TEMA 9

CONTRASTES NO PARAMÉTRICOS

9.6. *INTRODUCCIÓN*

- Para no exigir robustez sobre la hipótesis de normalidad de la distribución subyacente, se utilizan técnicas no paramétricas
- Dos situaciones
  - o Técnicas No Paramétricas en sentido estricto: estadístico sin referencia a parámetro poblacional
  - o Métodos de distribución libre: distribución del estadístico no depende de la de la población
- Utilizan frecuencias y no estadísticos
  - o Frecuencia absoluta n_i
  - o Frecuencia absoluta acumulada N_i
- Ejemplos
  - o Bondad de ajuste (muestra de una determinada población)
  - o Independencia entre niveles de dos factores

9.7. *CONTRASTES DE BONDAD DE AJUSTE*

- Objetivo: verificar si una muestra procede de una población con una determinada distribución de probabilidad

  H_0: SÍ procede de ξ

  H_1: NO procede de ξ

- Procedimiento: Ajuste de una ξ teórica ala muestra X

  H_0: ajuste bueno

  H_1: ajuste no bueno

9.7.1. *TEST χ^2*

9.7.1.1. ξ no tiene parámetros

- Procedimiento
  - o Se clasifican las muestras en *r* clases mutuamente excluyentes y que cubren todo el campo de variación de la distribución ξ

    Clases: A_1, A_2, A_3, ..., A_r

- Se construye la distribución de frecuencias absolutas de la muestra

x_i	**n_i**
A_1	n_1
…	
A_r	n_r

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

- Se construye la distribución de frecuencias absolutas esperadas E_i de la población

x_i	**E_i**
A_1	E_1
…	
A_r	E_r

$$\sum_{i=1}^{r} E_i = n$$

- Se establece la medida de discrepancia

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

- Se considera región crítica

$$X^2 \geq k$$

* Cálculo de E_i
    - Si H_0 cierta, se puede calcular $P(A_i)=p_i$, resultando $\sum_{i=1}^{r} p_i = 1$
    - Entonces, la distribución de las frecuencias muestrales es multinomial

$$P(n_1, n_2, ..., n_r) = \frac{n!}{n_1! n_2! ... n_r!} p_1^{n_1} p_2^{n_2} ... p_r^{n_r}$$

por ser cada muestra individual independiente

- Entonces, cada n_i tiene una distribución binomial $B(n, p_i)$, cuyo valor esperado será E_i:

$E_i=E(B(n, p_i))= np_i$ "Reparto de n según probabilidad de clase"

$$\sum_{i=1}^{r} E_i = \sum_{i=1}^{r} np_i = n\sum_{i=1}^{r} p_i = n(1) = n$$

- Distribución de X^2
  - Como $B(n, p_i)$, converge en Poisson $Po(np_i)$
  - Por Teorema Central del Límite

$$\xi_i = \frac{n_i - np_i}{\sqrt{np_i}} \to N(0,1)$$

$$\xi_i^2 = \frac{(n_i - np_i)^2}{np_i}$$

Entonces $X^2 = \sum_{i=1}^{r} \xi_i^2 \to \chi_r^2$

Pero como $\sum_{i=1}^{r} n_i = n \to \chi_r^2$

- Luego la región crítica es:

$$X^2 \geq k = \chi^2_{r-1,\alpha}$$

  - Observaciones: $E_i \geq 5$

- Ejemplo: Al nivel de significación del 5%, se puede decir que hay diferencia de opinión entre la población si de 5000 personas 2449 contestan SÍ y el resto NO a una pregunta

  H_0: ajuste bueno a p_i=1/2

  H_1: ajuste malo a p_i=1/2

A_i	n_i	p_i	$E_i = np_i$	X^2	
SÍ	2449	$\frac{1}{2}$	$\frac{1}{2}5000 = 2500$	$\frac{(2449-2500)^2}{2500}$	1.0404
NO	2551	$\frac{1}{2}$	$\frac{1}{2}5000 = 2500$	$\frac{(2551-2500)^2}{2500}$	1.0404
	5000		5000		**2.0808**

$k = \chi^2_{2-1,0.05} = 3.8415$

Como $X^2 \leq k$, no hay evidencia para rechazar H_0

9.7.1.2. ξ tiene k parámetros desconocidos

- Procedimiento
  - El mismo pero $X^2 \to \chi^2_{r-k-1}$
- Ejemplo: Una compañía de seguros registra los accidentes de automóvil de una ciudad durante 100 días, obteniendo la siguiente información:

Nº Accidentes	0	1	2	3	4	5
Nº días	40	34	16	7	2	1

Establecer una hipótesis acerca de la población que representa el evento número de accidentes por día y contrastarla al nivel de significación del 5%.

- H_0: ajuste bueno a Poisson (λ)

  H_1: ajuste malo

  Como $\lambda^* = \bar{x} = 1$
- $E_i = np_i$

  $p_i = P(\xi = i) = \frac{\lambda^i e^{-\lambda}}{i!}$

A_i	n_i	p_i	$E_i = np_i$	n_i	$E_i = np_i$	X^2	
0	40	$\frac{1^0 e^{-1}}{0!} = 0.3679$	36.79	40	36.79	$\frac{(40-36.79)^2}{36.79}$	0.2805
1	34	$\frac{1^1 e^{-1}}{1!} = 0.3679$	36.79	34	36.79	$\frac{(34-36.79)^2}{36.79}$	0.2113
2	16	0.1839	18.39	16	18.39	$\frac{(16-18.39)^2}{18.39}$	0.3116
3	7	0.0613	6.13	10	8.02	$\frac{(10-8.03)^2}{8.03}$	0.4906
4	2	0.0153	*1.53*				
5	1	0.0031	*0.31*				
6	0	0.0005	*0.05*				
	100	1	100	100	100		**1.2904**

