ESTAD_ T27. CONTRASTES de BONDAD de AJUSTE. CONTRASTE χ^2 de PEARSON. CONTR. de KOLMOGOROV_SMIRNOV. CONTRASTES de NORMALIDAD.

0_ INTRODUCCIÓN.

Antes de comentar a hablar de los contrastes de la bondad del ajuste se hace necesario comentar brevemente los contrastes no paramétricos: en qué consisten, qué los diferencian de los contrastes paramétricos, por qué y para qué se utilizan, ...

2 _ CONTRASTE χ^2 de PEARSON

Supongamos que la muestra que se quiere contrastar puede dividirse en r clases o categorías (A_1 — A_r) mutuamente excluyentes (cuantitativas → discretas o continuas, cualitativas → ordinal ó nominal), cuya unión configure el campo de variación de la variable poblacional.

Hipótesis nula

$$\begin{cases} H_0: X \text{ procede de una distrib. de probab. } P \\ H_1: H_0 \text{ es falsa} \end{cases}$$

Estadístico del contraste

Categoría	Frec. observadas	Frec. Esperadas	$\frac{(O_i-E_i)^2}{E_i}$
A_1	O_1	E_1	
A_2	O_2	E_2	
⋮	⋮	⋮	
A_r	O_r	E_r	
SUMA	n		Estad χ^2

Frec. Observadas → $O_i = n_i \neq$ nº elementos de la categoría A_i presentes en la muestra. $\Sigma O_i = n \equiv$ tamaño muestral

Frec. Esperadas → $E_i = n \cdot p_i$ / nº elementos de la categoría A_i que se espera en la muestra si se considera H_0 cierta.

Cuando la distrib. de probab. de H_0, P, se puede establecer la probab. de cada categoría:

$$P(A_i) = p_i \quad / \ \Sigma p_i = 1.$$

Al ser X una m.a.s., la probab. de la muestra sigue una distrib. multinomial (con indep.):

$$P(O_1 \ldots O_r) = \frac{n!}{O_1! \ldots O_r!} p_1^{O_1} \ldots p_r^{O_r}$$

donde $O_i \to B(n, p_i)$ / $E[O_i] = np_i = E_i$

1.- CONTRASTES de BONDAD del AJUSTE

En los contrastes paramétricos, ~~las contra~~ las hipótesis se refieren a valores concretos de los parámetros de la población desconocidos, que identifican la distribución de probabilidad de la población.

En todos los contrastes paramétricos se parte de la premisa de que se conoce la distribución de probabilidad de la población de partida de la cual se ha obtenido la muestra.

Contr. Paramétricos: distrib. conocida (familia), pero no determinada.

En la práctica, se tiene una muestra obtenida por m.a.s., y se desconoce la distrib. de probabilidad de la población de la que procede.

El objetivo de los contrastes de bondad de ajuste es verificar si la muestra procede de una población con una determinada distribución de probabilidad.

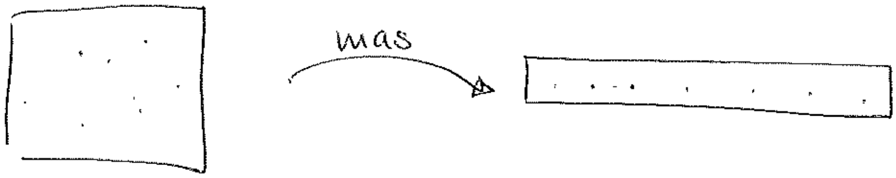
La hipótesis nula postula que la muestra se ajusta a una determinada distribución, que el ajuste es "bueno" en el sentido de que la información muestral no evidencia diferencias significativas (mayores a las atribuibles a la selección por mas de la muestra) que obliguen a rechazar la hipótesis nula.

La hipótesis alternativa postula el enunciado contrario, es decir, que el ajuste no es bueno.

Antes de continuar, hay que resaltar que la hipótesis nula que se formule no puede chocar de lleno con la información muestral, por lo que es conveniente previamente observar la muestra (histogramas de frecuencias) y conocer sus momentos muestrales (media y varianza).

