Test de Kolmogorov-Smirnov

Patricia Kisbye

FaMAF

3 de junio, 2010

Test de Kolmogorov-Smirnov

El test chi-cuadrado 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.

Test de Kolmogorov Smirnov

- Inconveniente: No es sencillo construir los intervalos a partir de las probabilidades.
- Se pierde información al agrupar los datos en intervalos.
- Aconsejable: Utilizar el test de Kolmogorov-Smirnov.

Test de Kolmogorov Smirnov

- Compara las funciones de distribución empírica de la muestra y la que se desea contrastar.
- Es aplicable a distribuciones continuas.
- Para distribuciones discretas, los valores críticos no están tabulados.
- Para distribuciones continuas, los valores críticos están tabulados para:
  - distribuciones con parámetros especificados,
  - algunas distribuciones con parámetros no especificados (normal, Weibull, gamma, exponencial).

Aplicación del test K-S

- Observar $Y_1, Y_2, \ldots, Y_n$ y considerar la distribución empírica

$$F_e(x) = \frac{\#\{i \mid Y_i \leq x\}}{n}.$$

- $F_e(x)$: proporción de valores observados menores o iguales a x.
- Hipótesis nula: $F_e(x)$ es "cercana" a $F(x)$.
- Estadístico de Kolmogorov-Smirnov

$$D \equiv \max_x |F_e(x) - F(x)|, \qquad -\infty < x < \infty.$$

Implementación

- Ordenar los datos observados $Y_1 = y_1,\ Y_2 = y_2, \ldots, Y_n = y_n$ en orden creciente:

$$y_{(j)} = j - \text{ésimo valor más pequeño}$$

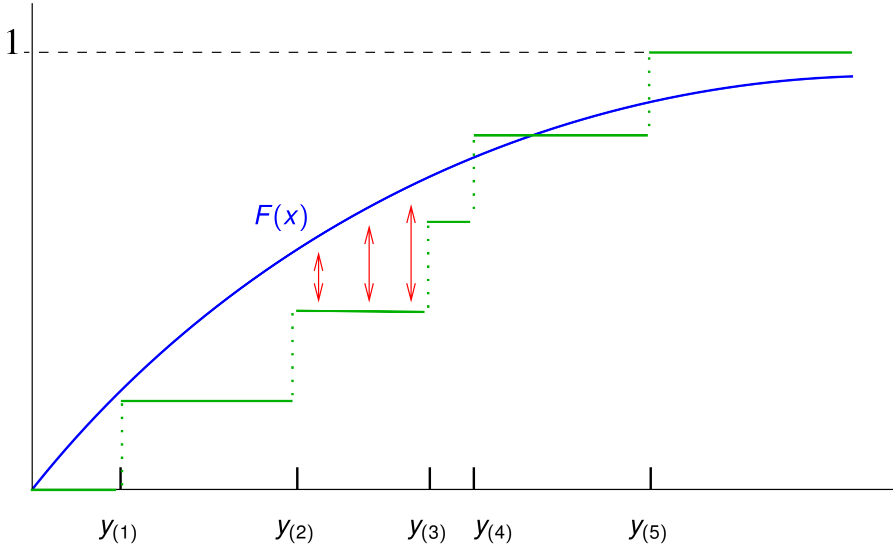
$$y_{(1)} < y_{(2)} < \cdots < y_{(n)}.$$

- Por ejemplo: $y_1 = 3$, $y_2 = 5$, $y_3 = 1$ y $n = 3$, entonces

$$y_{(1)} = 1,\ y_{(2)} = 3,\ y_{(3)} = 5.$$

Distribución empírica $\Rightarrow$ $F_e(x) = \begin{cases} 0 & x < y_{(1)} \\ \frac{1}{n} & y_{(1)} \leq x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & y_{(j)} \leq x < y_{(j+1)} \\ \vdots & \\ 1 & y_{(n)} \leq x \end{cases}$

