Asignatura .... Grupo ....

Apellidos .... Nombre ....

Ejercicio del día ....

Facultad de Ciencias Económicas y Empresariales   Estadística Teórica II

1. Según los dirigentes del partido A, la intención de voto del partido rival B, en Andalucía, es la misma que la que tiene en Madrid. Se realiza una encuesta a 100 personas en Andalucía de los que 25 mostraron su apoyo al partido B, y a otras 100 personas en Madrid de las que 30 se inclinaron por el partido B.

a) Construir un intervalo de confianza del 90% para la proporción de personas que votarían al partido B en Andalucía

b) ¿A cuántas personas habría que encuestar para obtener un margen de error o error de estimación $\pm$ 2%, al nivel de confianza anterior?.

c) Construir un intervalo de confianza al 90% para la diferencia de proporciones en la estimación del voto del partido B en las dos comunidades. ¿Podemos afirmar que los dirigentes del partido A tienen razón?.

<u>Solución:</u>

a) La característica en estudio en ambas comunidades es dicotómica, tenemos que construir un intervalo de confianza para el parámetro p_1 (proporción) de la variable aleatoria binomial asociada al estudio de la característica en la comunidad de Andalucía. Como el tamaño de la muestra es suficientemente grande, $n_1 = 100$, se puede utilizar la aproximación normal.

$$I_{1-\alpha}(p) = \left[\hat{p} \pm z_{(\alpha/2)}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right] \quad \begin{cases} \hat{p}_1 = 25/100 = 0{,}25 & \hat{q}_1 = 1-\hat{p}_1 = 0{,}75 \qquad n_1 = 100 \\ 1-\alpha = 0{,}90 \quad \alpha = 0{,}10 & \alpha/2 = 0{,}05 \quad z_{\alpha/2} = z_{0,05} = 1{,}645 \end{cases}$$

$$I_{0,90}(p_1) = \left[0{,}25 \pm (1{,}645)\sqrt{\frac{0{,}25\,.\,0{,}75}{100}}\right] = \left[0{,}179\ ;\ 0{,}321\right]$$

$$I_{0,90}(p_1) = \left[0{,}179\ ;\ 0{,}321\right] \equiv P\left[0{,}179 \le p_1 \le 0{,}321\right] = 0{,}90 = 1-\alpha$$

En Andalucía la intención de voto del partido B se encuentra entre el 17,9% y 32,1%, con un nivel de confianza del 90%.

b) La amplitud o longitud vendrá dado por la fórmula:

$$I = \left[\begin{pmatrix}\text{proporción}\\ \text{muestral}\end{pmatrix} \pm \begin{pmatrix}\text{error}\\ \text{muestral}\end{pmatrix}\right] \equiv \left[\hat{p}_x + z_{(\alpha/2)}\sqrt{\frac{\hat{p}_x.\hat{q}_x}{n}}\right] \quad \mapsto \quad \in^2 = \left(z_{\alpha/2}\sqrt{\frac{\hat{p}_x.\hat{q}_x}{n}}\right)^2$$

de donde, $n = \dfrac{(z_{\alpha/2})^2\,(\hat{p}_x.\hat{q}_x)}{\in^2}$

El caso más desfavorable será cuando $\hat{p}_x = \hat{q}_x = 0,5$.

Siendo $\in^2 = (\pm 0,02)^2 = 0,0004 \quad \mapsto \quad n = \dfrac{(1,645)^2\,(0,5.0,5)}{0,0004} \approx 1691$

c) Nos encontramos ante un intervalo de confianza para la diferencia de parámetros poblacionales $(p_1 - p_2)$ de dos distribuciones binomiales, con el tamaño de las muestras suficientemente grandes, $n_1 = n_2 = 100$, para utilizar la aproximación normal.

