Práctica 2. Inferencia Estadística.

Grado en Ciencia e Ingeniería de Datos. Universidad Rey Juan Carlos

Isaac Martín

2024-11-13

Introducción

Existen varias formulas que describen cómo calcular el tamaño muestra necesario para comparar una proporción con un valor de referencia. En esta práctica vamos a seguir los argumentos usados por Whitehead (1983) que permite derivar varias de estas fórmulas a partir de una teoría unificada, que además puede ser usada en diseños de experimentos.

Consideremos una secuencia $X_1, \ldots, X_n$ de variables aleatorias bernoulli i.i.d. con probabilidad de éxito igual a p. Queremos testear la hipótesis:

$$H_0 : p = p_0$$

Esto es equivalente a contrastar si un parámetro θ es igual a cero, siendo $\theta = g(p, p_0)$, y g una función que parametriza la diferencia entre p y p_0.

Se pide

1. Demostrar que la log-verosimilitud de p basada en n observaciones es:

$$\ell(p) = r \times log\left(\frac{p}{1-p}\right) + n \times log(1-p)$$

Donde $r = x_1 + x_2 + \ldots + x_n$ es la suma de las respuestas en las n observaciones. ¿Qué distribución sigue la suma de n variables aleatorias Bernoulli i.i.d?

El estadístico Z calculado como:

$$Z = \left[\frac{\partial}{\partial\theta} l(\theta)\right]_{\theta=0} = \left[\frac{\partial}{\partial\theta} p \frac{\partial}{\partial p} l(p)\right]_{\theta=0}$$

es una medida de la diferencia entre p y p_0.

Sea el estadístico $V = \left[\left(\frac{\partial}{\partial\theta} p\right)^2 \mathcal{I}(p)\right]_{\theta=0}$ la información que Z posee sobre θ, donde $\mathcal{I}(p)$ es la información de Fisher. En estadística, la cantidad de información de Fisher se define para un parámetro p y proporciona una medida de precisión en la estimación de ese parámetro en una muestra.

Sabiendo que $\mathcal{I}(p) = -E\left[\frac{\partial^2}{\partial p^2} l(p)\right]$.

Se pide:

2. Demuestra que en el caso de la variable Bernoulli de parámetro p, la información de Fisher sobre p en una muestra de tamaño n es igual a $\frac{n}{p(1-p)}$.

La fórmula para Z y V son diferentes dependiente de la forma en la que se parametriza la diferencia entre p y p_0.

Se pide:

3. Calcular las expresiones de Z y V para las siguientes parametrizaciones de θ:

- $\theta = log(\frac{p(1-p_0)}{p_0(1-p)})$
- $\theta = p - p_0$
- $\theta = arcsin\sqrt{p} - arcsin\sqrt{p_0}$

Constraste de hipótesis

Para valores de n suficientemente grandes y pequeños valores de θ, la distribución de Z es aproximadamente Normal con media $\theta \times V$ y varianza V.

Se pide:

4. Presentar un código en R, comentado convenientemente para el cálculo de Z y V en cada una de las parametrizaciones propuestas.
5. Simular para una muestra de tamaño 1000, el contraste de hipótesis siguiente:

$$H_0 : p = 0.3$$

$$H_1 : p > 0.3$$

Cálculo del tamaño muestral

Z se usa como estadístico de contraste con nivel de significatividad α y potencia $1 - \beta$. Si Z es mayor que un valor de referencia k, entonce la hipótesis nula se rechaza con nivel de significatividad α. En ese caso, puede concluirse que la proporción es superior a la establecida en la hipótesis de partida. Los requerimientos para este test (unilateral) son:

$$P(Z \geq k|H_0 : \theta = 0) = \alpha$$

$$P(Z \geq k|H_1 : \theta = \theta_R) = 1 - \beta$$

donde θ_R es la diferencia que, si está presente, se desea detectar. Un ensayo muestral fijo satisface estos requerimientos cuando la cantidad de información V está dada por:

$$V = \left(\frac{z_{\alpha/2} + z_\beta}{\theta_R} \right)^2$$

donde z_τ denota el percentil $100(1 - \tau)$ de una distribución normal de media 0 y desviación típica 1.

Se pide:

6. Emplear la fórmula anterior para obtener una fórmula para el tamaño muestral necesario para realizar el contraste con la potencia deseada en cada una de las parametrizaciones de θ consideradas anteriormente.

Potencia y error de tipo I

En nuestro estudio, p_0 es 0.003 y el valor de p que deseamos detectar es 0.006.

Se pide:

7. Emplear los resultados del último ejercicio con nivel $\alpha = 0.025$ y potencia 0.80 para estimar el tamaño de muestra necesario en este caso particular, para cada una de las parametrizaciones del estadístico.
8. Estimar los errores tipo I y la potencia del contraste basado en Z y V para cada una de los tamaños de muestra y parametrizaciones anteriores.
9. Emplear además los contrastes siguientes:

- Test de chi-cuadrado sin corrección de continuidad. 'chisq.test(x = c(r, n - r), p = c(p0, 1 - p0), correct = FALSE)'
- Prueba exacta. Función 'binom.test(r, n, p0, alternative = "greater")' .

10. Como se ha explicado en clase, an algunas situaciones, las suposiciones necesarias para aplicar pruebas paramétricas no se cumplen, por lo que es necesario utilizar pruebas no paramétricas. Sugiere qué prueba no parámtrica podría emplearse para comparar una proporción con un valor de referencia. Escribe el código en R para realizar dicha prueba.

Referencias

- Whitehead, J., & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. Biometrics, 227-236.