Inferencia sobre la diferencia de proporciones y el cociente de varianzas

Intervalos de confianza y contraste de hipótesis

Grado Biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

20 de noviembre de 2017

Inferencia: diferencia de propociones

Queremos responder preguntas tipo: el escarabajo tigre puede tener distintas coloraciones

Se toman muestras

- De los 500 escarabajos de Providence, Rhode Island, 95 eran negros
- De los 112 escarabajos en Aqueduct, Nueva York, 17 eran negros

1. **Intervalo de confianza** ¿En cuánto difiere la proporción poblacional de individuos negros en cada colonia?

2. **Contrastar hipótesis** sobre la proporción de individuos negros
   - ¿son significativamente diferentes? (bilateral)
   - ¿es significativamente mayor en una de las colonias? (unilateral)

Sección 9.1 del libro

Inferencia: diferencia de propociones

Distribución muestral de la diferencia de proporciones
Dos poblaciones cuyos individuos pueden presentar (éxito) o no (fracaso) una característica determinada

1. Dos muestras de tamaños n_1 y n_2
2. Proporciones muestrales $\hat{p}_i$
3. Además, por el Teorema del límite central, si

$$n_1 > 30, \quad n_2 > 30, \quad n_1 \cdot \hat{p}_1 > 5, \quad n_1 \cdot \hat{q}_1 > 5, \quad n_2 \cdot \hat{p}_2 > 5, \quad n_2 \cdot \hat{q}_2 > 5$$

entonces, $\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1}}\right)$ y $\hat{p}_2 \sim N\left(p_2, \sqrt{\frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}\right)$

¿Cómo se comporta la suma (o resta) de v.a. normales?

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}\right)$$

Ahora todo es análogo al caso de 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 propociones

Ejemplo: el de los escarabajo tigre.

- De los 500 escarabajos de Providence, Rhode Island, 95 eran negros
- De los 112 escarabajos en Aqueduct, Nueva York, 17 eran negros

Se pide

1. Calcula e interpreta el i.c. para la diferencia de proporciones, $\alpha = 0.1$
2. A la vista de los valores muestrales,

$$\hat{p}_1 = \frac{95}{500} = 0.19 \qquad \hat{p}_2 = \frac{17}{112} = 0.1517857$$

   Establece las hipótesis adecuadas sobre las proporciones de escarabajos negros en cada población y contrástalas

Inferencia: diferencia de propociones

Ejemplo: el escarabajo tigre.

1. Usamos las plantillas de intervalos de confianza

$$(-0.025,\ 0.101)$$

   INTERPRETACION

2. Usamos las plantillas de contraste de hipótesis para contrastar

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

   y también

$$H_0 : p_1 \leq p_2 \qquad H_1 : p_1 > p_2$$

Inferencia sobre dos poblaciones: varianzas

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

Se obtienen sendas muestras de longitudes $\bar{X}_1$ y $\bar{X}_2$;

1. **Intervalo de confianza** ¿Es posible comparar lo agrupados que están las longitudes en cada subgrupo? ¿en cuál hay más variabilidad?

2. **Contraste de hipótesis** Sobre las longitudes, vista la muestra,
   - ¿Son diferentes a nivel poblacional?
   - ¿Hay más dispersión en uno de los subgrupos a nivel poblacional?

Inferencia sobre dos poblaciones: varianzas

Ejemplo: Supongamos que $\sigma_1^2 = \dfrac{1}{1000}$, $\sigma_2^2 = \dfrac{1}{1000000}$. Entonces

$$\sigma_1^2 - \sigma_2^2 = 0.000999, \quad \text{pero} \quad \frac{\sigma_1^2}{\sigma_2^2} = 1000.$$

- La diferencia $\sigma_1^2 - \sigma_2^2$ es *sensible a la escala*
- El cociente no tiene ese problema

Si el cociente de dos números es cercano a uno, podemos asegurar que los dos números son realmente parecidos, con independencia de la escala de medida.

A menudo, lo más útil es usar:

- La **diferencia** para comparar medidas de centralización (medias, proporciones, medianas, etc.): **en cuántas unidades difieren**
- El **cociente** para comparar medidas de dispersión (varianzas, recorridos intercuartílicos, etc) **cuántas veces más disperso**

`Sección 9.3 del libro`

Distribución muestral para el cociente de varianzas de poblaciones normales

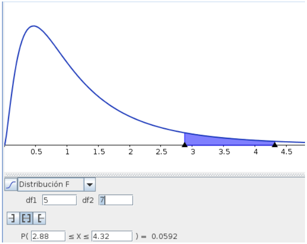
Dadas dos muestras de tamaños n_1 y n_2, sabemos que

$$(n_1 - 1)\frac{s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \qquad \text{y} \qquad (n_2 - 1)\frac{s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

dividiendo,

$$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$$

La distribución de probabilidad del miembro derecho se llama **F de Fisher-Snedecor con $n_1 - 1$ y $n_2 - 1$ grados de libertad**: F_{n_1-1,n_2-1}



