E2T8-9. CONTRASTES NO PARAMÉTRICOS

ESTADÍSTICA II – URJC

GRADO DE ECO Y ADE (2019/2020)

WWW.PACORABADAN.COM

ÍNDICE

1. PRUEBA CHI-CUADRADO

- Compara, a través del estadístico X^2, las posibles diferencias entre las frecuencias observadas en una distribución de una variable y las esperadas en razón de una determinada hipótesis.
- "Se dispone de una muestra cuyas observaciones pueden clasificarse en r clases o categorías mutuamente excluyentes: categorías que pueden venir dadas en intervalos de tipo cuantitativo o simplemente en una escala ordinal o nominal.

Categorías	Nº de elementos en la muestra	P (Categoría/ H_0 cierta)
A_1	n_1	π_1
A_2	n_2	π_2
…	…	…
A_i	n_i	π_i
…	…	…
A_r	n_r	π_r
$A_1 \cup A_2 \cup \cdots \cup A_r \equiv E$	N	1

- Bajo $H_0\ cierta\ , m.a.s\ \rightarrow$ observaciones independientes, las probabilidad de la intersección de las frecuencias muestrales es una ley multinomial

$$P(n_1 \cap n_2 \cap \cdots \cap n_r) = \frac{N!}{n_1!\, n_2! \ldots n_r!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_r^{n_r}$$

- Por tanto, cada frecuencia $n_i : B(N; \pi_i)$
- Y el valor esperado es $E[n_i] = n\pi_i = E_i$
- E_i es el número de observaciones de la clase A_i que cabe esperar se obtengan en la muestra, si la distribución de probabilidad de la población es la que se incluye en H_0 .

1. PRUEBA CHI-CUADRADO

$$\left.\begin{array}{c} H_0\ cierta \\ + \\ m.a.s\ (n)\ siendo\ n\ suf.grande \end{array}\right\} n_i \xrightarrow[N\to\infty]{} E_i$$

- X^2 Estadístico de Pearson (de bondad del ajuste):

$$X^2 = \sum_{i=1}^{r} \frac{(n_i - E_i)^2}{E_i} = \sum_{i=1}^{r} \frac{(n_i - nP_i)^2}{nP_i}$$

  - $(n_i - E_i)^2$ →evitar que signos contrarios compensen la media global.
  - $\frac{1}{E_i}$ →una discrepancia grande podría llevar a rechazar el modelo de probabilidad de H_0, aunque $\mathrm{P}(\mathrm{A_i})$ no fuera muy grande.

- X^2 Distribución: distribución conjunta binomial.
  - Las variables tipificadas $\xi_i = \frac{n_i - N\pi_i}{\sqrt{n\pi_i}} \xrightarrow[d:\ condicional\ r\ Poisson]{} z: N(0,1)$
  - Ligadura: $\sum n_i = N \to \sum P_i\,\xi_i = 0 \to X^2 = \sum_{i=1}^{r} \xi_i^2$
  - Por tanto es suma de normales al cuadrado $X^2 : \chi^2_{r-1}$

1. PRUEBA CHI-CUADRADO χ^2

$$\left.\begin{array}{c} H_0\ cierta \\ + \\ m.a.s\ (n)\ siendo\ n\ suf.grande \end{array}\right\} n_i \xrightarrow[N\to\infty]{} E_i$$

X^2 Estadístico: $X^2 = \sum_{i=1}^{r} \frac{(n_i - E_i)^2}{E_i} = \sum_{i=1}^{r} \frac{(n_i - n\pi_i)^2}{n\pi_i}$

- X^2 Distribución: $X^2 : \chi^2_{r-1}$
- Región crítica:
  - $\Delta X^2 \rightarrow$ diferencias altas entre las frecuencias observadas y las esperadas $\rightarrow$ $\mathbf{Rechazamos\ H_0}$
  - $R.C. \rightarrow \{X^2 \geq K\}\ donde\ \alpha = P[X^2 \geq K; si\ H_0\ cierta]$ … Lo encuentro en las tablas de χ^2_{r-1}
- Observaciones:
  1. X^2 tiene distribución asintótica a χ^2_{r-1} por lo que necesitamos que $E_i \geq 5\ \forall i$
  2. Test válido para poblaciones tanto continuas como discretas (categorías cualitativas y cuantitativas)
  3. Aconsejable que la probabilidad de las categorías sean aproximadamente iguales.