$$I_{1-\alpha}(p_1 - p_2) = \left[(\hat{p}_1 - \hat{p}_2) \pm z_{(\alpha/2)}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}\right]$$

$$\begin{cases} \hat{p}_1 = 25/100 = 0,25 \quad \hat{q}_1 = 1-\hat{p}_1 = 0,75 \quad n_1 = 100 \\ \hat{p}_2 = 30/100 = 0,3 \quad \hat{q}_2 = 1-\hat{p}_2 = 0,70 \quad n_2 = 100 \\ 1-\alpha = 0,90 \quad \alpha = 0,10 \quad \alpha/2 = 0,05 \quad z_{\alpha/2} = z_{0,05} = 1,645 \end{cases}$$

$$I_{0,90}(p_1 - p_2) = \left[(0,25 - 0,3) \pm (1,645)\sqrt{\frac{0,25\,.\,0,75}{100} + \frac{0,3\,.\,0,70}{100}}\right] = [-0,153\ ;\ 0,053]$$

El intervalo de confianza cubre el cero, lo que indica que no existe diferencia significativa entre la intención de voto del partido B en ambas comunidades, con lo cual los dirigentes del partido A tienen razón con una fiabilidad del 90%.

2. Sea una variable aleatoria X procedente de una población con densidad de probabilidad $N(\mu, 4)$. Se quiere contrastar la hipótesis nula $H_0 : \mu = \mu_0 = 10$ frente a la hipótesis alternativa $H_1 : \mu = \mu_1 = 12$, con un nivel de significación $\alpha = 0{,}05$, con un muestreo simple de tamaño 25.

Determinar:

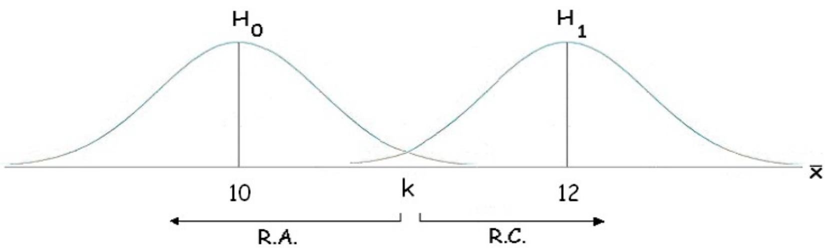
  a) La probabilidad de cometer el error tipo II
  b) La potencia del contraste

Solución:

Sea la variable aleatoria $X \approx N(\mu, 4)$

Las hipótesis sobre la media poblacional (contraste unilateral):

$$H_0 : \mu_0 = 10$$
$$H_1 : \mu_1 = 12$$



Regla de decisión $\begin{cases} \bar{x} \le k & \text{se acepta } H_0 \mapsto \text{ región aceptación (R.A.)} \\ \bar{x} > k & \text{no se acepta } H_0 \mapsto \text{ región crítica (R.C.)} \end{cases}$

La distribución de la media muestral $\bar{x}$, de tamaño 25, con la varianza poblacional $\sigma^2 = 16$ conocida:

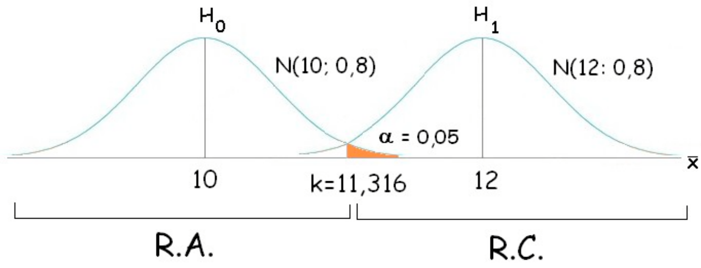
$$H_0 : \quad \bar{x} \in N(\mu_0, \frac{\sigma}{\sqrt{n}}) = N(10, \frac{4}{\sqrt{25}}) = N(10 ; 0{,}8)$$

$$H_1 : \quad \bar{x} \in N(\mu_1, \frac{\sigma}{\sqrt{n}}) = N(12, \frac{4}{\sqrt{25}}) = N(12 ; 0{,}8)$$

a) Para hallar el valor crítico 'k' recurrimos al Error Tipo I:

