Técnicas de validación estadística
Bondad de ajuste

Patricia Kisbye

FaMAF

20 de mayo, 2010

Pruebas de bondad de ajuste

- Dado un conjunto de observaciones, ¿de qué distribución provienen o cuál es la distribución que mejor ajusta a los datos?
- Si se realiza una simulación de datos por computadora, ¿podemos asegurar que responden a la distribución deseada?
- Para responder a estas preguntas, existen técnicas de validación estadística.
- Técnica: Prueba de hipótesis:

  H_0) Hipótesis nula. Los datos provienen de la distribución F.

  H_1) Hipótesis alternativa. Los datos no provienen de la distribución F.

Tests de hipótesis

- Test - Prueba - Contraste. Se utilizan para
- contrastar el valor de un parámetro. Ejemplo: la media de una población es 30. Intervalo de confianza.
- comparar dos parámetros. Ejemplo: la efectividad de el medicamento A es mejor que la de B.
- contrastar los datos con una distribución teórica. Ejemplo: los datos provienen de una distribución normal.
- contrastar hipótesis de homogeneidad. Ejemplo: el porcentaje de desempleados, ¿es igual en Bs. As., Córdoba y Rosario? Tablas de contingencia
- contrastar hipótesis de independencia. Ejemplo: ser varón o mujer, ¿influye en la preferencia de un producto?

Procedimiento en una prueba de hipótesis

- Plantear
  - H_0) Hipótesis nula, con la que se contrastan los datos de la muestra.
  - H_1) Hipótesis alternativa.
- Fijar un estadístico de prueba T.
- Fijar el o los valores críticos para el estadístico de prueba, que delimitan la zona de rechazo. (valor α).
- Tomar la muestra y calcular el estadístico de prueba.
- ¿Los datos evidencian que la hipótesis nula es falsa?
  - Sí. Se rechaza la hipótesis nula.
  - No. No se rechaza la hipótesis nula.

Errores

Dado que una prueba de hipótesis se trabaja con muestras, puede haber errores:

Errores	Rechazar H_0	no Rechazar H_0
H_0 verdadera	E_I	DC
H_0 falsa	DC	E_{II}

E_I : error de tipo *I*. $P(E_I) = \alpha$.

E_{II} : error de tipo *II*. $P(E_{II}) = \beta$.

DC : Decisión correcta.

Probabilidades

Probabilidades	Rechazar H_0	no Rechazar H_0
H_0 verdadera	α	$1-\alpha$
H_0 falsa	$1-\beta$	β

- α: es la probabilidad de equivocarse rechazando una hipótesis correcta. Es controlable.
- β: es la probabilidad de equivocarse no rechazando una hipótesis falsa. No se calcula fácilmente, y puede reducirse tomando muestras grandes.
- $1-\beta$: potencia del test.
- Control sobre β:
  - aumentar α
  - aumentar el tamaño de la muestra.
- Un test deseable debe tener $1-\beta > \alpha$: la probabilidad de rechazar debería ser mayor cuando H_0 es falsa.

Pruebas de bondad de ajuste

- Aplicación: contrastar los datos con una distribución.
- Test chi-cuadrado (ji-cuadrado):
  - Es aplicable a distribuciones continuas o discretas.
  - Compara las frecuencias observadas con las frecuencias esperadas.
- Test de Kolmogorov-Smirnov:
  - Es aplicable a distribuciones continuas.
  - Compara las distribuciones acumuladas observadas y esperadas.
- Aconsejable: Utilizar chi-cuadrado para discretas, y Kolmogorov Smirnov para continuas.

Test chi-cuadrado

- No se utilizan los valores de las observaciones sino las frecuencias.
- Se compara la distribución de las frecuencias de los datos observados con las frecuencias según la distribución teórica supuesta.

f_o : frecuencia observada.

f_e : frecuencia esperada.

$$T = \sum \frac{(f_o - f_e)^2}{f_e}.$$

Implementación

- Se observan n valores de v.a. independientes igualmente distribuidas, $Y_1, Y_2, \ldots, Y_n$: Por ejemplo, se generan n valores mediante simulación, o se tienen n observaciones.
- Llamamos Y a cualquiera de las Y_i.
- Se agrupan los datos en k intervalos adyacentes que cubran el rango de la variable Y:

$$[y_0, y_1), [y_1, y_2), \ldots, [y_{k-1}, y_k).$$

  Se puede elegir $y_0 = -\infty$ o $y_k = \infty$.
- N_j: cantidad de valores que cayeron en $[y_{j-1}, y_j)$. Es la frecuencia observada.

