Contraste de Hipótesis

Gherardo Varando

ISP - IPL
Departament d'Estadística i Investigació Operativa

Contraste de hipótesis, paramétrico

Dado un modelo paramétrico $\mathcal{M}$, dividimos los posibles parámetros Θ en dos subconjuntos, Θ_0 y Θ_1, de manera que:

$$\Theta_0 \cup \Theta_1 = \Theta \quad \Theta_0 \cap \Theta_1 = \emptyset$$

Y definimos:

- H_0: $\theta \in \Theta_0$ La hipótesis nula
- H_1: $\theta \in \Theta_1$ La hipótesis alternativa

Un contraste de hipótesis es un procedimiento que, a partir de algunas observaciones, nos dice si debemos **rechazar** o **no rechazar** la hipótesis nula.

Dadas observaciones del modelo:

$$X_1, \ldots, X_n$$

La forma habitual de construir un contraste de hipótesis es elegir una **estadística de contraste** apropiada (que es una función de las observaciones) y verificar si es **extrema**, y en ese caso rechazar la hipótesis nula H_0.

- Si $T(X_1, \ldots, X_n) > c$ entonces rechazamos H_0
- Si $T(X_1, \ldots, X_n) < c$ no rechazamos H_0

donde T es una estadística de contraste.

¿Cómo elegir el valor de c? En general, queremos elegir un valor que minimice algunos errores...

Errores en Contrastes de hipótesis

	Retener nula	Rechazar nula
H_0 verdadera	✓	Error tipo I
H_1 verdadera	Error tipo II	✓

- La **función de potencia** de un contraste se define como

$$\beta(\theta) = \mathrm{P}_\theta(T(X_1, \ldots, X_n) > c)$$

- El **tamaño** de un contraste es

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup \mathrm{P}(\text{error tipo I} | H_0)$$

- Un contraste se dice que tiene **nivel** α si su tamaño es menor o igual a α

Ejemplo: Gaussiana con varianza conocida

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

donde σ es conocida. Queremos contrastar

- $H_0 : \mu \leq 0$ por lo tanto $\Theta_0 = (\infty, 0]$
- $H_1 : \mu > 0$ por lo tanto $\Theta_1 = (0, +\infty)$

Consideremos la estadística de contraste $T = \overline{X}$ y el contraste correspondiente,

$$\text{rechazar } H_0 \text{ si } \overline{X} > c$$

Podemos obtener un contraste de nivel α si elegimos

$$c = \frac{\sigma z_\alpha}{\sqrt{n}}$$

donde, como de costumbre, z_α es el cuantil $(1 - \alpha)$ de la distribución normal estándar.

Valores p

La probabilidad de observar una estadística de contraste igual o más extrema que la estadística observada actual $T(x_1, \ldots, x_n)$ se llama **valor p**.

El valor p es el nivel más bajo al cual podemos rechazar H_0.
Una vez que calculamos el valor p de un contraste y fijamos un nivel α (por ejemplo, 0.01 o 0.05), hacemos lo siguiente:

- Rechazamos la hipótesis nula si $p \leq \alpha$.
- Mantenemos la hipótesis nula si $p > \alpha$

Valores p

La probabilidad de observar una estadística de contraste igual o más extrema que la estadística observada actual $T(x_1, \ldots, x_n)$ se llama **valor p**.

El valor p es el nivel más bajo al cual podemos rechazar H_0.
Una vez que calculamos el valor p de un contraste y fijamos un nivel α (por ejemplo, 0.01 o 0.05), hacemos lo siguiente:

- Rechazamos la hipótesis nula si $p \leq \alpha$.
- Mantenemos la hipótesis nula si $p > \alpha$

Observaciones

valor p	evidencia
< 0.01	evidencia muy fuerte en contra de H_0
0.01 - 0.05	evidencia fuerte en contra de H_0
0.05 - 0.10	evidencia débil en contra de H_0
> 0.1	poca o ninguna evidencia en contra de H_0

Advertencia

Un valor p grande no es una evidencia sólida **a favor de** H_0

No confundir el valor p con $P(H_0|datos)$. **El valor p no es la probabilidad de que la hipótesis nula sea cierta**

La Prueba de Wald

Sea $\hat{\theta}$ un estimador asintóticamente normal de θ. Y sea $\hat{se}$ el error estándar estimado de $\hat{\theta}$. La Prueba de Wald es una prueba de hipótesis para:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

El tamaño α de la **Prueba de Wald** es el siguiente:

- Rechazar H_0 si

$$\left| \frac{\hat{\theta} - \theta_0}{\hat{se}} \right| > z_{\alpha/2}$$

  donde, como es habitual, $z_{\alpha/2}$ es el cuantil $1 - \alpha/2$ de la distribución normal estándar.