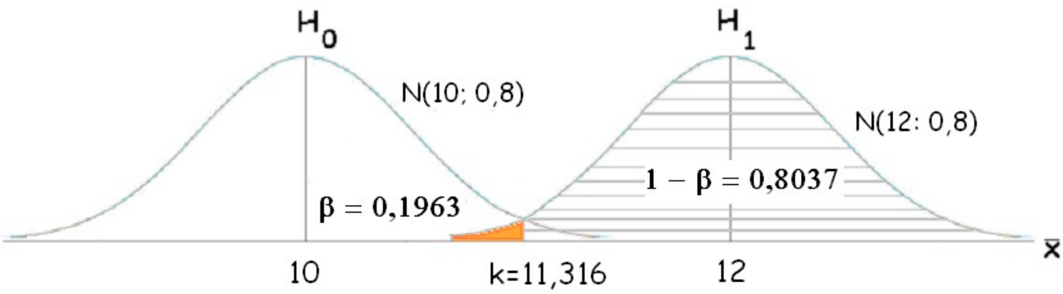
$$\alpha = P(\text{ET I}) = P(\text{Rechazar } H_0 \,/\, H_0 \text{ cierta}) = 0{,}05$$

$$\alpha = P(ET\ \ I) = P(\overline{x} > k \,|\, H_0 : \mu_0 = 10) =$$
$$= P\left(z > \frac{k-10}{0,8}\right) = 0,05$$
$$\frac{k-10}{0,8} = 1,645 \quad \Rightarrow \quad K = 11,316$$



Error Tipo II: $\beta = P(ET\ II) = P(\text{Aceptar } H_0 \ / \ H_0 \text{ falsa})$

$$\beta = P(ET\ \ II) = P(\overline{x} \le 11,316 \,|\, H_1 : \mu_1 = 12) = P\left(z \le \frac{11,316-12}{0,8}\right) = P(z \le -0,855) =$$
$$= P(z \ge 0,855) = 0,1963$$



b) Potencia del Contraste: $\text{Potencia} = P(\text{Rechazar } H_0 \,|\, H_0 \text{ falsa}) = 1 - \beta$

$$1-\beta = P(\text{Rechazar } H_0 \,|\, H_0 \text{ falsa}) = 1 - 0,1963 = 0,8037$$

3. Las latas de mejillones de una determinada marca indican que el peso escurrido de dicho producto es de 250 gr. No obstante, un consumidor está convencido de que el peso escurrido medio de dicho producto es menor que el que indican las latas. Si el peso escurrido sigue una ley normal con desviación típica 9 gr.

a) Determinar, si existe, la mejor región crítica para contrastar la mejor región crítica, con un nivel de significación del 5% y muestras aleatorias simples de tamaño 100.

b) Tomar una decisión acerca del rechazo o no de la hipótesis nula a partir de una muestra aleatoria simple de tamaño 100 en la cual se ha observado un peso escurrido promedio de 245 gr.

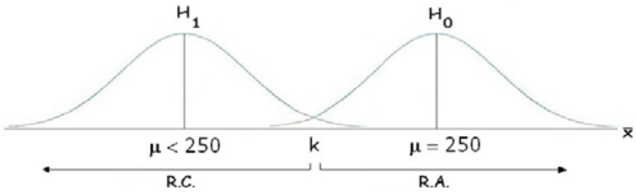
c) Determinar la función de potencia del contraste.

Solución:

a) Sea la variable aleatoria X = "peso escurrido de las latas de mejillones"

Se trata de un contraste unilateral:

$$H_0 : \mu = 250 \qquad H_1 : \mu < 250$$



La regla de decisión del muestreo: $\begin{cases} \overline{x} > k & \text{se acepta } H_0 \text{ (R.A.)} \\ \overline{x} \leq k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases}$

La variable aleatoria X en el muestreo, bajo la hipótesis nula, sigue una distribución $N\left[250, \frac{9}{\sqrt{100}}\right]$

