TEMA 5 - CONTRASTE DE HIPÓTESIS NO PARAMÉTRICOS

En el tema anterior hemos supuesto conocida la distribución de probabilidad de una determinada variable aleatoria y a partir de muestras obteníamos propiedades o valores de los parámetros de esa distribución. En este tema vamos a estudiar hipótesis que se refieren a la forma de la distribución que sigue la v.a. No existe una teoría general que englobe todos los casos, cada uno y ante cada hipótesis se introducen test específicos.

Una clasificación de los contrastes de hipótesis no paramétricos podría ser:

Contrastes no paramétricos
- Basados en una muestra
  - Aleatoriedad: ¿la muestra es aleatoria?
  - Localización: ¿disponemos de algún valor para localizar la distribución?
  - Bondad de ajuste: buscar una distribución a la que pertenezca la muestra
- Basados en $K \geq 2$ muestras
  - Independencia entre muestras y/o sus distribuciones
  - Igualdad: ¿las muestras pertenecen a la misma distribución?
  - Correlación o asociación entre muestras

Partición del espacio paramétrico.

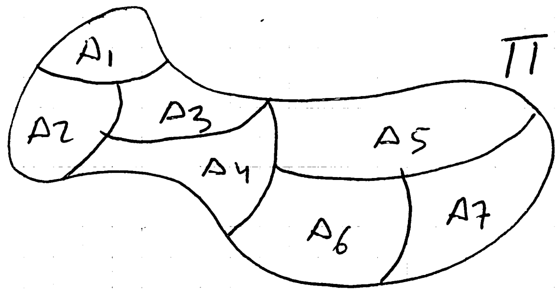
Sea X una v.a. con distribución $X \rightsquigarrow F(x, \theta)$

Sea Π el espacio paramétrico de θ. Una partición de Π es un conjunto de subconjuntos de Π, $A_1, A_2 \dots A_n$ tales que:

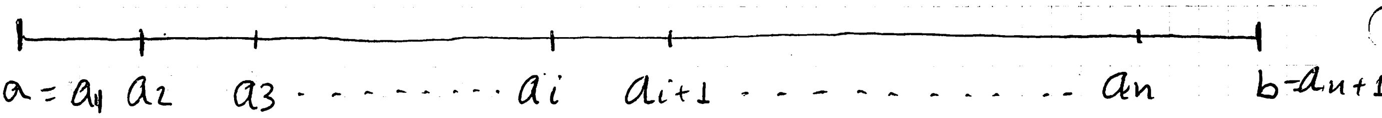
i) $A_1 \cup A_2 \cup \dots \cup A_n = \Pi$

ii) $A_i \cap A_j = \emptyset$ si $i \neq j$

Una representación esquemática de la partición de Π podría ser:



De forma más precisa, supongamos que el espacio paramétrico de θ es $\Pi = [a, b]$



Partimos el intervalo $[a, b]$ del siguiente modo:

$P = \{ a = a_1 < a_2 < a_3 < \dots < a_i < a_{i+1} < \dots < a_n \leq a_{n+1} = b \}$

Entonces si consideramos $S_i = [a_i, a_{i+1}]$, el conjunto $S_1, S_2, \dots S_n$ es una partición de Π

Indicador de diferencias

Sea X una v.a. con función de distribución conocida y sea $A_1, A_2, \dots A_K$ una partición del espacio de variación de X (E_X):

$E_X = A_1 \cup A_2 \cup \dots \cup A_K \ ; \ A_i \cap A_j = \emptyset \quad \forall i \neq j$

Si $P(A_i) = p_i$, sea n_i = nº veces que aparece A_i en n observaciones (n_i es la frecuencia de A_i). Como n_i se distribuye como $n_i \rightsquigarrow B(n, p_i) \Rightarrow E[n_i] = e_i = n \cdot p_i; \ i = 1 \dots K$

Llamamos indicador de diferencias o Ji-dos de Pearson a la expresión:

$$\chi^2_{exp} = \sum_{i=1}^{K} \frac{(n_i - e_i)^2}{e_i}$$

¿porqué se introduce este concepto?