- El valor p de la Prueba de Wald es

$$\text{valor p} \approx P(|Z| > |w|) = 2F_Z(-|w|)$$

donde $Z \sim N(0,1)$ es una variable aleatoria gaussiana estándar, y w es el valor observado de la estadística de Wald W,

$$w = \frac{\hat{\theta} - \theta_0}{\hat{se}}$$

Usando la Prueba de Wald, podemos responder a todas las preguntas de la siguiente manera:

- Sea $X_1, \ldots, X_n \sim Bernoulli(p)$, observaciones de Bernoulli en el lanzamiento de una moneda, ¿es la moneda justa?

$$H_0 : p = 0.5 \quad H_1 : p \neq 0.5$$

- Comparación de dos medias: consideremos dos muestras independientes de dos poblaciones con medias μ_1 y μ_2:

$$X_1, \ldots, X_m$$

$$Y_1, \ldots, Y_n$$

Consideremos la siguiente hipótesis:

$$H_0 : \mu_1 = \mu_2$$

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

La Prueba t de Student

Si los datos son Normales, es decir

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

Y queremos contrastar

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Podríamos usar la Prueba de Wald. Pero especialmente cuando el tamaño de la muestra es pequeño y σ es desconocida, es común utilizar en su lugar la prueba t de Student.

Prueba t de Student

Consideremos

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

Y

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s}$$

donde s es la desviación estándar empírica. Luego, bajo H_0, T sigue la distribución t de Student con $n-1$ grados de libertad y la prueba t de nivel α es:

- Rechazar H_0 si $|T| > t_{\alpha/2}$

- Si n es (moderadamente) grande, la prueba t es esencialmente idéntica a la Prueba de Wald
- La prueba t es exacta, bajo la suposición de distribución gaussiana

Prueba t de dos muestras

Ahora consideramos dos muestras de dos poblaciones distribuidas normalmente:

$$X_1, \ldots, X_m \sim N(\mu_1, \sigma^2)$$

$$Y_1, \ldots, Y_n \sim N(\mu_2, \sigma^2)$$

Queremos probar $H_0 : \mu_1 = \mu_2$

La estadística

$$T(X, Y) = \frac{\sqrt{\frac{n+m}{nm}}\,(\overline{X} - \overline{Y})}{s}$$

sigue una distribución t con $n + m - 2$ grados de libertad, donde s es la desviación estándar empírica combinada.

Entonces, la prueba t de dos muestras es

- Rechazar H_0 si

$$|T(X, Y)| > t_{\alpha/2}$$

donde $t_{\alpha/2}$ es el cuantil $1 - \alpha/2$ de la distribución t con $n + m - 2$ grados de libertad.

El valor p se obtiene con la siguiente fórmula,

$$\text{valor p} = 2F_T(-|T(X, Y)|)$$

donde F_T es la función de distribución acumulativa de la distribución t con $m + n - 2$ grados de libertad.

La función `t.test` realiza la prueba t de Student en R.

Prueba unilateral y prueba bilateral

Hasta ahora hemos definido la prueba bilateral, donde la hipótesis alternativa toma la forma de

$$H_1 : \theta \in (-\infty, \theta_0) \cup (\theta_0, +\infty)$$

o equivalentemente

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

De manera similar, podemos definir la prueba de Wald y la prueba t de Student para hipótesis unilaterales de la forma,

$$H_0 : \theta \leq \theta_0 \quad H_1 : \theta > \theta_0$$

o

$$H_0 : \theta \geq \theta_0 \quad H_1 : \theta < \theta_0$$

Tablas de contingencia

Una tabla de contingencia es una tabla que contiene conteos, es decir, realizaciones de una distribución multinomial.

	Tratamiento A	Tratamiento B
Recuperados	54	23
No recuperados	13	87

o

	Dado 1	Dado 2
1	54	23
2	13	87
3	23	7
⋮	⋮	⋮
6	17	15

Si consideramos los datos descritos como una tabla de contingencia, podemos hacer las siguientes preguntas,

- ¿Las columnas tienen una distribución idéntica? O, equivalentemente, ¿tienen un patrón de frecuencia idéntico? ¿Son las filas independientes de las columnas?
- ¿Las columnas siguen una distribución particular?

Este tipo de preguntas se pueden responder con la **prueba de chi-cuadrado de Pearson (χ^2)**.

En R, la función `chisq.test` realiza la prueba de chi-cuadrado de Pearson.

Prueba de bondad de ajuste

`chisq.test(x, p = p)` realiza la prueba de bondad de ajuste.

La prueba de bondad de ajuste verifica si la distribución que genera los recuentos de las celdas tiene probabilidades dadas por $p = (p_1, \ldots, p_k)$.