El valor crítico k, bajo la hipótesis nula, se determina con el nivel de significación α:

$$\alpha = P\,(ET\ \ I) = P\,(\text{Rechazar } H_0\ /\ H_0 \text{ cierta}) = P\,(\overline{x} \leq k\ /\ H_0 \text{ cierta}) =$$

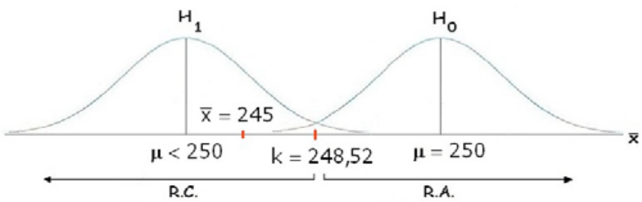
$$= P\left[\,\overline{x} \leq k\ /\ N(250\,;\,0{,}9)\,\right] = P\left[\frac{\overline{x} - 250}{0{,}9} \leq \frac{k - 250}{0{,}9}\right] = P\left[z \leq \frac{k - 250}{0{,}9}\right] = 0{,}05$$

observando las tablas de la N(0,1), y considerando que $z_{\alpha} = -z_{1-\alpha}$, se tiene:

$$\frac{k - 250}{0{,}9} = -1{,}96 \;\mapsto\; k = 248{,}52$$

La región crítica más potente, para muestras de tamaño 100, es $\overline{x} \leq 248{,}52$

b) Dado que $\overline{x} = 245 < 248{,}52$, el peso escurrido promedio se encuentra en la región de rechazo de la hipótesis nula.



c) La función potencia del contraste se establece como:

$$P(\mu) = P(\overline{x} \leq 248{,}52) = P\left[\ \overline{x} \leq k \ / \ N(\mu\,;\ \frac{9}{\sqrt{100}})\ \right] = P\left[\frac{\overline{x}-\mu}{0{,}9} \leq \frac{248{,}52-\mu}{0{,}9}\right] =$$

$$= P\left[z \leq \frac{248{,}52-\mu}{0{,}9}\right]_{\mu\,\leq\,250}$$

4. Para una muestra aleatoria simple de 350 días, el número de urgencias tratadas diariamente en un hospital A queda reflejado en la siguiente tabla:

Nº urgencias	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 - 30	Total días
Nº días	20	65	100	95	60	10	350

Contrastar, con un nivel de significación del 5%, si la distribución del número de urgencias tratadas diariamente en el hospital A se ajusta a una distribución normal.

<u>Solución.-</u>

Para ajustar los datos obtenidos a una distribución normal $N(\mu, \sigma)$ de parámetros desconocidos, se necesitan estimar los dos parámetros recurriendo a los estimadores máximo-verosímiles: $(\hat{\mu} = \overline{x}\ ,\ \hat{\sigma}^2 = \sigma_x^2)$, donde la variable aleatoria X = ' número de urgencias diarias'.

Se establece la hipótesis nula H_0 : 'La distribución empírica se ajusta a la normal'

Se acepta la hipótesis nula, a un nivel de significación α si

$$\chi^2_{k-p-1} = \underbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n \ < \ \underbrace{\chi^2_{\alpha\,;\,k-p-1}}_{\text{estadístico teórico}} \quad \text{donde} \quad \begin{array}{l} k \equiv \text{número intervalos} \\ p \equiv \text{número parámetros a estimar} \end{array}$$

1) Se obtiene la media y la desviación típica:

Intervalos	x_i	n_i	$x_i\,n_i$	$x_i^2\,n_i$
0 - 5	2,5	20	50	125
5 - 10	7,5	65	487,5	3656,25
10 - 15	12,5	100	1250	15625
15 - 20	17,5	95	1662,5	29093,75
20 - 25	22,5	60	1350	30375
25 - 30	27,5	10	275	7562,5
		$n=\sum_{i=1}^{6} n_i = 350$	$\sum_{i=1}^{6} x_i\,n_i = 5075$	$\sum_{i=1}^{6} x_i^2\,n_i = 86437,5$