Implementación

- p_j: Si $\hat{f}$ o $\hat{p}$ son la f.d.p. o f.p.m. a ajustar:

$$p_j = \int_{y_{j-1}}^{y_j} \hat{f}(x)\,dx \qquad \text{o} \qquad p_j = \sum_{y_{j-1} \le x_i < y_j} \hat{p}(x_i).$$

- np_j: es la frecuencia esperada. H_0): $N_j = n\,p_j$.
- Estadístico:

$$T = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}.$$

Si t es el valor observado del estadístico, se calcula

$$\text{valor } p = P_{H_0}(T \ge t)$$

que permite decidir si la hipótesis nula se rechaza o no.

Valor *p*

Si el nivel de significación del test es α,

- Valor $p < \alpha \Rightarrow$ se rechaza la hipótesis nula.
- Valor $p > \alpha \Rightarrow$ no se rechaza la hipótesis nula.
- Valor p próximo a $\alpha \Rightarrow$ se optimiza el cálculo del valor p: Simulación.

El valor p está relacionado con los valores críticos y el nivel de significación del test de la siguiente manera:
Para valores de n grandes, el estadístico

$$T = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

tiene aproximadamente una distribución χ^2.

El valor *p*

- Si se conocen todos los parámetros de la distribución, el número de grados de libertad es $k - 1$.
- En algunos casos hace falta estimar parámetros (λ en una Poisson, p en una binomial, etc.).
- Si se estiman m parámetros, el número de grados de libertad es

$$(k - 1) - m.$$

- Para estimar el valor p, puede utilizarse

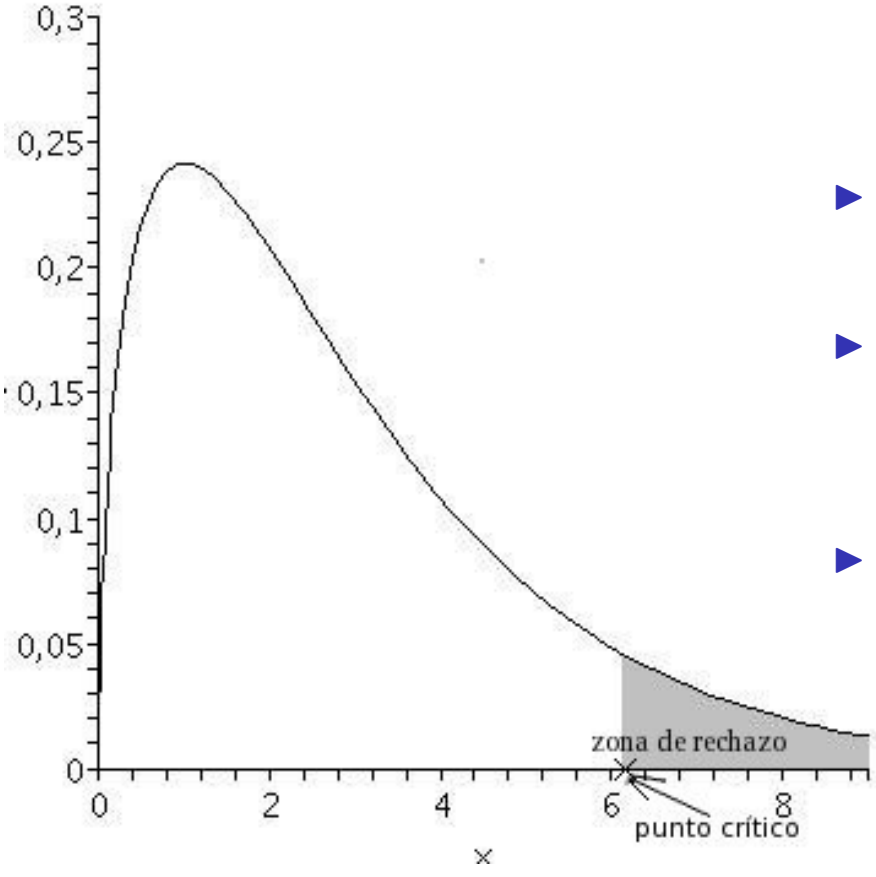
$$\text{valor } p = P_{H_0}(T \geq t) \approx P\left(\chi^2_{k-1-m} \geq t\right)$$