Podemos utilizar la prueba de bondad de ajuste chi-cuadrado para comprobar si un dado es justo, es decir, si $p = (1/6, \ldots, 1/6)$.

Podemos probar si bajo el Tratamiento A las probabilidades de recuperación o no son iguales.

$$H_0 : P(\text{Recuperados}|\text{Tratamiento A}) = 0.5$$

$$H_1 : P(\text{Recuperados}|\text{Tratamiento A}) \neq 0.5$$

Prueba de independencia

La función `chisq.test(x,y)` o `chisq.test(x)` (si `x` es una matriz o un arreglo) realiza la prueba de chi-cuadrado para la independencia, donde la hipótesis nula es que las probabilidades de las celdas son el producto de las probabilidades marginales de filas y columnas.

- Podemos probar si las distribuciones de dos dados son iguales
- Podemos probar si la probabilidad de recuperación es independiente del tratamiento recibido

$$H_0 : P(\text{Recuperación}) = P(\text{Recuperación}|\text{Tratamiento *})$$

Resultado asintótico del chi-cuadrado

Chi-cuadrado de Pearson

La estadística del chi-cuadrado de Pearson es:

$$T(X_1, \ldots, X_n) = \sum_{j=1}^{k} \frac{(X_j - np_{0,j})^2}{np_{0,j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

donde $E_j = \mathbb{E}(X_j) = np_{0,j}$ es el valor esperado de X_j bajo la hipótesis nula H_0.
Bajo H_0, la estadística de Pearson es asintóticamente distribuida como una χ^2_{k-1}, una distribución chi-cuadrado con $k-1$ grados de libertad.

Selección de modelos

Para un conjunto de datos dado, a menudo estamos interesados no solo en un solo modelo, sino en múltiples modelos candidatos.
Por ejemplo, para modelar los datos de ISI neural, hemos probado la distribución exponencial, la distribución gamma y la distribución inversa gaussiana.
A veces no podemos ver fácilmente qué modelo funciona mejor utilizando métodos descriptivos como gráficos Q-Q o histogramas y gráficos de densidad.
Ahora veremos tres métodos para realizar la selección de modelos

- Contraste de razón de verosimilitud
- Criterio de información de Akaike
- Criterio de información bayesiana

Contraste de razón de verosimilitud

Si los dos modelos estadísticos están anidados, es decir,

$$\mathcal{M}_1 \subset \mathcal{M}_2$$

Entonces podemos usar el contraste de razón de verosimilitud.

H_0 : el modelo $\mathcal{M}_1$ es suficiente para describir los datos

Utilizando la estadística,

$$\lambda = -2 \log \left(\frac{\max_{M_1} \mathcal{L}(\theta|X)}{\max_{M_2} \mathcal{L}(\theta|X)} \right)$$

Contraste de razón de verosimilitud

Supongamos que tenemos la siguiente situación:

$$H_0 : \theta \in \Theta_0 \subset \Theta$$

Podemos obtener la máxima verosimilitud en los dos subconjuntos y compararlos, para hacerlo calculamos la razón de verosimilitud.

$$\mathcal{L}_1 = \max_{\theta \in \Theta_0} \mathcal{L}(\theta) \quad \mathcal{L}_2 = \max_{\theta \in \Theta} \mathcal{L}(\theta)$$

Razón de verosimilitud

Definimos la estadística,

$$q(X) = q(X_1, \ldots, X_n) = \frac{\mathcal{L}_1}{\mathcal{L}_2}$$

- $\mathcal{L}_1 \leq \mathcal{L}_2$ y por lo tanto $q(X) \in (0, 1]$
- Los valores **pequeños** de q son **extremos**