La ventaja que tiene Ji-dos es que converge en ley a una distribución χ^2_{k-r-1}, siendo r = nº de parámetros estimados en la distribución de X, es decir

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \xrightarrow{L} \chi^2_{k-r-1}$$

Test de la bondad de ajuste

Sea X una v.a. cuya función de distribución es conocida, $X \rightsquigarrow F(x, \theta_1, \theta_2, \dots \theta_s)$. Tomamos una partición del espacio paramétrico de X: $A_1, A_2 \dots A_n$ y sea $p_i = P(A_i)$. Tomamos una muestra de tamaño n, $x_1, x_2, \dots x_n$. Pretendemos saber si la muestra pertenece a la distribución $X \rightsquigarrow F(x, \theta_1, \theta_2, \dots \theta_s)$

Sea n_i = nº veces que aparece A_i en la muestra.

Entonces $n_i \rightsquigarrow B(n, p_i)$. El indicador de diferencias es:

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{(n_i - e_i)}{e_i} \quad \text{donde } e_i = n \cdot p_i$$

El contraste que pretendemos hacer es:

$H_0: \{x_i\}_{1 \leq i \leq n} \rightsquigarrow F(x, \theta_1, \dots \theta_s)$

$H_1: \{x_i\}_{1 \leq i \leq n} \not\rightsquigarrow F(x, \theta_1, \dots \theta_s)$

La muestra pertenecerá a la distribución de la v.a X si $n_i \simeq e_i \Rightarrow n_i - e_i \simeq 0 \Rightarrow \chi^2_{exp} \simeq 0$. Entonces la región de rechazo será

$C = \{(x_1, x_2, \dots x_n) \in \mathbb{R}^n / \chi^2_{exp} \geq \delta\}$ para determinar δ tenemos en cuenta que

$P[(x_1, x_2, \dots x_n) \in C] = \alpha \Rightarrow P[\chi^2_{exp} \geq \delta] = \alpha \Rightarrow$

$\Rightarrow P[\chi^2_{exp} < \delta] = 1 - \alpha \Rightarrow \delta = \chi^2_{k-r-1, 1-\alpha}$ ya que $\chi^2_{exp} \rightsquigarrow \chi^2_{k-r-1}$

Luego la región crítica es:

$C = \{ (x_1, x_2, \cdots x_n) \in \mathbb{R}^n \ / \ \chi^2_{exp} \geq \chi^2_{k-r-1, 1-\alpha} \}$

Si $\chi^2_{exp} \geq \chi^2_{k-r-1, 1-\alpha}$ se rechaza la hipótesis nula $\Rightarrow$ la muestra no pertenece a la distribución de X

Si $\chi^2_{exp} < \chi^2_{k-r-1, 1-\alpha}$ no se rechaza la hipótesis nula, la muestra pertenece a la distribución de X.

Veamos un ejemplo:

Ejemplo: tras lanzar un dado 120 veces se han obtenido los siguientes resultados:

Resultado	1	2	3	4	5	6
nº de lanzamientos	23	17	20	22	17	21

¿Es el dado perfecto?

X = resultado del lanzamiento de un dado. El espacio paramétrico de X es $\{1, 2, 3, 4, 5, 6\}$. La función de probabilidad de X es:

$P[X = i] = \frac{1}{6} \quad \forall i = 1 \cdots 6$. Si llamamos A_i = obtener i al lanzar un dado $i = 1, 2, \cdots 6$ el contraste que hemos de realizar para saber si el dado está trucado es:

$H_0: \ p_i = \frac{1}{6} \quad \forall i = 1, 2, \cdots 6$

$H_1: \ p_i \neq \frac{1}{6} \quad \forall i = 1, 2, \cdots 6$ donde $p_i = P(A_i)$

Calculemos χ^2_{exp}

A_i	n_i	$e_i = n \cdot p_i$	$n_i - e_i$	$\frac{(n_i - e_i)^2}{e_i}$
1	23	$120 \cdot \frac{1}{6} = 20$	3	$\frac{9}{20}$
2	17	20	-3	$\frac{9}{20}$
3	20	20	0	0
4	22	20	2	$\frac{4}{20}$
5	17	20	-3	$\frac{9}{20}$
6	21	20	1	$\frac{1}{20}$

Obtenemos:

$$\chi^2_{exp} = \sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = \frac{32}{20} = 1'6$$

Si suponemos un nivel de significación del 5%:
$\chi^2_{k-r-1,1-\alpha} = \chi^2_{6-0-1,0'95} = \chi^2_{5,0'95} = 11'1$
Como $\chi^2_{exp} \not> \chi^2_{5,0'95}$ no se puede rechazar la hipótesis nula, aceptamos que el dado no está trucado.
Hay que hacer la siguiente observación: para que sea cierto $\chi^2_{exp} \xrightarrow{L} \chi^2_{k-r-1}$ ha de verificarse que $n \cdot p_i > 5$ o bien $n \cdot (1-p_i) > 5$. En este caso esto se verifica ya que $120 \cdot \frac{1}{6} = 20 > 5$.