$k = \chi^2_{4-1-1,0.05} = 5.99$

Como $X^2 \le k$, no hay evidencia para rechazar H_0

9.6. *TABLAS DE CONTINGENCIA*

- Definición: Una tabla de contingencia *r*x*c* es una tabla de doble entrada donde se recoge la frecuencia conjunta nij con que aparece el nivel i-ésimo de un factor A con *r* niveles y el nivel j-ésimo de un factor B con *c* niveles:

	B_1 … B_c	
A_1		$n_{1.}$
…	n_{ij}	
A_r	n_{rc}	$n_{r.}$
	$n_{.1}$ $n_{.c}$	n

9.6.1. *CONTRASTE DE INDEPENDENCIA*

- Hipótesis a contrastar:

  H_0: Independencia entre niveles de A y B

  H_1: Dependencia entre niveles de A y B

- Condición de independencia:

  $p_{ij}=p_{i.}p_{.j}=P(A_i \cap B_j)=P(A_i)P(B_j)$

- Estadístico de contraste:

  Como $p^*_{i.MV}=\frac{n_{i.}}{n}$ y $p^*_{.j\,MV}=\frac{n_{.j}}{n}$

$$E_{ij}{}^{*}=np_{ij}{}^{*}=n\frac{n_{i.}}{n}\frac{n_{.j}}{n}=\frac{n_{i.}n_{.j}}{n}$$

$$X^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-E_{ij}\right)^2}{E_{ij}}\rightarrow\chi^2_{rc-(r+c-2)-1}=\chi^2_{(r-1)(c-1)}$$

- Región crítica:

$$X^2\geq k=\chi^2_{(r-1)(c-1),\alpha}$$

- Ejemplo: Una empresa dedicada a la venta de automóviles desea determinar si la edad de sus clientes potenciales puede explicar o no la preferencia por tres modelos que va al lanzar al mercado. Consultados 200 clientes habituales, la información obtenida es la siguiente

		MODELO I	II	III
	20-30	10.00	40.00	10.00
EDAD	30-40	30.00	30.00	20.00
	40-50	10.00	30.00	20.00

¿Aceptaría la empresa a la vista de esta información que la edad media explicará la preferencia por el modelo al nivel de significación del 5%?

Solución:

H$_0$: Independencia entre niveles de Edad y Modelo

H$_1$: Dependencia entre niveles de Edad y Modelo

OBSERVADA

	1	2	3	4	5	
1	10.00	40.00	10.00			60.00
2	30.00	30.00	20.00			80.00
3	10.00	30.00	20.00			60.00
4						0.00
5						0.00
	50.00	100.00	50.00	0.00	0.00	200.00

TEÓRICA

	1	2	3	4	5	
1	15.00	30.00	15.00	0.00	0.00	60.00
2	20.00	40.00	20.00	0.00	0.00	80.00
3	15.00	30.00	15.00	0.00	0.00	60.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00
	50.00	100.00	50.00	0.00	0.00	200.00

DIFERENCIA

	1	2	3	4	5	
1	1.67	3.33	1.67	0.00	0.00	6.67
2	5.00	2.50	0.00	0.00	0.00	7.50
3	1.67	0.00	1.67	0.00	0.00	3.33
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00
	8.33	5.83	3.33	0.00	0.00	17.5000

$$k = \chi^2_{(3-1)(3-1),0.05} = 9.4877$$

$\therefore$Como $X^2 \geq k$, hay suficiente evidencia para rechazar la hipótesis de independencia

9.6.2. *CONTRASTE DE HOMOGENEIDAD*

- *c* muestras clasificadas según *r* niveles:
- Hipótesis a contrastar:

  H$_0$: Las *c* muestras SÍ corresponden a la misma población

  H$_1$: Las *c* muestras NO corresponden a la misma población

- Problema: Para examinar las consecuencias de un conjunto de medidas de política económica se efectúan dos sondeos en dos momentos distintos entre un grupo de 2500 empresas dedicadas a la exportación obteniéndose los siguientes resultados:

	SONDEO I	SONDEO II
INCREMENTO	60%	58%
DISMINUCIÓN	40%	42%

¿Puede afirmarse al nivel de significación del 5% que se han modificado las políticas de exportación?

Solución:

H_0: Las *2* muestras SÍ corresponden a la misma población, o no ha habido cambio de políticas

H_1: Las *2* muestras NO corresponden a la misma población

OBSERVADA

	1	2	3	4	5	
1	1500.00	1450.00				2950.00
2	1000.00	1050.00				2050.00
3						0.00
4						0.00
5						0.00
	2500.00	2500.00	0.00	0.00	0.00	5000.00

TEÓRICA

	1	2	3	4	5	
1	1475.00	1475.00	0.00	0.00	0.00	2950.00
2	1025.00	1025.00	0.00	0.00	0.00	2050.00
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00
	2500.00	2500.00	0.00	0.00	0.00	5000.00

DIFERENCIA

	1	2	3	4	5	
1	0.42	0.42	0.00	0.00	0.00	0.85
2	0.61	0.61	0.00	0.00	0.00	1.22
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00
	1.03	1.03	0.00	0.00	0.00	2.0670

$$k = \chi^2_{(2-1)(2-1),0.05} = 3.8415$$

$\therefore$Como $X^2 \leq k$, no hay suficiente evidencia para rechazar la hipótesis de homogeneidad, con lo que no se aprecian cambios significativos de políticas de exportación