Gráficamente



Estadístico de Kolmogorov-Smirnov

$$\boxed{D \equiv \sup_{-\infty<x<\infty} |F_e(x) - F(x)|}$$

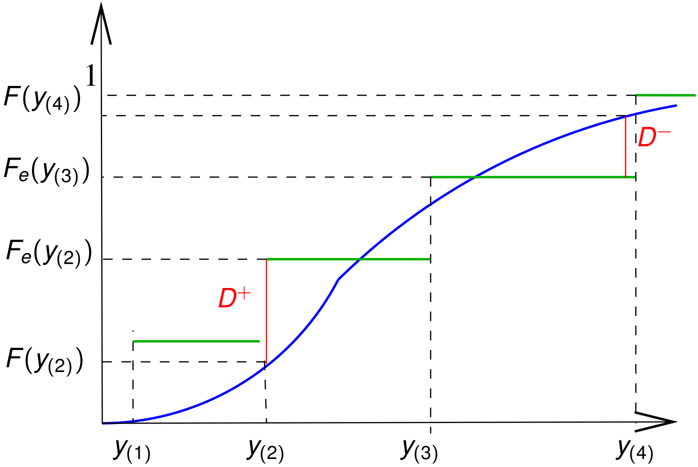
- Podemos considerar las diferencias $F_e(x) - F(x)$ y $F(x) - F_e(x)$ y analizar sus valores máximos (supremos).

$$D^+ = \sup_{-\infty<x<\infty} \{F_e(x) - F(x)\}\,, \qquad D^- = \sup_{-\infty<x<\infty} \{F(x) - F_e(x)\}\,.$$

- $F_e(y_{(n)}) = 1$. Por lo tanto, $D^+ \geq 0$.
- $F_e(x) = 0$ si $x < y_{(1)}$, por lo que $D^- \geq 0$.

$$\boxed{D = \max\{D^+, D^-\}}$$

El estadístico D



Cálculo de D

Notemos que:

- D^+ se alcanza en el límite inferior de algún intervalo, ya que $F(x)$ es creciente y $F_e(x)$ es constante en $[y_{(j-1)}, y_{(j)})$:

$$\boxed{D^+ = \max_{1\leq j\leq n}\left\{F_e\left(y_{(j)}\right) - F\left(y_{(j)}\right)\right\} = \max_{1\leq j\leq n}\left\{\left(\frac{j}{n}\right) - F\left(y_{(j)}\right)\right\}}$$

- D^- es el límite por izquierda calculado en el extremo derecho de algún intervalo, ya que $F_e(x)$ es discontinua en tal punto:

$$F_e\left(y_{(j)}\right) = \frac{j}{n} = F_e\left(y_{(j)} - \epsilon\right) + \frac{1}{n}, \qquad \epsilon \text{ pequeño.}$$

$$\boxed{D^- = \max_{1\leq j\leq n}\left\{F\left(y_{(j)}\right) - F_e\left(y_{(j-1)}\right)\right\} = \max_{1\leq j\leq n}\left\{F\left(y_{(j)}\right) - \frac{j-1}{n}\right\}}$$

Estadístico de Kolmogorov-Smirnov

$$\boxed{D = \max_{1 \le j \le n} \left\{ \frac{j}{n} - F(y_{(j)}),\ F(y_{(j)}) - \frac{j-1}{n} \right\}} \leftarrow \text{Estadístico de K-S}$$

- Elegir un grado de significación (nivel de rechazo) α.
- Tomar la muestra y ordenar los datos observados.
- Calcular el estadístico D en los datos observados.
- Valor observado: $D = d$.
- Calcular el valor $p = P_F(D \geq d)$.
  - valor $p < \alpha$: se rechaza H_0.
  - valor $p > \alpha$: no se rechaza H_0.
- ¿Cómo calcular el valor p?
- ¿Cuál es la distribución del estadístico D?

Estimación del valor p

$P_F\,(D \geq d)$ no depende de la distribución F.