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i\,n_i}{350} = 14,5 \qquad \sigma_x^2 = \frac{\sum_{i=1}^{6}(x_i-\bar{x})^2\,n_i}{350} = \frac{\sum_{i=1}^{6} x_i^2\,n_i}{350} - (\bar{x})^2 = 36,71 \qquad \sigma_x = 6,06$$

2) Se procede al ajuste de una distribución normal $N(14,5\ ;\ 6,06)$ hallando las probabilidades de cada uno de los intervalos:

Intervalos	n_i	p_i	$e_i = p_i\,n$	$(n_i - e_i)^2$	$(n_i - e_i)^2 / e_i$
0 - 5	20	0,0498	17,43	6,6	0,38
5 - 10	65	0,1714	59,99	25,1	0,42
10 - 15	100	0,3023	105,81	33,76	0,32
15 - 20	95	0,2867	100,35	28,62	0,29
20 - 25	60	0,1396	48,86	124,1	2,54
25 - 30	10	0,0366	12,81	7,9	0,62
	$n = 350$				$\sum_{i=1}^{6}(n_i - e_i)^2 / e_i = 4,57$

$$P(0 < x < 5) = P\left[\frac{0-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{5-14,5}{6,06}\right] = P(-2,39 < z < -1,57) =$$
$$= P(1,57 < z < 2,39) = P(z > 1,57) - P(z > 2,39) = 0,0582 - 0,00842 = 0,04978$$

$$P(5 < x < 10) = P\left[\frac{5-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{10-14,5}{6,06}\right] = P(-1,57 < z < -0,74) =$$
$$= P(0,74 < z < 1,57) = P(z > 0,74) - P(z > 1,57) = 0,2296 - 0,0582 = 0,1714$$

$$P(10 < x < 15) = P\left[\frac{10-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{15-14,5}{6,06}\right] = P(-0,74 < z < 0,08) =$$
$$= P(0,08 < z < 0,74) = 1 - P(z > 0,74) - P(z > 0,08) = 1 - 0,4681 - 0,2296 = 0,3023$$

$$P(15 < x < 20) = P\left[\frac{15-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{20-14,5}{6,06}\right] = P(0,08 < z < 0,91) =$$
$$= P(z > 0,08) - P(z > 0,91) = 0,4681 - 0,1814 = 0,2867$$

$$P(20 < x < 25) = P\left[\frac{20-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{25-14,5}{6,06}\right] = P(0,91 < z < 1,73) =$$
$$= P(z > 0,91) - P(z > 1,73) = 0,1814 - 0,0418 = 0,1396$$

$$P(25 < x < 30) = P\left[\frac{25-14,5}{6,06} < \frac{x-14,5}{6,06} < \frac{30-14,5}{6,06}\right] = P(1,73 < z < 2,56) =$$
$$= P(z > 1,73) - P(z > 2,56) = 0,0418 - 0,0052 = 0,0366$$

3) Se calculan las frecuencias esperadas, multiplicando las probabilidades por el número total $e_i = p_i . n$

4) Se calcula el estadístico de contraste χ^2, donde el número de grados de libertad es $k - p - 1 =$ (nº intervalos) $-$ (nº parámetros a estimar) $- 1 = 6 - 2 - 1 = 3$, con lo cual,

$$\chi_3^2 = \sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = 4,57$$

Por otra parte, el estadístico teórico $\chi^2_{0,05\,;\,3} = 7,815$

Como $\chi_3^2 = 4,57 < \chi^2_{0,05\,;\,3} = 7,815$, se acepta la hipótesis nula a un nivel de significación del 5%. En consecuencia, la variable aleatoria número de urgencias en el hospital A sigue una distribución $N(14,5\,;\,6,06)$.

5. Las compañías de seguros de automóviles suelen penalizar en sus primas a los conductores más jóvenes, con el criterio que éstos son más propensos a tener un mayor número de accidentes. En base a la tabla adjunta, con un nivel de significación del 5%, contrastar si el número de accidentes es independiente de la edad del conductor.