Por ejemplo, en una distrib. de Poisson media y varianza coinciden, por lo que $\bar{x}$ y S^2 no pueden ser muy distintos si queremos formular que la muestra procede de una Poisson.

Población, $\xi \to F(x)$ Muestra, $X = \{x_1 \dots x_n\}$



H_0: La muestra procede de una población con función de distribución $F_0(x)$.

H_1: La muestra no procede de la población con función de distribución $F_0(x)$.

$$H_0 : F(x) = F_0(x)$$
$$H_1 : F(x) \neq F_0(x)$$

Los contrastes de bondad de ajuste más utilizados son:

- Test χ^2 de bondad de ajuste
- Test G^2 de razón de verosimilitud
- Test de Kolmogorov - Smirnov
- Test de normalidad de Lilliefors.
- Test de normalidad de Shapiro - Wilk

2. CONTRASTE χ^2 de PEARSON

Es el más antiguo y conocido de los tests de bondad de ajuste, y fue introducido por Pearson en 1900, para utilizarlo con datos nominales.

Compara las frecuencias muestrales con las probabilidades teóricas en caso de ser H_0 cierta.

Se pueden considerar dos casos:

a) Todos los parámetros de la distribución de la población son conocidos, o se han estimado previamente a partir de una muestra distinta a la del contraste.

b) Uno o varios parámetros de la distribución no son conocidos, y hay que estimarlos a partir de la misma muestra del contraste.

a) Test χ^2 cuando no se estiman parámetros:

Supongamos la v.a. ξ con distribución de probabilidad conocida, $F_0(x)$ (bajo H_0 cierta), cuyo campo de variación se puede particionar* en r clases o categorías mutuamente excluyentes, que pueden ser intervalos o escala nominal u ordinal.

Bajo H_0, podemos conocer las probabilidades p_i, de que una observación pertenezca a la categoría A_i.

$$p_i = P(\xi \in A_i) \quad , \quad i = 1 \ldots r$$

$$\sum_{i=1}^{r} p_i = 1$$

En la muestra, podemos observar el nº de observaciones que caen en cada categoría A_i.

$$O_i \equiv \text{nº observaciones que } \xi \in A_i = n_i$$

$$\sum_{i=1}^{r} O_i = \sum_{i=1}^{r} n_i = n$$

---

*Partición: $\Omega = A_1 \cup \ldots \cup A_r \ / \ A_i \cap A_j = \emptyset \ , \forall i \neq j$.

mas ⟹ observaciones independientes ⟹ bajo H_0 cierta, la (ni observaciones de la categoría A_i), probabilidad de la muestra viene dada por una distribución multinomial a través de la expresión:

$$P(n_1, \dots, n_r) = \frac{n!}{n_1! \cdot \dots \cdot n_r!} p_1^{n_1} \cdot \dots \cdot p_r^{n_r}$$

donde cada n_i tiene una distrib. marginal $B(n, p_i)$, con valor esperado:

$$E(O_i) = E(n_i) = np_i = E_i \equiv \text{frec. esperada de la categoría } A_i, \text{ bajo } H_0 \text{ cierta.}$$
$$i = 1 \dots r$$

Debido al muestreo, es lógico que los valores de las frecuencias observadas O_i y los de las frecuencias esperadas E_i no coinciden, pero para tamaños muestrales suficientemente grandes, si H_0 es cierta, las discrepancias no serán muy grandes.

Discrepancia entre la frec. observada y la frec. esperada para cada categoría:

$$O_i - E_i = n_i - n \cdot p_i \quad , i = 1 - r$$

Para evitar la posible compensación de discrepancias positivas con las discrepancias negativas, elevamos al cuadrado.