Inferencia sobre dos poblaciones: varianzas

Dadas dos poblaciones normales y muestras independientes de tamaños n_1 y n_2, el **intervalo de confianza al nivel** $nc = (1-\alpha)$ **para el cociente de varianzas** $\frac{\sigma_1^2}{\sigma_2^2}$ es:

$$\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{n_1-1,n_2-1;\alpha/2}},\ \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\right)$$

con $F_{n_1-1,n_2-1,\alpha}$ el valor crítico α de la F_{n_1-1,n_2-1}.

Ejemplo: dadas dos muestras de poblaciones normales con, $n_1 = 55$, $n_2 = 47$, $s_1 = 3$, $s_2 = 4.1$ calcular el IC al nc 95 % para el cociente de las varianzas

$$IC_{\sigma_1^2/\sigma_2^2} = \left(\frac{3}{4.1}\cdot\frac{1}{f_{54,46;0.025}},\ \frac{3}{4.1}\cdot\frac{1}{f_{54,46;0.975}}\right) = (0.303,\ 0.93)$$

donde:

$f_{54,46;0.025} =$ `qf(0.975, df1 = 54, df2 = 46)` = 1.7665018

$f_{54,46;0.975} =$ `qf(0.025, df1 = 54, df2 = 46)` = 0.5733775

INTREPRETACION

Ejemplo: Para las especies que aparecen en el fichero `crabs` de la librería `asbio` de `R` se quiere estimar la relación entre la variabilidad de la longitud del lóbulo frontal (`FL`) en cada una de las especies analizadas (naranja y azul). `aggregate(crabs$FL ~ crabs$sp, FUN = sd )`

```
crabs$sp  crabs$FL
       B 3.019610
       O 3.275575
```

Si se toma otra muestra, ¿qué se puede esperar? usa un nivel de confianza del 95 %

- Cálculos con la plantilla.
- INTERPRETACION

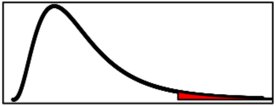
Inferencia sobre dos poblaciones: varianzas

Contraste sobre el cociente de varianzas con nivel de significación α.

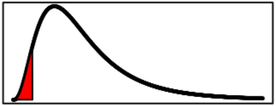
Región de rechazo: se usa el estadístico

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

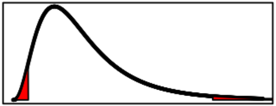
(a) $H_0 : \sigma_1^2 \leq \sigma_2^2 \qquad H_1 : \sigma_1^2 > \sigma_1^2$. Región de rechazo: $Y > F_{n_1-1,n_2-1,\alpha}$



(b) $H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_1 : \sigma^2 < \sigma_0^2$. Región de rechazo: $Y < F_{n_1-1,n_2-1,1-\alpha}$



(c) $H_0 : \sigma^2 = \sigma_0^2 \; H_1 : \sigma^2 \neq \sigma_0^2$. Región de rechazo: $Y \notin \left(F_{n_1-1,n_2-1,1-\alpha/}, F_{n_1-1,n_2-1,\alpha/2}\right)$



Ejemplo 1: dadas dos muestras de poblaciones normales con, $n_1 = 55$, $n_2 = 47$, $s_1 = 3$, $s_2 = 4.3$ contrasta la hipótesis

$$H_0 : \sigma_1^2 \geq \sigma_2^2 \qquad\qquad H_1 : \sigma_1^2 < \sigma_2^2$$

Ejemplo 2: contrasta la igualdad de varianzas en el ejemplo del fichero `crabs`

Inferencia sobre dos poblaciones: varianzas

El cálculo del **p-valor** es:

(a) $H_0 : \sigma_1^2 \leq \sigma_2^2 \qquad H_a : \sigma_1^2 > \sigma_2^2$.
p-valor=$P\left(F_{k_1,k_2} > s_1^2/s_2^2\right)$ (cola derecha)

(b) $H_0 : \sigma_1^2 \geq \sigma_2^2 \qquad H_a : \sigma_1^2 < \sigma_2^2$.
p-valor=$P\left(F_{k_1,k_2} < s_1^2/s_2^2\right)$ (cola izquierda).

(c) $H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_a : \sigma_1^2 \neq \sigma_2^2$.
p-valor=$2 \cdot P\left(F_{k_1,k_2} > s_1^2/s_2^2\right)$ ¡¡siempre que sea $s_1^2/s_2^2 \geq 1$!! Si se tiene $s_1^2/s_2^2 < 1$, cambiar s_1 por s_2.