Resultado

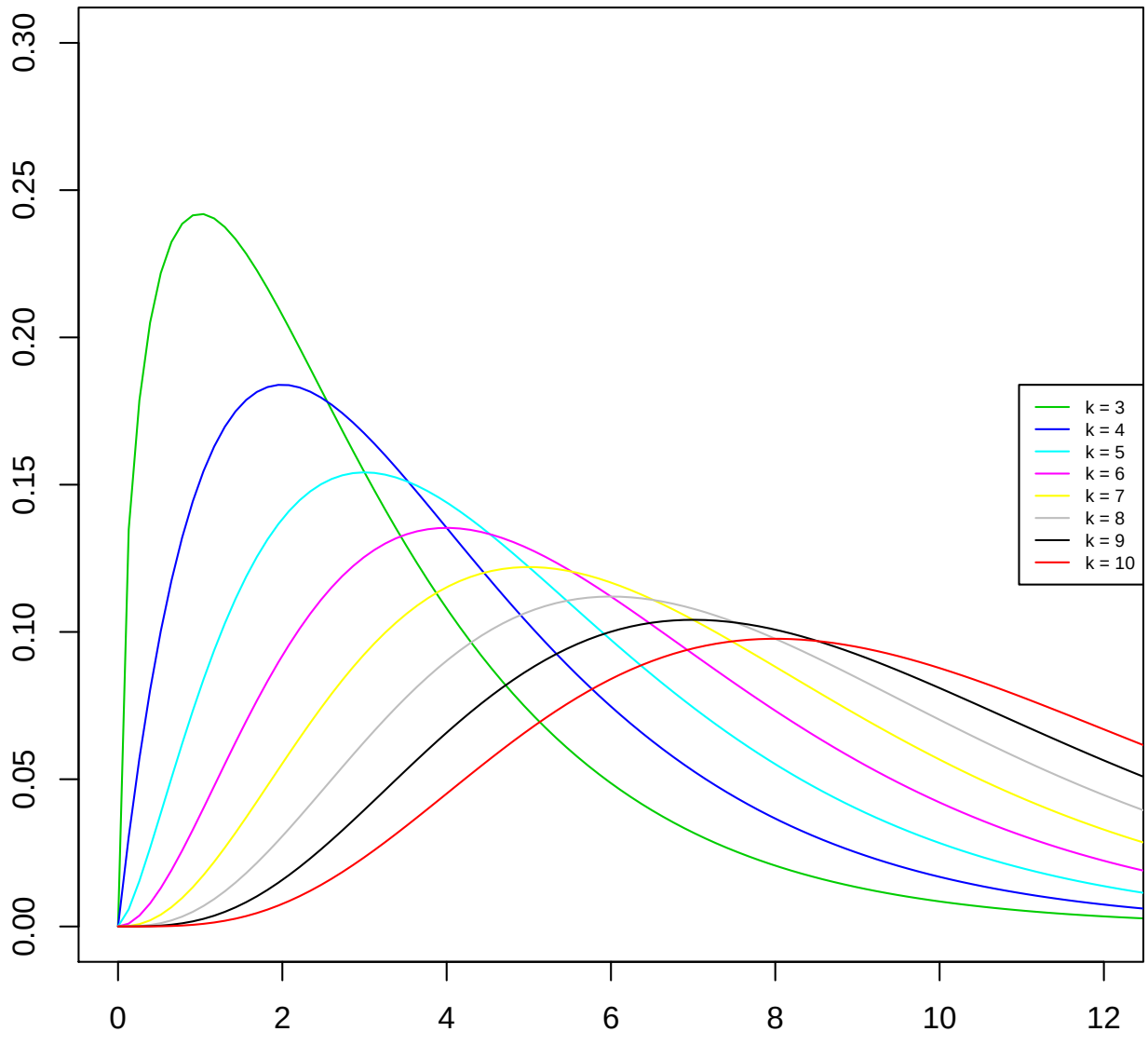
Bajo la hipótesis nula $H_0 : \theta \in \Theta_0$ tenemos,

$$\lambda(X_1, \ldots, X_n) = -2 \log\left(q(X)\right) \approx \chi^2_{d-d_0}$$

Es decir, $-2\log(q(x))$ es asintóticamente una chi cuadrado con $d - d_0$ grados de libertad, donde d es la dimensión de Θ y d_0 es la dimensión de Θ_0

Así que el contraste de razón de verosimilitud a un nivel α es:

- Rechazamos H_0 si $\lambda(X_1, \ldots, X_n) > \chi^2_{d-d_0;\alpha}$ aquí $\chi^2_{d-d_0;\alpha}$ es el cuantil superior de la distribución chi cuadrado con $d - d_0$ grados de libertad
- El valor p es $= P(\chi^2_{d-d_0} > \lambda)$


0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
12
k = 3
k = 4
k = 5
k = 6
k = 7
k = 8
k = 9
k = 10

Criterio de Información de Akaike (AIC)

El criterio de información de Akaike (AIC) se basa en una estimación de la divergencia Kullback-Leibler (KL) entre nuestro modelo $\mathcal{M}$ y el verdadero modelo de generación de datos $\mathcal{M}^*$. Si nuestro modelo tiene k parámetros, entonces,

$$AIC = -2\log(\mathcal{L}) + 2k$$

Elegimos el modelo que obtiene el valor AIC más bajo entre los modelos candidatos, aquí los modelos no necesitan ser anidados.

Criterio de Información Bayesiana (BIC)

El Criterio de Información Bayesiana (BIC) es similar al AIC, pero ahora aproxima la distribución posterior del modelo $\mathcal{M}$ dado los datos observados, utilizando una distribución previa uniforme sobre los modelos.

$$BIC = -2\log(\mathcal{L}) + k\log(n)$$

Elegimos el modelo que obtiene el valor BIC más bajo entre los modelos candidatos, aquí los modelos no necesitan ser anidados.