Para evitar la influencia de grandes discrepancias entre valores grandes, se pondera cada discrepancia por la inversa de la frecuencia esperada

---

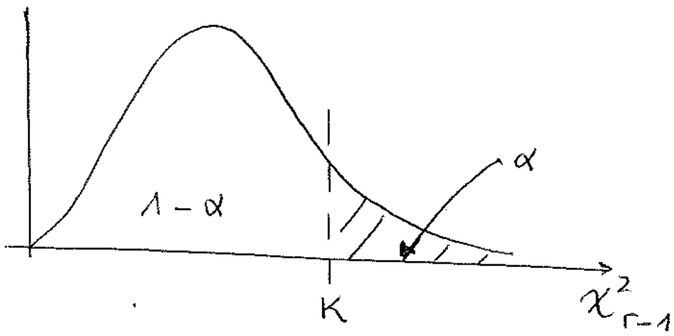
Una discrepancia grande podría llevar a rechazar H_0, aunque la categoría donde se haya dado sea poco probable.

No es lo mismo $(O_i - E_i)^2 = 25$ para $E_i = 500$ que para $E_i = 5$.

El estadístico propuesto por Pearson es la suma de los chicuadrados de las discrepancias ponderados por las inversas de las esperadas.

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \longrightarrow \chi^2_{r-1}$$

Bajo H_0 cierta, se distribuye asintóticamente según una χ^2 con $K-1$ grados de libertad.



$$\alpha = P(\chi^2_{r-1} \geq K /_{H_0 \text{ cierta}}) \Rightarrow K \text{ (tablas)}$$

Regla de decisión: $\begin{cases} \chi^2 > K & \Rightarrow \text{rechazo } H_0 \\ \chi^2 \leq K & \Rightarrow \text{acepto } H_0 . \end{cases}$

OBSERVACIONES:

1 _ Al ser la distribución de χ^2 asintótica, se tiene que verificar $E_i > 5$ para todas las categorías. En caso contrario, agrupar. Reagrupamiento $\Rightarrow$ reducción del nº de grados de libertad.

2 _ El test χ^2 puede utilizarse para distrib. continuas y discretas.

3 _ Para mejor ~~ar~~ calidad de la ~~dis~~ aproximación a través de la distrib. asintótica, es conveniente construir las categorías para que tengan probabilidades aprox. iguales.

---

la distribución se basa en aproximaciones de distrib. binomiales a normales → CRAMER, 1963
ROHATGI, 1996

El grado de libertad se pierde por la restricción implícita $\sum_{i=1}^{r} n_i = n$

b) Test χ^2 cuando se estiman parámetros:

Cuando la distrib. de la población no está totalmente especificada, se pueden formular las hipótesis:

$$\begin{cases} H_0: F(x) = F_0(x;\theta_1 \dots \theta_K) \\ H_1: F(x) \neq F_0(x;\theta_1 \dots \theta_K) \end{cases}$$

donde $\theta_1 \dots \theta_K$ son parámetros desconocidos.

Fisher demostró que si previamente se estiman los parámetros θ_j (máxima verosimilitud) con la misma información muestral del contraste, el estadístico

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \xrightarrow[\text{distrib. asintótica}]{} \chi^2_{r'-K-1}$$

donde r' = nº categorías (reagrupadas, si procede)
K = nº parámetros a estimar

Categorías	Frec. observadas $O_i = n_i$	Frec. esperadas $E_i = np_i$	Discrepancias $O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
A_1				
⋮				
$A_{r'}$				
Suma	n	n		χ^2

* Para calcular $p_i \to p_i = P(\xi \in A_i)$, acudir a las tablas de la distrib. ó a la función de cuantía / densidad

* Si alguna categoría A_i / $E_i \leq 5 \Rightarrow$ reagrupar $\Rightarrow$ menos grados de libertad en la distrib. asintótica

3_ CONTRASTE de KOLMOGOROV-SMIRNOV

En 1933, Kolmogorov y Smirnov propusieron este test de bondad de ajuste para distribuciones continuas y aplicable a muestras pequeñas.

Un test de bondad de ajuste parte de una muestra aleatoria procedente de una población desconocida y pretende contrastar la hipótesis nula de que esa función de distribución tiene una forma especificada.

H_0: La m.a.s. procede de una población con f. distrib. $F_0(x)$.

Si la hipótesis nula es cierta, las diferencias entre $F_n(x)$ y $F_0(x)$ para una muestra de tamaño suficientemente grande, no serán significativas.