1.1. TEST DE BONDAD DEL AJUSTE ($\chi^2_{R'-k-1}$)

- Contrasta la hipótesis que se refieren al tipo de **distribución de probabilidad de la población**.
- $\boldsymbol{H_0}$ **es no paramétrica y simple**, y
  - se establece en función de observaciones de la población, es decir, observando la nube de puntos, un histograma de frecuencias, etc. o teniendo en cuenta medidas representativas de la distribución muestral.
  - ¡No se pueden establecer hipótesis en contra de los datos!

$$\begin{cases} H_0: la\ población\ sigue\ una\ distribución\ determinada\ (\ con\ o\ \text{sin}\ parámetro) \\ H_1: H_0\ no\ es\ cierta. \end{cases}$$

- No se puede formar H_1 de otra forma, por ejemplo, otra distribución
- Procedimiento: se trata de comparar como son los datos (frecuencia empíricas) con como deberían ser en función de las probabilidades teóricas (frecuencias teóricas) :
  - Si no conocemos el/los parámetros poblacionales θ (nos lo dice la H_0) , los estimamos por los estimadores máximo verosímiles $\hat{\theta}_{MV}$.

1. TEST DE BONDAD DEL AJUSTE ($\chi^2_{R'-k-1}$)

Nº clases (rango)	Realizaciones	Frecuencias empíricas	Probabilidades Teóricas	Frecuencias teóricas	$\frac{(n_i-\varphi_i)^2}{\varphi_i}$
1	x_1	n_1	$P_1 = P(\xi = x_1)$	$\varphi_1 = NP_1$	$\frac{(n_1-\varphi_1)^2}{\varphi_1}$
2	x_2	n_2	$P_2 = P(\xi = x_2)$	$\varphi_2 = NP_2$	$\frac{(n_2-\varphi_2)^2}{\varphi_2}$
3	x_3	n_3	$P_3 = P(\xi = x_3)$	$\varphi_3 = NP_3$	$\frac{(n_3-\varphi_3)^2}{\varphi_3}$
…	…	…	…	…	…
R	x_R	n_R	$P_R = P(\xi = x_R)$	$\varphi_R = NP_R$	$\frac{(n_R-\varphi_R)^2}{\varphi_R}$
		$N = \sum_{i=1}^{R} n_i$	1	N	$I_d = \sum \frac{(n_i-\varphi_i)^2}{\varphi_i}$

P_i: las obtenemos según la distribución de H_0

n_i: son las frecuencias observadas

φ_i : frecuencias que deberíamos haber obtenido bajo H_0 cierta

$\mathbb{CN}$: $\varphi_i \geq 5$ si esto no se cumple, tenemos que agrupar por intervalos, lo que modificaría la tabla.

$\frac{(n_i-\varphi_i)^2}{\varphi_i}$ nos sirve para comparar las frecuencias teóricas con las observadas

$$Bajo\ H_0\ cierta \rightarrow I_d: \chi^2_{R'-k-1}\ siendo \begin{cases} R'el\ nº\ de\ clases\ agrupadas \\ k\ el\ número\ de\ parámetros\ estimados \end{cases}$$

$$R.C. \equiv Id \geq d_\alpha \qquad d_\alpha\ ? \rightarrow P\left(\chi^2_{R'-k-1} > d_\alpha\right) = \alpha$$

1. TEST DE BONDAD DEL AJUSTE ($\chi^2_{R'-k-1}$)

Cuantía siniestro (en miles) C_i	Número de siniestros n_i	Probabilidades Teóricas P_i	Frecuencias teóricas $\varphi_i = NP_1$	$\frac{(n_i - \varphi_i)^2}{\varphi_i}$
$C_i < 60$	10	1/5	20	5
(60,80)	25	1/5	20	1′25
(80,100)	25	1/5	20	1′25
(100,120)	20	1/5	20	0
$C_i > 120$	20	1/5	20	0
	$N = 100$ (tamaño muestral)	1	N	$I_d = 7'5$

Problema 1. La cuantía de los pagos por siniestros en miles de € se distribuye como recoge la tabla (parte verde). Contraste a un nivel de significación del 5% si podemos considerar que hay equiprobabilidad en la cuantía del desembolso para las categorías propuestas.

$$I_d: X^2_{R'-k-1} = X^2_{5-0-1} = X^2_4$$

$$R.C. \equiv Id \geq 9{,}4877$$

$$P(X^2_4 > d_\alpha) = 5\% \rightarrow d_\alpha = 9'4877$$

=INV.CHICUAD(0,95;4)

En nuestro caso $I_d = 7'5 < 9{,}4877 \rightarrow$ No rechazo H_0

1. TEST DE BONDAD DEL AJUSTE ($\chi^2_{R'-k-1}$)

Problema 2. Observados 4000 individuos que han estado expuestos a un riesgo durante cierto periodo de tiempo se sabe que 1870 de ellos han sufrido determinado tipo de accidente. ¿ Puede mantenerse la hipótesis de que hay la misma probabilidad de sufrir este accidente de que no sufrirlo con un nivel de confianza del 99%?