Test χ^2 de independencia

Nos planteamos contrastar la independencia de dos características de una población (por ejemplo nº de hijos - nº de dormitorios de la vivienda familiar, estatura - peso, etc)
Sea (X, Y) una población bidimensional de modo que X puede presentarse en r- modalidades $\{x_1, x_2, \dots x_r\}$ e Y puede presentarse en s modalidades $\{y_1, y_2, \dots y_s\}$. Esto equivale a decir que $\{x_1, x_2 \dots x_r\}$ es partición de X, $\{y_1, y_2, \dots, y_s\}$ partición de Y.
En una muestra se obtiene la siguiente tabla de frecuencia

$X \backslash Y$	y_1	y_2	....	y_j	....	y_s	$n_{i\cdot}$
x_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1s}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	....	n_{2j}	....	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	....	$\vdots$	....	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	....	$\vdots$	....	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	....	n_{rj}	....	x_{rs}	
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	....	$n_{\cdot j}$	....	$n_{\cdot s}$	n

donde n_{ij} = nº de veces que aparece la modalidad (x_i, y_j) al repetir n-veces el experimento. Entonces $n_{ij} \rightsquigarrow B(n, p_{ij})$ con lo que $E[n_{ij}] = n \cdot p_{ij} = e_{ij}$

Se tiene que

$p_{i\cdot} = P[X = x_i]$ $\qquad p_{\cdot j} = P[Y = y_j]$ Una estimación de $p_{i\cdot}$ y $p_{\cdot j}$ es:

$$p_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad p_{\cdot j} = \frac{n_{\cdot j}}{n}$$

Si X e Y son independientes $p_{ij} = p_{i\cdot} \cdot p_{\cdot j} = \dfrac{n_{i\cdot}\, n_{\cdot j}}{n^2}$

$$\Rightarrow e_{ij} = n \cdot p_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

Se define el indicador de diferencias como:

$$\chi^2_{exp} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}}$$

Se tiene que $\chi^2_{exp} \xrightarrow{L} \chi^2_{r\cdot s - q - 1}$ siendo q el número de parámetros estimados.

Como:

$\sum_{i=1}^{r} p_{i\cdot} = 1$ se estiman $r-1$ parámetros

$\sum_{j=1}^{s} p_{\cdot j} = 1$ se estiman $s-1$ parametros

Luego $q = r-1+s-1 = r+s-2$ y por tanto:

$$\chi^2_{exp} \xrightarrow{L} \chi^2_{rs-r-s+2-1} = \chi^2_{rs-r-s+1} = \chi^2_{(r-1)(s-1)}$$

Si X e Y son independientes $e_{ij} \simeq n_{ij} \Rightarrow n_{ij} - e_{ij} \simeq 0$ $\Rightarrow \chi^2_{exp} \simeq 0$. Si rechazamos la hipótesis nula entonces χ^2_{exp} no es cero $\Rightarrow \chi^2_{exp} \geq K$

Luego la región de rechazo se caracteriza porque se verifica que $\chi^2_{exp} \geq K$. Si el nivel de significación es α entonces:

$P[\chi^2_{exp} \geq k] = \alpha \Rightarrow P[\chi^2_{exp} < k] = 1-\alpha$ y

como $\chi^2_{exp} \rightsquigarrow \chi^2_{(r-1)(s-1)} \Rightarrow k = \chi^2_{(r-1)(s-1),1-\alpha}$

luego la región de rechazo se caracteriza porque:
$\chi^2_{exp} \geq \chi^2_{(r-1)(s-1),1-\alpha}$

Si $\chi^2_{exp} \geq \chi^2_{(r-1)(s-1),1-\alpha} \Rightarrow$ Se rechaza H_0
Si $\chi^2_{exp} \leq \chi^2_{(r-1)(s-1),1-\alpha} \Rightarrow$ No se rechaza H_0.