- Se toma como punto crítico $\chi^2_{k-m-1,1-\alpha}$:

$$P\left(\chi^2_{k-m-1} \geq \chi^2_{k-m-1,1-\alpha}\right) = \alpha.$$

Test chi-cuadrado

- Si el valor observado cae en la "zona de rechazo", se rechaza la hipótesis nula.



- valor $p < \alpha$: se rechaza la hipótesis nula.
- Equivalentemente, si el valor observado es mayor que el valor crítico, se rechaza H_0.
- Si no, no hay evidencias para rechazar H_0.

Ejemplo: tiempos entre arribos

- Se tiene el registro de $n = 219$ tiempos entre arribos, y se utiliza la prueba chi-cuadrado para ajustar a una distribución exponencial

$$\hat{F}(x) = 1 - e^{-x/0.399}, \qquad x \geq 0.$$

- Se han construido $k = 10$ intervalos, con $p_j = 0.1$.
- $np_j = 21.9$. (≥ 5).

  H_0): Los datos provienen de una distribución exponencial con media 0.399.

  H_1): Los datos no provienen de una distribución exponencial con media 0.399.

Ejemplo: tiempos entre arribos

j	Intervalo	N_j	np_j	$\frac{(N_j - np_j)^2}{np_j}$
1	$[0, 0.042)$	19	21.9	0.384
2	$[0.42, 0.089)$	28	21.9	1.699
3	$[0.089, 0.142)$	26	21.9	0.768
4	$[0.142, 0.204)$	12	21.9	4.475
5	$[0.204, 0.277)$	25	21.9	0.439
6	$[0.277, 0.366)$	14	21.9	2.850
7	$[0.366, 0.480)$	22	21.9	0.000
8	$[0.480, 0.642)$	29	21.9	2.302
9	$[0.642, 0.919)$	20	21.9	0.165
10	$[0.919, \infty)$	24	21.9	0.201
				$T = 13.283$

Ejemplo: tiempos entre arribos

- H_0: la distribución es exponencial con media 0.399.
- Dado que los parámetros son todos conocidos, se utiliza una χ^2 con $9 = 10 - 1$ grados de libertad.
- $\chi^2_{9,\,0.90} = 14.684$ es mayor que 13.283, no se rechaza la hipótesis al nivel $\alpha = 0.10$.
- Equivalentemente, valor $p \approx P(\chi^2_9 > 13.283) \sim 0.2$

valor $p > 0.10$ no se rechaza la hipótesis

- Al nivel $\alpha = 0.10$, el test no da razones para concluir que la distribución no se ajuste a una exponencial con $\lambda = 0.399$.

- $\chi^2_{9,\,0.75} = 11.389$ es menor que 13.283, se rechaza la hipótesis al nivel $\alpha = 0.25$.

Ejemplo: cantidades de demanda

Se tienen registros de cantidades de demanda de un producto, y se quiere testear el ajuste de estos datos a una distribución geométrica con $p = 0.346$.

$$P(X \leq x) = 1 - (0.654)^x, \qquad x = 1, 2, \ldots.$$

- Se han construido $k = 3$ intervalos.
- Como la distribución es discreta, los intervalos son esencialmente subconjuntos de valores de la variable.
- En este caso se han elegido:

$$I_1 = \{1\}, \; I_1 = \{2, 3\}, \; I_1 = \{4, 5, \ldots\}.$$

H_0): Los datos provienen de una distribución geométrica con $p = 0.346$.