La función de distribución empírica de la muestra es:

$$F_n(x) = \frac{N(x)}{n} \quad / \quad N(x) = \text{nº observaciones muestrales} \leq x .$$

Para medir la diferencia entre ambas funciones, K y S propusieron la distancia máxima vertical entre sus gráficas correspondientes. Esa medida viene dada por el estadístico de Kolmogorov-Smirnov:

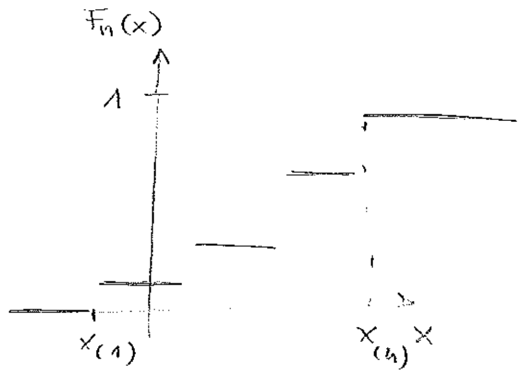
$$D_n = \max_{-\infty < x < +\infty} |F_n(x) - F_0(x)|$$

que tiene las siguientes propiedades:

$X = \{x_1 \ldots x_n\}$ m.a.s. $\longrightarrow$ Ordenamos la muestra: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$

y definimos la f.d. empírica $F_n(x)$, como:

$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{k}{n} & X_{(k)} \leq x < X_{(k+1)}\ ;\ k = 1 \ldots n-1 \\ 1 & x \geq X_{(n)} \end{cases}$$



- Por el teorema de Glivenko - Cantelli, la f.d. empírica converge casi seguro a la f.d. poblacional, por lo que:

$$P\left(\lim_{n\to\infty} |F_n(x) - F(x)| = 0 \right) = 1 \sim P\left(\lim_{n\to\infty} D_n = 0 \right) = 1$$

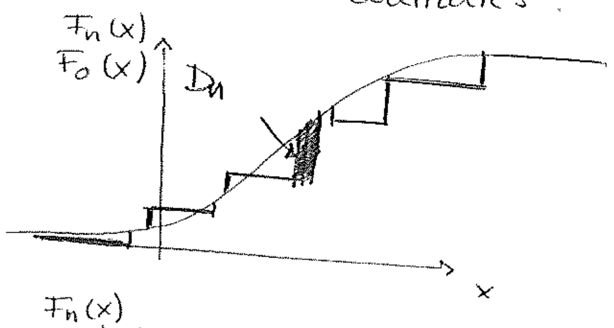
- La distrib. de D_n es libre, no depende del modelo fijado para $F(x)$ en la hipótesis nula H_0.
  - La v.a. $y = F(x)$, por la propiedad de transformación integral, siempre tiene una distrib. uniforme continua en $[0,1]$, y $D_n = \max_{-\infty<x<+\infty} |F_n(x) - y|$
  - La distrib. asintótica de D_n tb. es una distrib. libre de $F(x)$.

El test de K-S se puede formular mediante tres contrastes:

a) $H_0 : F(x) = F_0(x)$

$H_1 : F(x) \neq F_0(x)$

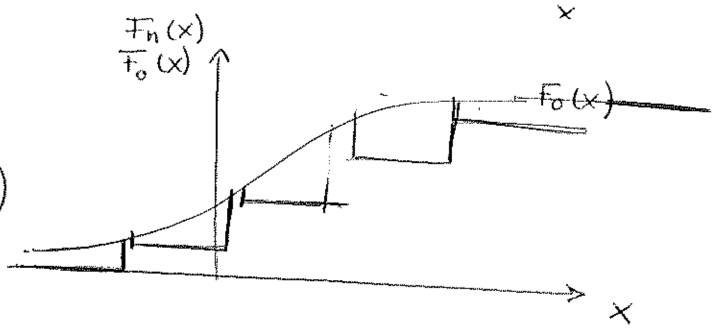
$D_n = \max_{-\infty<x<+\infty} |F_n(x) - F_0(x)|$



b) $H_0 : F(x) \geq F_0(x)$

$H_1 : F(x) < F_0(x)$

$D_n^{+} = \max_{-\infty<x<+\infty} \left(F_0(x) - F_n(x) \right)$



c) $H_0 : F(x) \leq F_0(x)$

$H_1 : F(x) > F_0(x)$

$D_n^{-} = \max_{-\infty<x<+\infty} \left\{ F_n(x) - F_0(x) \right\}$

En todos los casos, los valores críticos del test K-S para una muestra están tabulados, en función del tamaño de la muestra y del nivel de significación α.