Accidente C_i	Número de siniestros n_i	Probabilidades Teóricas P_i	Frecuencias teóricas $\varphi_i = NP_1$	$\frac{(n_i-\varphi_i)^2}{\varphi_i}$
Si	1870	0,5	2000	$8'45$
No	2130	0,5	2000	$8'45$
	$N = 4000$ (tamaño muestral)	1	N=4000	$I_d = 16'9$

$$I_d: X^2_{R'-k-1} = X^2_{2-0-1} = X^2_1$$

$$P(X^2_1 > d_\alpha) = 1\% \rightarrow d_\alpha = 6{,}63$$

=INV.CHICUAD(0,99;1)

$$R.C. \equiv Id \geq 6{,}63$$

En nuestro caso $I_d = 16,9 > 6,63 \rightarrow$ Rechazo H_0

1. TEST DE BONDAD DEL AJUSTE ($\chi^2_{R'-k-1}$)

Problema 3. Para realizar reformas en una gasolinera, necesitamos conocer la distribución de probabilidad de la llegada de los vehículos al día. Para contrastar esta hipótesis, se realiza un m.a.s. de tamaño. 907. Con estos resultados contraste con un nivel de significación del 5%, que esta muestra se ajusta a una distribución paramétrica conocida.

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	P_i	$\varphi_i = nP_i$	$E_i > 5$	$\frac{(n_i-\varphi_i)^2}{\varphi_i}$
0	195	0	0	0,20	184,78	184,78	0,56
1	301	301	301	0,32	293,98	293,98	0,17
2	210	420	840	0,26	233,86	233,86	2,43
3	125	375	1125	0,14	124,02	124,02	0,01
4	45	180	720	0,05	49,33	49,33	0,38
5	22	110	550	0,02	15,70	15,70	2,53
6	7	42	252	0,00	4,16	5,33	2,53
7	1	7	49	0,00	0,95		
8	1	8	64	0,00	0,19		
9	0	0	0	0,00	0,03		
	907	1443	3901	1,00	906,99	906,99	8,61

- $H_0: \xi: Poisson(\lambda) \rightarrow k = 1$
- Utilizamos $\overline{x} = \hat{\lambda}_{MV} = \frac{1443}{907} = 1'59$

$$P_i = P(\xi = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Cuando φ_i<5, agrupamos :

  5'33=4'16+0'95+019+0,03
- Al agrupar pasamos de R=10 a R'=7
- $2'53 = \frac{((7+1+1+0)-5{,}33)^2}{5'33}$

$I_d: X^2_{R'-k-1} = X^2_{7-1-1} = X^2_4$

$P(X^2_4 > d_\alpha) = 5\% \rightarrow d_\alpha = 14{,}1$ | =INV.CHICUAD(0,95;7)

$R.C. \equiv Id \geq 14{,}1$

En nuestro caso $I_d = 8{,}61 < 14{,}1 \rightarrow$ No rechazo H_0

1.2. TEST DE INDEPENDENCIA EN TABLAS DE CONTINGENCIA ($\chi^2_{(h-1)(k-1)}$)

- Contrasta la hipótesis de que **dos variables nominales** (Por tanto, se puede aplicar también al resto de tipo de variables: ordinales, de escala y de razón) con distribución muestral conocida de frecuencias conjuntas **son estadísticamente independientes**.
- **H_0 es no paramétrica y simple**, $\begin{cases} H_0: A, B \text{ son estadísticamente independientes} \\ H_1: H_0 \text{ es falsa} \end{cases}$
- Bajo la hipótesis de independencia, la probabilidad de aparición de A_i y A_j de manera conjunta es el producto de sus probabilidades :$P_{ij} = P(A_i \cap A_j) = P(A_i) * P(A_j) = P_{i*}P_{*j}$, teniendo en cuenta que $\sum P_{i*} = \sum P_{*j} = 1$
- Los estimadores maximoverosímiles de estas probabilidades son $\widehat{P}_{i_{MV}} = \frac{n_{i*}}{N}; \widehat{P}_{j_{MV}} = \frac{n_{*}j}{N}$
- Para efectuar el contraste de independencia, utilizamos el estadístico $X^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
  - O_{ij} = frecuencias observadas= n_{ij} (empíricos muestrales)
  - E_{ij} = frecuencias esperadas = $n * \hat{P}_{i*} * \hat{P}_{*j}$ (teóricos)
- Distribución del estadístico X^2: $X^2_{nº \text{ de observaciones } - nº \text{ parametros estimado}} = X^2_{hk-(h-1)-(k-1)-1} = X^2_{(h-1)(k-1)}$

