \bar{x}_2$:

Con $H_0 : \mu_1 - \mu_2 = \delta$ podemos tener H_1 como habitualmente la definimos con $<, >$ ó \neq .

Y tenemos al estadístico de contraste:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Contraste de hipótesis para la diferencia de medias con varianza conocida

Ejemplo

Recientemente se ha instalado el pago sin contacto a través de móvil en el supermercado XYZ. El gerente desea saber si el tiempo medio de pago con el método de pago estándar es más largo que con el nuevo.

El gerente ha reunido la siguiente info. de muestra:

	Tipo de cliente	Media muestral	Desv.std.(poblacional)	N
	Estándar	5.50 mins	0.40 mins	50
	Contact-less	5.30 mins	0.30 mins	100

Contraste de hipótesis para la diferencia de medias con varianza conocida

Ejemplo

El gerente ha reunido la siguiente info. de muestra:



Tipo de cliente	Media muestral	Desv.std.(poblacional)	N
Estándar	5.50 mins	0.40 mins	50
Contact-less	5.30 mins	0.30 mins	100

- $Z = \frac{\bar{x}_e - \bar{x}_{cl}}{\sqrt{\sigma_e^2/n_e + \sigma_{cl}^2/n_{cl}}} = \frac{5,5 - 5,3}{\sqrt{0,40^2/50 + 0,30^2/100}} = 3,13$
- 3,13 es mayor que el valor critico al 0.05 (2,33) por lo que se rechaza la hipótesis nula . La diferencia de 0.2 minutos es demasiado grande para haber ocurrido por casualidad por lo que se concluye que el método contact-less es más rápido.

Contraste de hipótesis para la diferencia de medias con varianzas desconocidas e iguales(Pooled t-test)

Cuando las varianzas son desconocidas, se debe realizar previamente una prueba estadística para verificar si éstas son iguales o diferentes.

Para hacerlo usamos la distribución F. Aquí tomaremos esa información como dada.

En el caso en el que no conocemos las varianzas de las poblaciones pero estas son iguales, usamos el estadístico t:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}$$

donde s_p es un estimador combinado de s_1 y s_2 : $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

Y los grados de libertad vienen dados por : $n_1 + n_2 - 2$

Contraste de hipótesis para la diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)

Ejemplo

Una empresa fabrica y ensambla cortadoras de césped que se envían a distribuidores en todo España y Francia. Se han propuesto dos procedimientos diferentes para montar el motor en el marco del cortacésped.

Para evaluar los dos métodos, se decidió realizar un estudio de tiempo y movimiento. Una muestra de cinco empleados fue cronometrada usando el método 1 y seis usando el método 2.

Los resultados, en minutos, se muestran a continuación. ¿Hay alguna diferencia en los tiempos medios de montaje? Use el nivel de significancia 0,10.

Método 1	2	4	9	3	2	
Método 2	3	7	5	8	4	3

Contraste de hipótesis para la diferencia de medias con varianzas desconocidas e iguales (Pooled t-test)

Ejemplo

$$H_o : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Tenemos que calcular s_1 y s_2 para calcular s_p dado que:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{(1/n_1 + 1/n_2)}}$$

$$\text{Como } s_p = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{(5-1)2,9155^2 + (6-1)2,0976^2}{5+6-2} = 6,2222$$

entonces,

$$t \approx \frac{4-5}{2,5\sqrt{1/5+1/6}} = -0,662$$

Tomamos la decision de no rechazar H_o pues 0,662 es menor que 1,833($t_{0,10/2,n_2+n_1-2}$). Así, concluimos que no hay diferencia en los tiempos medios de ambos métodos.

Contraste de hipótesis para la diferencia de medias con varianzas desconocidas y desiguales

Si estamos ante el mismo caso que antes pero esta vez las varianzas poblacionales no son iguales, entonces usamos:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

pero esta vez con los grados de libertad vienen dados por:

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1)] + [(s_2^2/n_2)^2/(n_2-1)]}$$