Ejemplo: cantidades de demanda

j	Intervalo	N_j	np_j	$\frac{(N_j - np_j)^2}{np_j}$
1	$\{1\}$	59	53.960	0.471
2	$\{2, 3\}$	50	58.382	1.203
3	$\{4, 5, \dots\}$	47	43.658	0.256
				$T = 1.930$

- Los parámetros de la distribución son conocidos.
- Se utiliza una χ^2 con $2 = 3 - 1$ grados de libertad.
- $\chi^2_{2,0.90} = 4.605$. No se rechaza la hipótesis nula a un nivel de $\alpha = 0.10$.
- Equivalentemente, valor $p \approx P(\chi^2_2 > 1.930) \sim 0.6$
- valor $p > 0.10$, no se rechaza la hipótesis nula.

Ejemplo

Una v.a. puede tomar los valores 1,2,3,4,5. Testear la hipótesis que estos valores son equiprobables.

$$H_0) \;\; p_i = 0.2, \text{ para cada } i = 1, \ldots, 5.$$

- Se toma una muestra de tamaño $n = 50$.
- Se obtienen los siguientes valores:

$$N_1 = 12, \quad N_2 = 5, \quad N_3 = 19, \quad N_4 = 7, \quad N_5 = 7.$$

- $np_i = 50 \cdot 0.2 = 10$ para cada $i = 1, \ldots, 5$.

Ejemplo

- Estadístico:

$$\begin{aligned} T &= \frac{(12-10)^2+(5-10)^2+(19-10)^2+(7-10)^2+(7-10)^2}{10} \\ &= 12.8 \end{aligned}$$

- valor $p \approx P(\chi^2_4 > 12.8) = 0.0122$.
- Para este valor de p, se rechaza la hipótesis que todos los valores son igualmente probables.

Simulación del valor p

- Si el valor p es próximo a α significa que el valor observado t es próximo al valor crítico.
- ¿Se rechaza o no se rechaza?
- Es conveniente tener una estimación más exacta para p.
- Método: simulación.

Implementación en el caso discreto

- H_0: $P(Y = y_j) = p_j$, para todo $j = 1, \ldots, k$.
- Generar n v.a. independientes con probabilidad de masa p_i, $1 \leq i \leq k$.
- Evaluar el estadístico T.
- Repetir el procedimiento r veces y calcular la proporción de valores mayores que t.

Implementación

- Generar $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_n^{(1)}$ independientes, que tomen los valores $1, 2, \ldots, k$ con probabilidad de masa

$$P\left(Y_i^{(1)} = j\right) = p_j.$$

- $N_j^{(1)} = \#\{i \mid Y_i^{(1)} = j\}$.
- Evaluar el estadístico T para este conjunto de valores:

$$T^{(1)} = \sum_{i=1}^{k} \frac{(N_i^{(1)} - np_i)^2}{np_i}$$

- Repetir el procedimiento r veces, para obtener $T^{(1)}, T^{(2)}, \ldots, T^{(r)}$.

$$\text{valor } p = P_{H_0}(T \geq t) \approx \frac{\#\{i \mid T_i \geq t\}}{r}$$

Estimación del valor p en el caso continuo

Implementación en el caso continuo

- H_0: Las v.a. $Y_1, Y_2, \ldots, Y_n$ tienen distribución continua F.
- Particionar el rango de $Y = Y_j$ en k intervalos distintos:

$$[y_0, y_1), [y_1, y_2), \ldots, [y_{k-1}, y_k),$$

- Considerar las n v.a. discretizadas $Y_1^d, Y_2^d, \ldots, Y_n^d$ dadas por

$$Y_j^d = i \qquad \text{si } Y_i \in [y_{j-1}, y_j).$$

- La hipótesis nula es entonces
  $H_0)\ P(Y_j^d = i) = F(y_i) - F(y_{i-1}), \qquad i = 1, \ldots, k.$
- Proceder ahora como en el caso discreto.
- Es aconsejable utilizar el test de Kolmogorov-Smirnov.