$$RC \equiv X^2 \geq K \; ; \; \alpha = P\left(X^2_{(h-1)(k-1)} \geq k\right)$$

1.2. TEST DE INDEPENDENCIA EN TABLAS DE CONTINGENCIA ($\chi^2_{(h-1)(k-1)}$)

Problema 4. Realice un contraste de independencia de las categorías A y B con un nivel de confianza del 5%

$$\widehat{P}_{i_{MV}} = \frac{n_{i*}}{N}; \widehat{P}_{j_{MV}} = \frac{n_{*}j}{N}$$

O_{ij}	B_1	B_2	P_{i*}
A_1	15	10	25
A_2	8	17	25
P_{*j}	23	27	N=50

E_{ij}	B_1	B_2	P_{i*}
A_1	$\frac{25*23}{50} = 11{,}5$	$\frac{25*27}{50} = 13{,}5$	25
A_2	$\frac{25*23}{50} = 11{,}5$	$\frac{25*27}{50} = 13{,}5$	25
P_{*j}	23	27	N=50

A_iB_j	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
1,1	$\frac{(15-11{,}5)^2}{11{,}5} = \frac{12'25}{11{,}5} = 1{,}065$
1,2	0'907
2,1	1'065
2,2	0'907
$X^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	**3'945**

$$H_0 \text{ cierta: } P(X_1^2 > d_\alpha) = 5\% \rightarrow d_\alpha = 3{,}84$$

=INV.CHICUAD(0,95;1)

$$R.C. \equiv Id \geq 3{,}84$$

En nuestro caso $d_\alpha = 3'945 > 3{,}84 \rightarrow$ ***Rechazo*** H_0

1.3. TEST DE HOMOGENEIDAD ($\chi^2_{(h-1)(k-1)}$)

- También parte de una tabla de doble entrada en la que aparecen:
  - Modalidades: $A_1 \dots A_h$ de un factor A
  - Muestras: $1 \dots k$ obtenidas y clasificadas según las modalidades.
- Se pretende contrasta la H_0 de que todas las muestras proceden de la misma población, esto es, los resultados en las muestras son homogéneos.

O_{ij}		Muestras				
		1	2	…	k	n_{i*}
Factor A	A_1	n_{11}	n_{12}	…	n_{1k}	n_{1*}
	A_2	n_{21}	n_{22}	⋮	n_{2k}	n_{2*}
	…	…	…	n_{ij}	…	…
	A_h	n_{h1}	n_{h2}	⋮	n_{hk}	n_{h*}
		n_{*1}	n_{*2}	…	n_{*k}	N

- n_{ij}: número de elementos de la muestra j que presentan la modalidad i-ésima (A_i) del factor A.
- n_{i*}: número total de elementos muestrales que presentan A_i
- n_{*j}: número de elementos que tienen la muestra j-ésima.
- N: número total de observaciones muestrales.

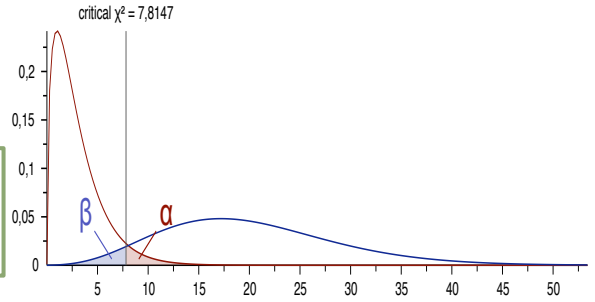
- Bajo H_0 cierta, (H_0: todas las muestras proceden de la misma población),
  - la probabilidad de aparición de cada modalidad del factor A ha de ser: $P(A_h) = \frac{P_h}{\sum P_i = 1}$
  - Las frecuencias esperadas para cada modalidad en la población será, por tanto, $E_{ij} = n_{*j}\widehat{P_i} = \frac{n_{i*}n_{*j}}{N}$;
  - n_{*j} es el número de elementos de cada una de las muestras.