- El estadístico D depende de las n observaciones $Y_1, Y_2, \ldots, Y_n$:

$$D = \sup_x |F_e(x) - F(x)| = \sup_x \left| \frac{\#\{i \mid Y_i \leq x\}}{n} - F(x) \right|$$

- Si Y tiene distribución F entonces

$$F(Y) \sim \mathcal{U}(0,1).$$

- Como F es una función creciente, entonces

$$Y_i \leq x \qquad \text{implica} \qquad F(Y_i) \leq F(x).$$

Estimación del valor p

$$D = \sup_x \left| \frac{\#\{i \mid F(Y_i) \leq F(x)\}}{n} - F(x) \right|$$

Equivalentemente, se puede reemplazar

- $F(Y_i)$ por U_i, v.a. uniformemente distribuida en $(0, 1)$, y
- $F(x)$ por $y \in [0, 1]$.

$$\boxed{D = \sup_{0 \leq y \leq 1} \left| \frac{\#\{i \mid U_i \leq y\}}{n} - y \right|}$$

Estimación del valor p

$$\boxed{D = \sup_{0 \le y \le 1} \left| \frac{\#\{i \mid U_i \le y\}}{n} - y \right|}$$

- Esta expresión no depende de la distribución F.

$$\text{valor } p = P_F(D \ge d) = P_U(D \ge d), \qquad U \sim \mathcal{U}(0,1).$$

- Puede estimarse mediante simulación:
  - Generar n números aleatorios U_i, $1 \le i \le n$,
  - Evaluar D y comparar con el valor observado d de la muestra original.

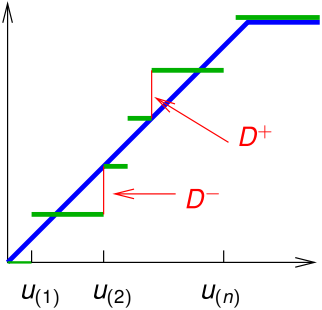
$$\sup_{0 \le y \le 1} \left| \frac{\#\{i \mid U_i \le y\}}{n} - y \right| \ge d$$

  - Repetir el procedimiento r veces.
- Se estima el valor p como la proporción de veces que se cumple la desigualdad $D \ge d$.

Estimación del valor p

$$\sup_{0 \le y \le 1} \left| \frac{\#\{i \mid U_i \le y\}}{n} - y \right| \ge d$$

- Para calcular este supremo, procedemos como para el cálculo de d.



- Ordenar $u_{(1)}, \ldots, u_{(n)}$.
- Calcular

$$\max_{1 \le j \le n} \left\{ \frac{j}{n} - u_{(j)}, u_{(j)} - \frac{j-1}{n} \right\}$$

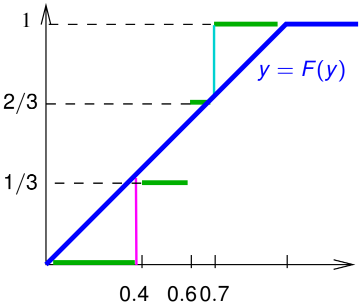
Ejemplo

- Si $n = 3$ y $U_1 = 0.7$, $U_2 = 0.6$, $U_3 = 0.4$, entonces

$$U_{(1)} = 0.4, \qquad U_{(2)} = 0.6, \qquad U_{(3)} = 0.7,$$

y el valor D para este conjunto de datos es

$$D = \max\left\{\frac{1}{3} - 0.4,\ \frac{2}{3} - 0.6,\ 1 - 0.7, 0.4, 0.6 - \frac{1}{3}, 0.7 - \frac{2}{3}\right\} = 0.4$$



Ejemplo

- Se quiere probar la hipótesis que una determinada distribución es exponencial con media 100

$$F(x) = 1 - e^{-x/100}.$$

Los valores ordenados para una muestra de tamaño 10 para esta distribución son:

$$66,\ 72,\ 81,\ 94,\ 112,\ 116,\ 124,\ 140,\ 145,\ 155,$$

¿qué conclusión puede obtenerse?

Ejemplo

j	valores	$F(j/n)$	$\frac{j}{n} - F\left(\frac{j}{n}\right)$	$\frac{j-1}{n} - F\left(\frac{j}{n}\right)$
1	66	0,48	-0,38	0,48
2	72	0,51	-0,31	0,41
3	81	0,56	-0,26	0,36
4	94	0,61	-0,21	0,31
5	112	0,67	-0,17	0,27
6	116	0,69	-0,09	0,19
7	124	0,71	-0,01	0,11
8	140	0,75	0,05	0,05
9	145	0,77	0,13	-0,03
10	155	0,79	0,21	-0,11
			$d = 0,48315$	

- Calcular el valor p mediante simulaciones.
- Si el p valor es 0.012, se rechaza la hipótesis nula.