La región crítica viene determinada por el valor D_α tq

$$P(D_n > D_\alpha / H_0) = \alpha$$

4 _ COMPARACIÓN de χ^2 y K-S.

- KS es de más fácil aplicación que χ^2.
- KS no se ve afectado por reagrupaciones de las categorías. χ^2 pierde información y su distrib. pierde grados de libertad.
- KS es aplicable a muestras pequeñas, pues utiliza la distrib. exacta del estadístico D_n; mientras que χ^2 necesita muestras grandes pues se base en la distrib. asintótica del estadístico.
- KS tiene mayor potencia de contraste que χ^2, pero para n sufic. grande las potencias son similares.
- χ^2 se adapta fácilmente cuando hay que estimar parámetros, KS no se puede aplicar, salvo para la distrib. normal.
- χ^2 es aplicable a poblaciones con distrib. discretas y continuas y cuando la información muestral es cualitativa. KS sólo es válido para distrib. continuas.

5_ CONTRASTES de NORMALIDAD

Todos los contrastes de normalidad coinciden en tener como hipótesis nula que la muestra procede de una distrib. normal.

5.1_ Contraste de normalidad de LILLIEFORS:

En 1967, Lilliefors una modificación de KS para contrastar la hipótesis compuesta de normalidad → H_0: la muestra procede a una población normal, sin especificar su media y su varianza.

A partir de una m.a.s. $X = \{x_1 \ldots x_n\}$ se estiman la media y la varianza a partir de sus estimadores media muestral y cuasivarianza muestral:

$$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Cada valor muestral se tipifica, $X_i \longrightarrow Z_i = \frac{X_i - \bar{X}}{s}$, $i = 1 \ldots n$

y los cálculos se hacen sobre los valores normalizados.

H_0: la muestra procede de una normal, con μ, σ desconoc.

H_1: la muestra no procede de una normal.

Utilizamos el estadístico de KS sobre los valores normalizados:

$$D'_n = \max_{-\infty < z < \infty} |F_n(z) - F_0(z)|$$

La región crítica viene dada por D'_α tq $P(D'_n > D'_\alpha / H_0) = \alpha$ en donde D'_α está tabulado para diferentes valores de n y α.

5.2 _ Contraste de normalidad de SHAPIRO_WILK :

Trata de contrastar si una muestra procede de una población ~~muestra~~ normal con parámetros desconocidos sin tener que hacer ninguna especificación sobre sus parámetros

Este test es muy útil para muestras de tamaño pequeño ($n<50$) con una apreciable potencia del contraste .

A partir de una m.a.s. $X=\{x_1 - x_n\}$ se ordenan los valores muestrales :

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

El estadístico de Shapiro_Wilk viene dado por :

$$W = \frac{\sum_{i=1}^{K} a_i (X_{(n-i+1)} - X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad (*)$$

donde : $K = n/2$ si n par , $K = \frac{n-1}{2}$ si n impar .

a_i = coef. calculados y tabulados

(*) En Martín_Pliego :

$$W = \frac{\sum^{K} \left[a_{n-i+1} (X_{(n-i+1)} - X_{(i)})\right]^2}{\sum^{n} (X_{(i)} - \bar{X})^2}$$

La región crítica viene dada por :

$$P[W > W_\alpha / H_0] = \alpha$$

donde W_α está tabulado en función de n y α .

5_3. Contraste de normalidad de BERA y JARQUE:

Es el contraste propuesto por Novales en su libro de Econometría. Se basa en las propiedades de simetría y curtosis de la distrib. normal.