Veamos un ejemplo

Ejemplo: Se clasifican 6.800 personas según el color del cabello (Y) y de los ojos (X) con el siguiente resultado:

X \ Y	Rubio	Castaño	Negro	Rojo	$n_{i.}$
Azules	1.768	807	189	47	2.811
Verdes	946	1.387	746	53	3.132
Pardos	115	438	288	16	857
$n_{.j}$	2.829	2.632	1.223	116	6.800

Estudiar al 5% si existe independencia entre una y otra característica.

Solución:

En primer lugar vamos a calcular

$$\chi^2_{exp} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}\,n_{.j}}{n}\right)^2}{\frac{n_{i.}\,n_{.j}}{n}}$$

, calculando previamente el valor $\frac{n_{i.}\,n_{.j}}{n}$

	$n_{.1} = 2829$	$n_{.2} = 2632$	$n_{.3} = 1223$	$n_{.4} = 116$
$n_{1.} = 2811$	1.169'458	1.088'02	505'566	47'952
$n_{2.} = 3132$	1.303'004	1.212'268	563'299	53'428
$n_{3.} = 857$	356'537	331'709	154'134	14'619

Ahora calculamos $\dfrac{\left(n_{ij} - \dfrac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\dfrac{n_{i\cdot}\, n_{\cdot j}}{n}}$

X \ Y	Rubio	Castaño	Negro	Rojo	
Azules	306'34	72'585	198'222	0'019	577'166
Verdes	97'814	25'185	59'257	0'003	182'259
Pardos	163'63	34'059	116'263	0'130	314'082
					1.073'507

Luego $\chi^2_{exp} = 1073'507$

$\chi^2_{(r-1)(s-1),\alpha} = \chi^2_{6,0'05} = 12'6$

Como $\chi^2_{exp} > \chi^2_{6,0'05}$ se rechaza la hipotesis nula.

Recordemos que el test que hemos de hacer es

H_0: X, Y son independientes

H_1: X, Y son dependientes

Luego las variables X e Y son dependientes.

Test χ^2 de homogeneidad

Test χ^2 de homogeneidad

Queremos comprobar si un conjunto de muestras proceden de una misma población X. Supongamos que la variable aleatoria X posee r-modalidades $x_1, x_2 \dots x_r$ (esto es tanto como decir que $x_1, x_2, \dots, x_n$ es una partición del espacio paramétrico de X). Si tomamos S muestras las frecuencias que corresponde a cada una se refleja en la siguiente tabla:

muestra \ x_i	x_1	x_2	$\dots$	x_j	$\dots$	x_r	n_i
1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1r}	n_1
2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2r}	n_2
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ir}	n_i
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
S	n_{S1}	n_{S2}	$\dots$	n_{Sj}	$\dots$	n_{Sr}	n_S
m_j	m_1	m_2	$\dots$	m_j	$\dots$	m_r	n

Queremos hacer el contraste:

H_0: las muestras son homogeneas (proceden de la misma población)
H_1: las muestras no son homogeneas

Pero decir que las muestras son homogeneas equivale a decir que $x_1, x_2, \dots x_r$ es independiente de la muestra tomada por lo que el estadístico de contraste es:

$$\chi^2_{exp} = \sum_{i=1}^{S} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_i \cdot m_j}{n}\right)^2}{\frac{n_i \cdot m_j}{n}}$$

Si α es el nivel de significación, la región critica está en los valores para los que $\chi^2_{exp} > \chi^2_{(r-1)(S-1),\alpha}$

• Si $\chi^2_{exp} > \chi^2_{(r-1)(s-1),\alpha}$ se rechaza la hipótesis nula, no podemos afirmar que las muestras sean homogeneas

• Si $\chi^2_{exp} < \chi^2_{(r-1)(s-1),\alpha}$ aceptamos que las muestras puedan ser homogeneas.

A la hora de ver si se cumple $np > 5$, para poder afirmar todo esto, hay que tener en cuenta que este valor es $\frac{n_i \cdot m_j}{n}$ ya que:

$P[x_j] = P[X_j = x_j] = \frac{m_j}{n}$ y como el tamaño de la muestra es $n_i \Rightarrow n \cdot p = n_i \cdot \frac{m_j}{n} = \frac{n_i \cdot m_j}{n}$

Veamos un ejemplo.