Pruebas de bondad de ajuste si hay parámetros no especificados

Caso discreto: test chi-cuadrado

Dadas n observaciones, $Y_1, \ldots, Y_n$, éstas se agrupan en k intervalos distintos. La hipótesis nula está dada por

$$H_0)\ P(Y_i = j) = p_j, \text{ para } 1 \leq j \leq k,\ i = 1 \ldots n..$$

- En algunos casos se tiene alguna hipótesis sobre la forma de la distribución pero no sobre los parámetros de la misma: media, desviación estándar, varianza, etc.
- Esto puede implicar que se desconozca p_j:

$$P(Y = j) = \frac{e^{-\lambda}\lambda^j}{j!} \quad \Rightarrow \quad ¿\lambda \text{ desconocido?}$$

- En este caso, se estiman el o los parámetros desconocidos a partir de la muestra.

El caso discreto

- A partir de estas estimaciones, se obtienen las probabilidades estimadas: $\hat{p}_j$.
- El estadístico es el siguiente:

$$T = \sum_{j=1}^{k} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j}$$

  - N_j: cantidad de observaciones en el j-ésimo intervalo.
  - $\hat{p}_j$: probabilidad estimada, según H_0, que Y_j caiga en la región j.
- Si el valor observado del estadístico es t, y se han debido estimar m parámetros:

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

Ejemplo

En un período de 30 días se registraron 6 días sin accidentes, 2 con un accidente, 1 con dos accidentes, 9 con 3 accidentes, 7 con 4 accidentes, 4 con 5 accidentes y 1 con 8 accidentes. Realizar una prueba de hipótesis para determinar si el número de accidentes sigue una distribución de Poisson.

- Estimamos la media λ de la distribución:
  número de accidentes =
  $6 \cdot 0 + 2 \cdot 1 + 1 \cdot 2 + 9 \cdot 3 + 7 \cdot 4 + 4 \cdot 5 + 1 \cdot 8 = 87.$

$$\hat{\lambda} = \frac{\text{número de accidentes}}{\text{total de días}} = \frac{87}{30} = 2.9$$

$$\hat{p}_{j+1} = P(Y = j) = e^{-2.9} \frac{(2.9)^j}{j!}, \quad j = 0, 1, 2, \ldots$$

Ejemplo

- Se establecen los k intervalos. Elegimos $k = 6$:

$$I_1 = \{0\} \qquad I_3 = \{2\} \qquad I_5 = \{4\}$$
$$I_2 = \{1\} \qquad I_4 = \{3\} \qquad I_6 = \{5, 6, 7, \dots\}$$

$$\hat{p}_1 = 0.0500 \qquad \hat{p}_2 = 0.1596 \qquad \hat{p}_3 = 0.2312$$
$$\hat{p}_4 = 0.2237 \qquad \hat{p}_5 = 0.1622 \qquad \hat{p}_6 = 0.1682$$

- Frecuencias observadas: $N_1 = 6$, $N_2 = 2$, $N_3 = 1$, $N_4 = 9$, $N_5 = 7$, $N_6 = 5$.
- Frecuencias esperadas: $30\,\hat{p}_j$, $1 \leq j \leq 6$.
- Estadístico:

$$T = \sum_{j=1}^{6} \frac{(N_j - 30\,\hat{p}_j)^2}{30\,\hat{p}_j} = 19.887.$$

Valor p

- El valor observado del estadístico es $t = 19.887$.
- Como se estimó 1 parámetro, y se consideraron 6 intervalos, se estima el valor p utilizando una distribución χ^2 con $6 - 1 - 1 = 4$ grados de libertad:

$$\text{valor } p \approx P(\chi^2_4 > 19.887) = 0.0005.$$

- Conclusión: se rechaza la hipótesis nula.

Simulación para determinar el valor *p*

- La hipótesis nula no especifica completamente la distribución.
- El procedimiento es similar al caso anterior, pero los parámetros deben estimarse nuevamente en cada simulación.

