

Intervalos de confianza y contrastes de hipótesis

Diferencia de medias en dos poblaciones

Grado de Biología sanitaria - UAH

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

29 de noviembre de 2017

Inferencia: diferencia de medias

Queremos responder preguntas tipo: estudiamos la longitud del caparazón de cangrejos de la especie *Leptograpsus variegatus* agrupados según su color (azul o naranja).



A la vista de las **longitudes muestrales** $\bar{X}_1$ y $\bar{X}_2$:

Intervalo de confianza

- 1 **A nivel poblacional**, ¿en cuantas unidades difieren μ_B y μ_O ?
- 2 Fijado un nivel de confianza, ¿puedo esperar que $\mu_B > \mu_O + 4mm$?

Contraste de hipótesis sobre las longitudes, vista la muestra,

- 1 ¿Son **significativamente** diferentes **a nivel poblacional**?

Notación

- n_1 y n_2 el tamaño de cada muestra
- $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales

Inferencia: diferencia de medias

Distribución muestral para la diferencia de medias

Suponer que las poblaciones

- 1 son independientes
- 2 se distribuyen conforme a una normal
- 3 conocemos la varianza en cada población

El teorema central del límite afirma que

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right), \quad \text{y que} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right).$$

Sabemos que la diferencia $\bar{X}_1 - \bar{X}_2$ es una normal. Concretamente:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Los intervalos de confianza, las regiones de rechazo, y el p-valor se calculan de forma análoga a como lo hicimos para una población.

Las expresiones concretas están en tabla B.1 (pág 576), B.2 (pág 577) y B.3 (pág 578) del libro

Inferencia: diferencia de medias

Distribución muestral para la diferencia de medias

Suponer que las

- 1 son independientes
- 2 las muestras son “grandes”: $n_1 > 30, n_2 > 30$
- 3 NO conocemos las varianzas

El teorema central del límite afirma que

$$\bar{X}_1 \sim N\left(\mu_1, \frac{s_1}{\sqrt{n_1}}\right), \quad \text{y que} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{s_2}{\sqrt{n_2}}\right).$$

Sabemos que la diferencia $\bar{X}_1 - \bar{X}_2$ es una normal. Concretamente:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

Los intervalos de confianza, las regiones de rechazo, y el p-valor se calculan de forma análoga a como lo hicimos para una población.

Las expresiones concretas están en tabla B.1 (pág 576), B.2 (pág 577) y B.3 (pág 578) del libro

Inferencia: diferencia de medias

Ejemplo: Para las especies que aparecen en el fichero `crabs` de la librería `MASS` de R se quiere probar si hay diferencias significativas entre la longitud del lóbulo frontal (`FL`) en cada una de las especies analizadas (naranja y azul).

- 1 Contrasta las hipótesis

$$H_0 : \mu_O = \mu_B, \quad H_1 : \mu_O \neq \mu_B$$

al nivel de significación $\alpha = 0,05$

Las muestras tienen tamaño 100. PLANTILLA

- 2 ¿A partir de qué diferencia de medias es significativo el contraste?
Recuerda: si H_0 es cierta, $\mu_O = \mu_B$
- 3 Si se rechaza el contraste bilateral, ¿es necesario hacer otro contraste unilateral para ordenar las medias (al nivel de significación fijado)?

Inferencia: diferencia de medias

Ejemplo: Sobre el ejemplo anterior, estimar la diferencia en la longitud del lóbulo frontal (FL) de las dos especies con un n.c del 95 %.
Las muestras tienen tamaño 100.
PLANTILLA

$$IC_{\mu_B - \mu_O} = (-3,9, -2,2)$$

Interpretar:

- 1 ¿qué se puede esperar?
- 2 Complementa muy bien la información del p-valor

Inferencia: diferencia de medias

Distribución muestral para la diferencia de medias

Suponer que las muestras

- 1 son independientes
- 2 son **muestras “pequeñas”**: $n_1 < 30$ o $n_2 < 30$ y de poblaciones **normales**
- 3 NO conocemos las **varianzas**, pero podemos suponerlas **distintas**

Podemos usar que

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_f$$

y el número de grados de libertad f se calcula con la aproximación de Welch

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{\frac{s_1^4}{n_1^2}}{(n_1 - 1)} \right) + \left(\frac{\frac{s_2^4}{n_2^2}}{(n_2 - 1)} \right)}$$

Los intervalos de confianza, las regiones de rechazo, y el p-valor se calculan de forma análoga a como lo hicimos para una población.

Las expresiones concretas están en tabla B.1 (pág 576), B.2 (pág 577) y B.3 (pág 578) del libro

Inferencia: diferencia de medias

Distribución muestral para la diferencia de medias

Suponer que

- 1 son independientes
- 2 son **muestras “pequeñas”**: $n_1 < 30$ o $n_2 < 30$ y de **poblaciones normales**
- 3 NO conocemos las **varianzas**, pero podemos suponerlas **iguales**

Podemos usar que

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{X_1} - \mu_{X_2})}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Los intervalos de confianza, las regiones de rechazo, y el p-valor se calculan de forma análoga a como lo hicimos para una población.

Las expresiones concretas están en tabla B.1 (pág 576), B.2 (pág 577) y B.3 (pág 578) del libro

Inferencia: diferencia de medias

Distribución muestral para la diferencia de medias, **muestras pareadas**

Esto sucede cuando las dos muestras se toman sobre los mismos individuos. Por ejemplo:

- 1 Frecuencia cardíaca antes y después de hacer ejercicio, medida en los mismos individuos
- 2 Frente una erupción cutánea, se aplica a todos los individuos un remedio sólo en parte de la zona afectada

Ahora las muestras NO son independientes

Se define la variable variable aleatoria

$$D = X_1 - X_2$$

Los intervalos de confianza, las regiones de rechazo, y el p-valor se calculan de forma análoga a como lo hicimos para una población.

Las expresiones concretas están en tabla B.1 (pág 576), B.2 (pág 577) y B.3 (pág 578) del libro

Inferencia: diferencia de medias

Ejemplo: El tratamiento anticoagulante por vía oral se realiza con dicumarínicos. Se midió el porcentaje de nivel de protrombina (que determina el grado de coagulación) a 9 pacientes, antes y después del tratamiento con dicho fármaco y se obtuvieron los siguientes resultados

Antes	95	85	77	100	92	67	81	94	99
Después	80	90	72	102	109	66	73	80	92
Después-antes	-15	5	-5	2	17	-1	-8	-14	-7

Determina si el tratamiento es efectivo así como una medida de su efecto

- 1 Muestra pequeña
- 2 σ desconocida
- 3 Si podemos suponer que el porcentaje de nivel de protrombina se distribuye de acuerdo con una normal

Usar una t de Student con 9 – 1 grados de libertad

Tratamiento efectivo $\Leftrightarrow H_0 : \mu_{\text{antes}} \leq \mu_{\text{despus}}, \quad H_1 : \mu_{\text{antes}} > \mu_{\text{despus}},$
 $\alpha = 0,01$

¿Es efectivo el tratamiento?: IC para la diferencia de medias