Ejemplo: Cada concurso de plazas de profesores titulares de universidad es valorado por uno de los cinco tribunales existentes (A, B, C, D ó E). Los resultados son los siguientes:

	APTO	NO APTO	TOTAL
A	15	15	30
B	9	5	14
C	0	6	6
D	24	10	34
E	2	14	16
	50	50	100

¿Existen diferencias significativas al 5% entre la forma de valorar de cada tribunal?

En primer lugar vamos a calcular la frecuencia esperada $\frac{n_i \cdot m_j}{n}$

	APTO	NO APTO
A	15	15
B	7	7
C	3	3
D	17	17
E	8	8

→ problema $np \not> 5$

Hemos de agrupar las frecuencias de B y C tomando una nueva muestra a la que llamamos B+C

	APTO	NO APTO	TOTAL
A	15	15	30
B+C	9	11	20
D	24	10	34
E	2	14	16
	50	50	100

(Tabla inicial)

Sobre esta última tabla calculamos $\frac{n_i m_j}{n}$

	APTO	NO APTO
A	15	15
B+C	10	10
D	17	17
E	8	8

Ahora calculamos χ^2_{exp}

	APTO	NO APTO	
A	0	0	0
B+C	0'1	0'1	0'2
D	2'882	2'2882	5'764
E	4'5	4'5	9

$\Rightarrow \chi^2_{exp} = 14'964$

Se tiene que $\chi^2_{(2-1)(4-1),0'05} = \chi^2_{3,0'05} = 7'81$

Al ser $\chi^2_{exp} = 14'964 > \chi^2_{3,0'05}$ se rechaza la hipótesis nula, no podemos afirmar que las muestras sean homogeneas.

Test de ajuste de Kolmogorov-Smirnov

Pretendemos saber si una determinada muestra pertenece a una población concreta que tiene una función de distribución $F_0(X)$. La v.a. X debe ser continua y la distribución $F_0(X)$ no debe tener parámetros desconocidos. Ordenamos la muestra en sentido creciente de modo que sus valores son $x_1 < x_2 < \dots < x_n$. Definimos la función de distribución experimental como:

$$F_n(x_i) = \frac{1}{n} \sum_{j=1}^{i} n_j$$

siendo (x_i, n_i) la distribución de frecuencias en la muestra.

Se define la diferencia superior como:

$$a_i = \max \{ |F_n(x_i) - F_0(x_i)|, \; i = 1 \dots n \}$$

Se define la diferencia inferior como:

$$b_i = \max \{ |F_n(x_{i-1}) - F_0(x_i)| \}$$

Si $D = \max \{ a_i, b_i \}_{1 \le i \le n}$, el estadístico de contraste para es test:

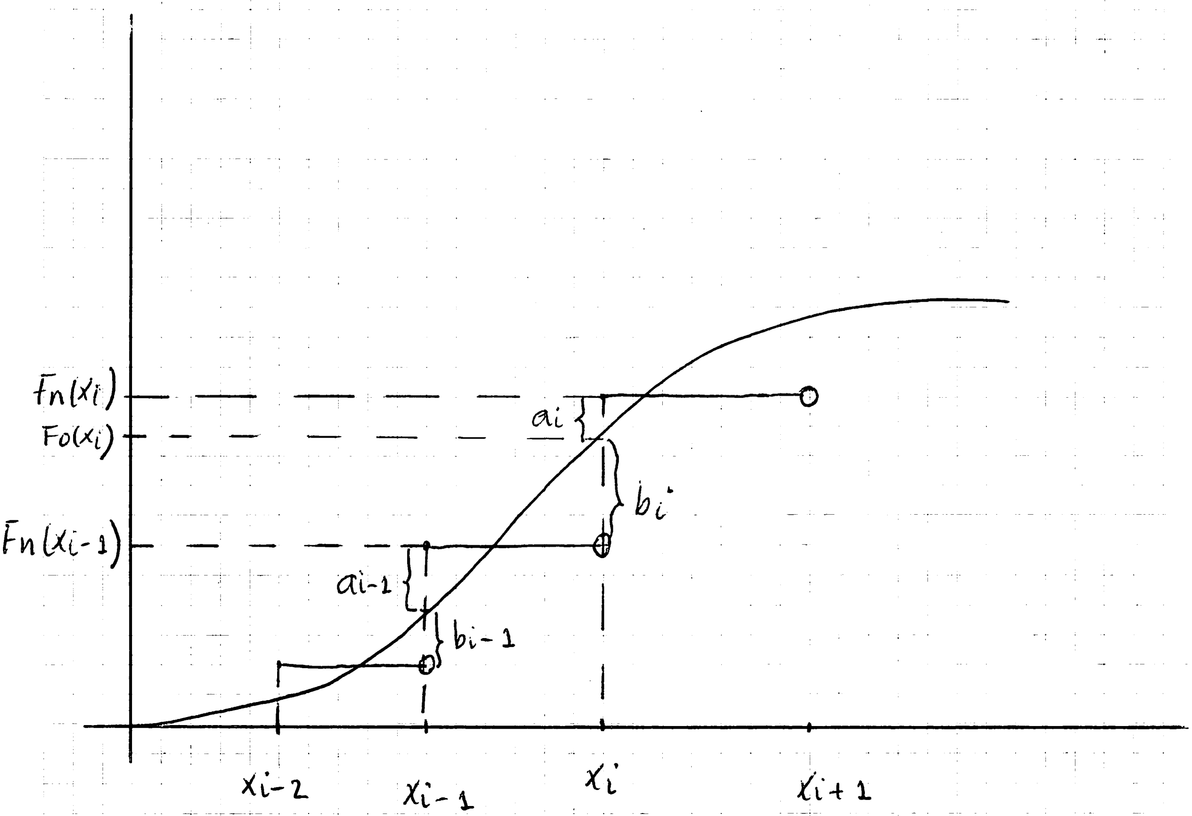
$H_0: \{x_1, x_2, \dots x_n\} \rightsquigarrow F_0(X)$

$H_1: \{x_1, x_2, \dots x_n\} \not\rightsquigarrow F_0(X)$ es precisamente D,

de forma más clara:

- Si $D > D_{n,\alpha}$ (tabla K-S) $\Rightarrow$ se rechaza la hipótesis nula.
- Si $D < D_{n,\alpha}$ (tabla K-S) $\Rightarrow$ se acepta la hipótesis nula.

De forma gráfica sería:



Veamos un ejemplo

Ejemplo: A partir de las siguientes observaciones

3'0050	5'3400	6'2914	5'8420	3'3656
5'2543	4'9195	5'9864	5'0541	4'4578
3'7732	4'4012	3'6964	4'2160	3'5893
6'2350	6'5838	6'4358	3'0599	

¿Se ajusta la distribución poblacional correspondiente a una $U(3,7)$?

Solución:

Queremos hacer el contraste:

$H_0: X \rightsquigarrow U(3,7)$

$H_1: X \not\rightsquigarrow U(3,7)$

Ordenamos los valores de la muestra de menor a mayor, calculando después a_i y b_i

x_i	n_i	$F_n(x_i)$	$F_0(x_i)$	a_i: $\lvert F_n(x_i) - F_0(x_i) \rvert$	b_i: $\lvert F_n(x_{i-1}) - F_0(x_i) \rvert$
3'0050	1	1/19	0'0012	0'05138	0'00125
3'0599	1	2/19	0'0150	0'09028	0'03765
3'3656	1	3/19	0'0914	0'06649	0'01386
3'5893	1	4/19	0'1473	0'06319	0'01056
3'6964	1	5/19	0'1741	0'08905	0'03642
3'7732	1	6/19	0'1933	0'12248	0'06985
4'2160	1	7/19	0'3040	0'06441	0'01178
4'4012	1	8/19	0'3503	0'07074	0'01811
4'4578	1	9/19	0'3645	0'10922	0'05659
4'9195	1	10/19	0'4799	0'04642	0'00621
5'0541	1	11/19	0'5135	0'06540	0'01277
5'2543	1	12/19	0'5636	0'06798	0'01535
5'3400	1	13/19	0'5850	0'09919	0'04656
5'8420	1	14/19	0'7105	0'02632	0'02631
5'9864	1	15/19	0'7466	0'04285	0'00978
6'2350	1	16/19	0'8088	0'03333	0'01930
6'2914	1	17/19	0'8229	0'07186	0'01923
6'4358	1	18/19	0'8590	0'08839	0'03576
6'5838	1	19/19	0'8960	0'10402	0'05139