- El estadístico de contraste es

$$X^2 = \sum_{\forall i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Bajo H_0 cierta Bajo X^2: $\chi^2_{(h-1)(k-1)}$

$$\alpha = P[Rechazar\ H_0;\ H_0\ cierta] = P\left[X^2 \geq d_\alpha; \chi^2_{(h-1)(k-1)}\right] = P[\chi^2_1 \geq d_\alpha] \rightarrow d_\alpha en\ tablas.\ \boldsymbol{RC \equiv X^2 \geq d_\alpha}$$



1.3. TEST DE HOMOGENEIDAD ($\chi^2_{(h-1)(k-1)}$)

Problema 5. De 110 árboles frutales elegidos al azar, 50 fueron abonados con un fertilizante y 60 no fueron abonados. En los 110 árboles se analiza la producción actual con respecto a la del año anterior, obteniéndose los siguientes resultados.

Producción	Efecto	Abono	Sin Abono
	↑	20	35
	=	20	15
	↓	10	10

Contraste con un nivel de confianza del 10% que la producción de los árboles abonados es la misma que la producción obtenida en los árboles sin abonar.

$$Poblaciones \begin{cases} X_1\ (abonados) & m.a.s.(50) \\ X_2(\text{sin } abonar) & m.a.s.(60) \end{cases}$$

H_0: las dos muestras proceden de la misma población, es decir, los resultados son homogéneos

	O_{ij}	1	2	
↑	A_1	20	35	55
=	A_2	20	15	35
↓	A_3	10	10	10
		50	60	110

E_{ij}	1	2
A_1	25	30
A_2	15'9	19'1
A_3	9'1	10'9

O_{ij}	E_{ij}	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
20	25	1
35	30	1'2
20	15'9	1'06
15	19'1	0'88
10	9'1	0'09
10	10'9	0'07
		4'30

$$X^2 = \sum_{\forall i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 4'30$$

$$H_0\ cierta, X^2: \chi^2_{(3-1)(2-1)=}\chi^2_2$$
$$RC\ X^2 \geq d_\alpha$$

$$\alpha = 0'1$$
$$= P(Rechazar\ H_0; H_0\ cierta), =$$
$$= P(X^2 \geq d_\alpha; X^2: \chi^2_2)$$
$$= P(\chi^2_2 \geq d_\alpha) \rightarrow d_\alpha = 4'605$$
$$RC\ X^2 \geq d_\alpha$$

=INV.CHICUAD(0,95;2)

En nuestro caso $I_d = 4,3 \leq 4,605 \rightarrow$ No Rechazo H_0

2.1. PRUEBA BINOMIAL

- Esta prueba se puede considerar como un caso particular de la prueba χ^2, en el sentido de que compara las frecuencias observadas de cada categoría de una variable dicotómica con las frecuencias esperadas de una distribución binomial.
- Observaciones:
  - 1. La probabilidad de ocurrencia esperada se refiere siempre a la primera categoría.
  - 2. Estadístico asintótico: la proporción muestral $\hat{\pi}_{MV} \rightarrow N\left(n\pi, \sqrt{n\pi(1-\pi)}\right)$lo que pone en duda la efectividad del test para tamaños muestrales pequeños.
  - Se pueden contrastar tantas variables como queramos, pero se contrastan de manera independiente.

2.2. TEST DE RACHAS

- El objetivo de este contraste es verificar que las observaciones muestrales constituyen una muestra aleatoria procedente de una población continua.
- El test de rachas parte de una variable dicotómica (Si-no, cierto-falso, hombre-mujer, activo-inactivo, etc…) y pretende medir hasta que punto el valor de una observación puede influir en la siguiente.
- Para ello definimos racha como secuencia de observaciones iguales. El mayor número de rachas es un indicador de la independencia (aleatoriedad) de la distribución de las observaciones (si son muchas) o de la dependencia de las mismas (si son pocas).
- Se puede aplicar la convergencia asintótica del estadístico del test de rachas (número de rachas) a una normal mediante **la aproximación de Wald-Wolfovitz**.
  - Inconveniente propio de las distribuciones asintóticas: tamaños muestrales pequeños.

$$R \xrightarrow[d\ n\to\infty]{} N\left(\mu = \frac{2N_+N_-}{N} + 1;\ \sigma = \sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}\right)$$

Ejecución del test:

1. Seleccionamos un punto de corte (**mediana**, moda, media, u otro)
2. Separamos los elementos en dos categorías(> punto corte, < punto de corte)
3. Contamos tantas rachas como cambios de categorías se produzcan a lo largo de la extracción: el resultado depende del orden de la muestra…
4. Fijamos α y comparamos el estadístico con su distribución asintótica.