Edad del conductor	Número de accidentes al año				
	0	1	2	3	4
25 o menos	10	10	20	40	70
26 - 35	20	10	15	20	30
más de 36	60	50	30	10	5

Solución.-

La hipótesis nula H_0: 'El número de accidentes sufridos por los conductores no depende de la edad del conductor'

Se acepta H_0:

$$\chi^2_{(k-1).(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(k-1).(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha;(k-1).(m-1)}$$

Región de rechazo de la hipótesis nula: $R = \left\{\chi^2_{(k-1).(m-1)} \geq \chi^2_{\alpha\,;\,(k-1).(m-1)}\right\}$

Se forma la tabla de contingencia 3 x 5 donde cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\,;\,j=1,\cdots,m}$ tiene una frecuencia teórica o esperada en caso de independencia

$$e_{ij} = p_{ij}\,n = \frac{n_{x_i}\,n_{y_j}}{n}$$

Edad del conductor	Número de accidentes por año					
	0	1	2	3	4	n_{x_i}
25 o menos	10 $e_{11}=33{,}75$	10 $e_{12}=26{,}25$	20 $e_{13}=24{,}37$	40 $e_{14}=26{,}25$	70 $e_{15}=39{,}37$	150 (150)
26 - 35	20 $e_{21}=21{,}37$	10 $e_{22}=16{,}62$	15 $e_{23}=15{,}44$	20 $e_{24}=16{,}62$	30 $e_{25}=24{,}94$	95 (95)
más de 36	60 $e_{31}=34{,}87$	50 $e_{32}=27{,}12$	30 $e_{33}=25{,}19$	10 $e_{34}=27{,}12$	5 $e_{35}=40{,}69$	155 (155)
n_{y_j}	90	70	65	70	105	400

donde,

$$e_{11} = \frac{150.90}{400} = 33,75 \qquad e_{21} = \frac{95.90}{400} = 21,37 \qquad e_{31} = \frac{155.90}{400} = 34,87$$

$$e_{12} = \frac{150.70}{400} = 26,25 \qquad e_{22} = \frac{95.70}{400} = 16,62 \qquad e_{32} = \frac{155.70}{400} = 27,12$$

$$e_{13} = \frac{150.65}{400} = 24,37 \qquad e_{23} = \frac{95.65}{400} = 15,44 \qquad e_{33} = \frac{155.65}{400} = 25,19$$

$$e_{14} = \frac{150.70}{400} = 26,25 \qquad e_{24} = \frac{95.70}{400} = 16,62 \qquad e_{34} = \frac{155.70}{400} = 27,12$$

$$e_{15} = \frac{150.105}{400} = 39,37 \qquad e_{25} = \frac{95.105}{400} = 24,94 \qquad e_{35} = \frac{155.105}{400} = 40,69$$

Estadístico observado: $\chi^2_{(3-1).(5-1)} = \chi^2_8 = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{n_{ij}^2}{e_{ij}} - n =$

$$= \left(\frac{10^2}{33,75} + \frac{10^2}{26,25} + \frac{20^2}{24,37} + \frac{40^2}{26,25} + \frac{70^2}{39,37}\right) + \left(\frac{20^2}{21,37} + \frac{10^2}{16,62} + \frac{15^2}{15,44} + \frac{20^2}{16,62} + \frac{30^2}{24,94}\right) +$$

$$+ \left(\frac{60^2}{34,87} + \frac{50^2}{27,12} + \frac{30^2}{25,19} + \frac{10^2}{27,12} + \frac{5^2}{40,69}\right) - 400 = 143,51$$

Estadístico teórico: $\chi^2_{0,05\,;(3-1).(5-1)} = \chi^2_{0,05\,;8} = 15,507$

Como $\chi^2_8 = 143,51 > 15,507 = \chi^2_{0,05\,;8}$ se rechaza la hipótesis nula de independencia entre la edad del conductor y el número de accidentes; es decir, la edad influye significativamente en el número de accidentes al año.