Valor p con parámetros estimados

- El modelo

- H_0) Los datos de la muestra $Y_1, Y_2, \ldots, Y_n$ provienen de una distribución determinada, salvo por un conjunto de parámetros desconocidos $\theta_1, \ldots, \theta_m$.

Primer paso

- $\hat{\theta}_j$: estimación de θ_j a partir de la muestra, $j = 1, 2, \ldots, m$.
- $\hat{p}_j$: si la distribución tiene parámetros $\hat{\theta}_1, \ldots, \hat{\theta}_m$.
- Estadístico T:

$$T = \sum_{j=1}^{k} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j}$$

- $t \leftarrow$ valor observado del estadístico T.

Valor p con parámetros estimados

<u>Simulación</u>

- Objetivo: estimar el valor p.
- $\hat{F}$: distribución propuesta en H_0, con los parámetros estimados según la muestra.

El procedimiento consiste en repetir r veces los siguientes pasos:

1. Generar $Y_1, \dots, Y_n \sim \hat{F}$.
2. Calcular $N_j = \#\{i \mid Y_i \in I_j\}$, $j = 1, \dots, k$.
3. $\hat{\theta}_{1,sim}, \dots, \hat{\theta}_{m,sim}$: estimaciones de los parámetros a partir de los valores Y_i generados.
4. $\tilde{p}_j(sim)$, probabilidades si la distribución tiene parámetros $\hat{\theta}_{1,sim}, \dots, \hat{\theta}_{m,sim}$.
5. Calcular T^*:

$$T^* = \sum_{j=1}^{k} \frac{(N_j - n\tilde{p}_j)^2}{n\tilde{p}_j}$$

Valor p con parámetros estimados

Luego de r pasos se han obtenido r valores para T^*:

$$T_1^*, T_2^*, \ldots, T_r^*.$$

$$\boxed{\text{valor } p \approx \frac{\#\{j \mid T_j^* \geq t\}}{r}}$$

Ejemplo

- Parámetro estimado: $\hat{\lambda} = 2.9$.
- Valor del estadístico según la muestra: $t = 19.887$
- Simulación:
  1. Generar 30 v.a. Poisson con media 2.9.
  2. $\hat{\lambda}_{sim}$: estimación de λ según esta muestra.
  3. p_i^*: Probabilidad de tomar el valor i según una Poisson de parámetro $\hat{\lambda}_{sim}$.
  4. Calcular T^*.
  5. valor p: proporción de valores de T^* mayores a 19.887.

Test de Kolmogorov-Smirnov si hay parámetros no especificados

Caso continuo

H_0): Las v. a. $Y_1, \ldots, Y_n$ provienen de una distribución F con parámetros desconocidos $\theta_1, \ldots, \theta_m$.

- Tomar una muestra $Y_1, \ldots, Y_n$.
- Estimar los parámetros a partir de la muestra: $\hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$.
- Calcular el estadístico a partir de la distribución con parámetros estimados:
$$D = \sup_x \left| F_e(x) - F_{\hat{\Theta}}(x) \right| .$$
- $d \leftarrow$ valor de D observado.
$$\text{valor } p \approx P_{F_{\hat{\Theta}}}(D \geq d) = P_U(D \geq d).$$
- Este valor sobreestima el valor de p.

Simulación del valor p

- Si $p < \alpha$, se rechaza la hipótesis nula.
- Si está próximo o es mayor que α, se optimiza la estimación del valor p.
- Optimización: Luego de calcular d, a partir de la muestra:
  1. Generar $Y_1, \ldots, Y_n$ según la distribución $F_{\hat{\Theta}}$.
  2. $\hat{\Theta}_{sim} = (\hat{\theta}_{1,sim}, \ldots, \hat{\theta}_{m,sim})$: estimación de los parámetros según los datos simulados.
  3. $F_{e,sim}$: distribución empírica de los datos simulados.
  4. Calcular el estadístico D^*:

$$D^* = \sup_x \left| F_{e,sim}(x) - F_{\hat{\Theta}_{sim}}(x) \right|$$

.

- Repetir el procedimiento r veces, para obtener $D_1^*, \ldots, D_r^*$:

$$\boxed{\text{valor } p \approx \frac{\#\{j \mid D^* \geq d\}}{r}}$$