2.3. PRUEBA DE KOLMOGOROV-SMIRNOV PARA UNA MUESTRA

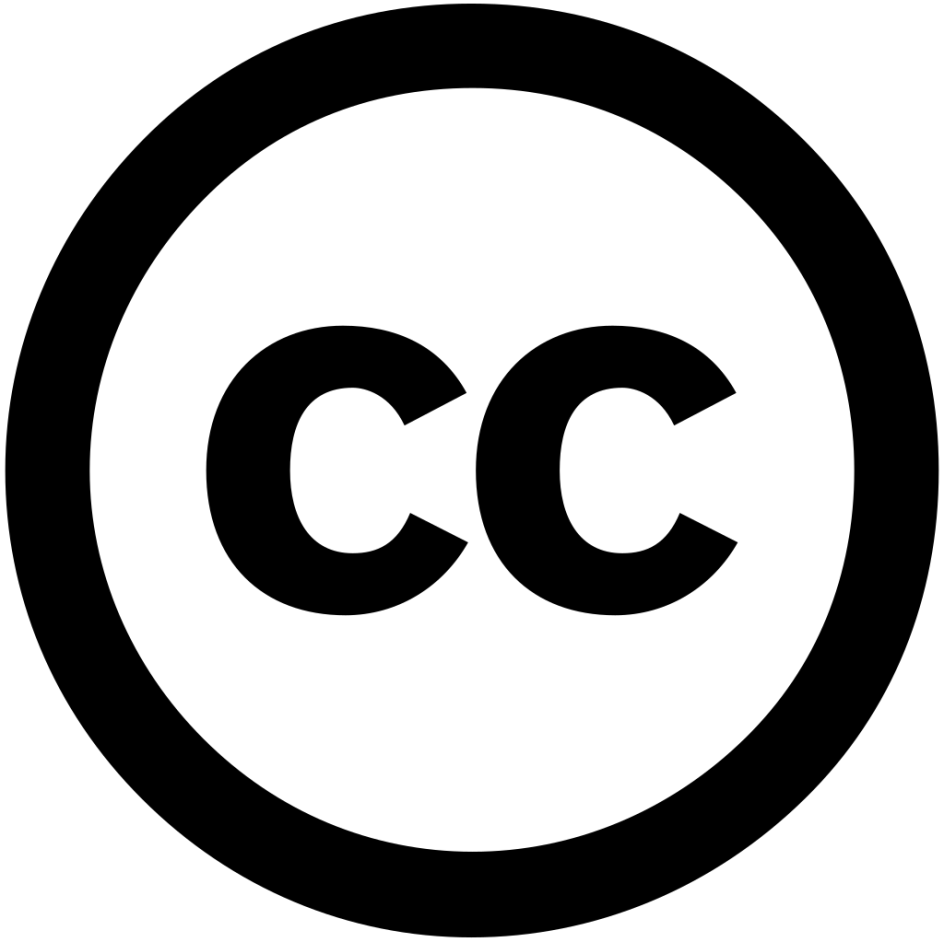
- H_0: la muestra procede de una población con función de distribución conocida F(x) de tipo continuo.
- $H_0\ cierta \rightarrow |F_n(x) - F(x)|\ pequeñas\ si\ n$ suficientemente grande $F_n(x_i) = \frac{N_i}{N}$ (función de distribución empírica de la muestra = Frecuencia relativa acumulada $N_i = \sum_{i=1}^{i} n_i$)
- Estadístico de Kolmogorov-Smirnov

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

Que cumple

1. $P\left(\lim_{n\to\infty} D_n = 0\right) = 1$
2. La distribución de $D_n\ indep.de\ F(x)$
3. La distribución asintótica de $D_n\ tb\ indep. F(x)$

$$RC \rightarrow \alpha = P(D_n \geq k; H_0\ cierta) = P(D_n \geq k; \xi: F(x))$$



www.pacorabadan.com

FUENTES Y RECURSOS

- Martín Pliego, Fundamentos de Inferencia Estadística, Editorial AC, 3ªEd, 2005
- Apuntes de don Antonio Franco Rodríguez de Lázaro y Pilar Ordás del Amo.
- Otros recursos en www.pacorabadan.com
- Aula virtual.