Los cálculos los hemos realizado con la función de distribución teórico $F_0(x)$ hallado como sigue:

La función de densidad de $X \rightarrow U(3,7)$ es

$$f(x) = \begin{cases} \frac{1}{4} & \text{si } 3 \leq x < 7 \\ 0 & \text{en el resto} \end{cases}$$

La función de distribución es

$$F_0(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{3}^{x} \frac{1}{4}\,dx = \frac{x-3}{4}$$

. Por tanto

$$F_0(x) = \begin{cases} \frac{x-3}{4} & \text{si } 3 \leq x < +\infty \\ 0 & \text{si } -\infty < x < 3 \end{cases}$$

Se tiene que $\max\{a_i, b_i\}_{1\le i\le 19} = 0'12248 = D_n$

como $D_{19, 0'05} = 0'301 \Rightarrow D_n < D_{n,\alpha}$ ya que $D_n = 0'12248 < D_{19,0'05} = 0'301$ se acepta la hipótesis nula, no podemos rechazar la hipótesis nula.

Test de normalidad de Lilliefors

Dada una muestra $x_1, x_2, \dots x_n$ pretendemos averiguar si procede de una normal $N(\mu, \sigma)$ con parámetros μ y σ desconocidos.

A partir de la muestra estimamos los parámetros

$$\hat{\mu} = \bar{x} = \frac{\sum_1^n x_i}{n} \qquad \hat{\sigma} = \sqrt{S_c^2} = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}}$$

Realizamos el contraste:

$H_0: X \rightarrow N(\hat{\mu}, \hat{\sigma})$

$H_1: X \not\rightarrow N(\hat{\mu}, \hat{\sigma})$

Ordenamos la muestra de menor a mayor valor de $\{x_i\}_{1\le i\le n}$, sea n_j = frecuencia, en la muestra de x_j.

El estadístico de contraste es:

$$D_n = \max\{a_i, b_i\}_{1\le i\le n} \quad \text{donde:}$$

$a_i = \max\{|F_n(x_i) - F_0(x_i)|, 1\le i\le n\}$

$b_i = \max\{|F_n(x_{i-1}) - F_0(x_i)|, 1\le i\le n\}$

$F_n(x_i) = \frac{1}{n}\sum_{j=1}^{i} n_j$

$F_0(x) = P(X \le x)$ = función de distribución de $N(\hat{\mu}, \hat{\sigma})$

Si $D_n > D_{n,\alpha}$ (tabla de Lilliefors) $\Rightarrow$ R H_0

Si $D_n < D_{n\alpha} \Rightarrow$ ~~R~~ H_0

Ejemplo:

A partir de las siguientes observaciones:

179, 100, 45, 384, 230, 100, 320, 80, 220, 320, 210, 159

¿se ajusta la distribución correspondiente a una normal?

Solución

Hallamos la media $\bar{x} = 195'5833$ y la cuasi-varianza $s_c = 106'330923$

Hacemos el contraste:

$$H_0 : X \rightsquigarrow N(195'5833, 106'330923)$$

$$H_1 : X \not\rightsquigarrow N(195'5833, 106'330923)$$

Ordenamos de menor a mayor los datos de la muestra y calculamos ai y bi

x_i	n_i	$F_n(x_i)$	$F_0(x_i)$	$\underbrace{\lvert F_n(x_i) - F_0(x_i) \rvert}_{a_i}$	$\underbrace{\lvert F_n(x_{i-1}) - F_0(x_i) \rvert}_{b_i}$
45	1	0'0833	0'07836	0'00497	0'07836
80	1	0'1667	0'13852	0'02815	0'05518
100	2	0'3333	0'18435	0'14899	0'01768
159	1	0'4167	0'36540	0'05126	0'03207
179	1	0'5000	0'43803	0'06197	0'02137
210	1	0'5833	0'55392	0'02941	0'05392
220	1	0'6667	0'59081	0'07586	0'00748
230	1	0'7500	0'62691	0'12309	0'03976
320	2	0'9167	0'87902	0'03765	0'12902
384	1	1	0'96180	0'03820	0'04513
	$n = 12$				

$D_n = \max\{a_i, b_i\} = 0'14899$

$D_{n,\alpha} = D_{12,0'05} = 0'242$ (tabla de Lilliefors)

$D_n < D_{12,0'05}$ no se puede rechazar la hipótesis nula