eman ta zabal zazu
EUSKAL HERRIKO UNIBERTSITATEA
UNIVERSIDAD DEL PAIS VASCO

Departamento de Matemática Aplicada
Plaza de la Casilla, 3
48012 Bilbao

INDUSTRIA INGENIARITZA TEKNIKOKO UNIBERTSITATE ESKOLA
ESCUELA UNIVERSITARIA DE INGENIERÍA TÉCNICA INDUSTRIAL
BILBAO

MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

CONVOCATORIA DE MAYO (2011)

EJERCICIO 1

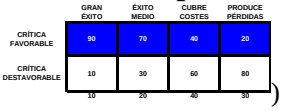
El director de publicaciones de una editorial trata de decidir si debe publicar un nuevo texto de estadística. Los anteriores libros de texto publicados indican que el 10 % son grandes éxitos, el 20 % tienen un éxito moderado, el 40 % cubren los costos y el 30 % producen pérdidas. Antes de tomar una decisión se somete la obra a la decisión de los críticos. En casos anteriores recibieron críticas favorables el 90 % de los grandes éxitos, el 70 % de los de éxito moderado, el 40 % de los que cubrieron costos y el 20 % de los que produjeron pérdida.

El espacio muestral se organizar de la siguiente manera:

	GRAN ÉXITO	**ÉXITO MEDIO**	**CUBRE COSTES**	**PRODUCE PÉRDIDAS**
CRÍTICA FAVORABLE	**90**	**70**	**40**	**20**
CRÍTICA DESTAVORABLE	**10**	**30**	**60**	**80**
	10	**20**	**40**	**30**

donde cada una de las 8 zonas representa sucesos incompatibles (mutuamente excluyentes). A partir de dicha información:

(**A**) ¿Qué proporción aproximada de libros recibe crítica favorable?

Aplicando el teorema de la probabilidad total se tiene que (zona azul en la siguiente figura 

)

$$p(CF) = p(CF \mid GE)\,p(GE) + p(CF \mid EM)\,p(EM) + p(CF \mid CC)\,p(CC) + p(CF \mid PP)\,p(PP) =$$
$$0.90 \times 0.10 + 0.70 \times 0.20 + 0.40 \times 0.40 + 0.20 \times 0.30 = 0.45$$

(**B**) Si el texto recibe crítica favorable, ¿cuál es la probabilidad de que tenga éxito?

Aplicando el teorema de Bayes (zona roja en el espacio muestral correspondiente)

	GRAN ÉXITO	ÉXITO MEDIO	CUBRE COSTES	PRODUCE PÉRDIDAS
CRÍTICA FAVORABLE	90	70	40	20
CRÍTICA DESTAVORABLE	10	30	60	80
	10	20	40	30

$$p\left[(GE \cup EM)\,|\,CF\right] = p(GE\,|\,CF) + p(EM\,|\,CF) = \frac{p(CF\,|\,GE)\,p(GE)}{p(CF)} + \frac{p(CF\,|\,EM)\,p(EM)}{p(CF)} =$$

$$\frac{0.90 \times 0.10 + 0.70 \times 0.20}{0.45} = \frac{0.23}{0.45} = 0.5111$$

(**C**) ¿Cuál es la probabilidad de que un texto tenga crítica favorable y éxito moderado a la vez?

Análogamente (ahora se trata de calcular la zona verde)

	GRAN ÉXITO	ÉXITO MEDIO	CUBRE COSTES	PRODUCE PÉRDIDAS
CRÍTICA FAVORABLE	90	70	40	20
CRÍTICA DESTAVORABLE	10	30	60	80
	10	20	40	30

$$p(EM \cap CF) = p(EM\,|\,CF)\,p(EM) = 0.70 \times 0.20 = 0.14$$

EJERCICIO 2

Se sabe que la probabilidad de que un estudiante de enseñanza primaria presente escoliosis es 0.004. De los siguientes 1875 estudiantes que se revisen:

Se trata de un proceso binomial de características $\mathrm{B}(n = 1875;\ p = 0.004)$ de la que no de disponen tablas, siendo la variable aleatoria X = "número de estudiantes de enseñanza primaria que presentan escoliosis". Ahora bien, dado que $np = 7.5 > 4$ y $nq = 1867.5 >> 4$. Las unidades del ejercicio son "estudiantes de enseñanza primaria que presentan escoliosis". En consecuencia, se puede utilizar la distribución normal para aproximar la distribución binomial inicial de acuerdo a

$$\mathrm{B}(n = 1875;\ p = 0.004) \ \rightarrow \mathrm{N}\left(\mu = np = 7.5;\ \sigma = \sqrt{npq} = 2.7331\right)$$

(**A**) ¿Cuál es el número medio de estudiantes que presentan escoliosis?

Se sabe que la esperanza matemática o valor esperado de una variable aleatoria X binomial viene dada por

$$E[X] = np = 7.5\ estudiantes$$

(**B**) ¿Cuál es la desviación estándar?

Análogamente, la desviación típica de una variable aleatoria binomial es

$$\sigma[X] = \sqrt{npq} = 2.7331\ estudiantes$$

(**C**) Encuentra la probabilidad de que al menos ocho alumnos y como máximo once alumnos presenten el problema.

Es decir:

$$\text{Ï}\,(8 \le X \le 11) \underset{\substack{\text{correccion}\\ \text{por continuidad}}}{=} \text{Ï}\,(7.5 \le X \le 11.5) \underset{B \to N}{=} \text{Ï}\,(z_1 = \frac{7.5-7.5}{2.7331} = 0 \le Z \le z_1 = \frac{11.5-7.5}{2.7331} = 1.4635) =$$

$$\square(z_1 = 0 \le Z \le z_2 = 1.4635) = \square(Z \le z_2 = 1.4635) - 0.5 = 0.928334693\text{-}0.5 = 0.428334693$$

Si se calcula a partir del experimento binomial original el resultado al que se llega es

$$\text{Ï}\,(8 \le X \le 11) = \text{Ï}\,(X \le 11) - \text{Ï}\,(X \le 7) = F(11) - F(7) = 0.396677281$$

EJERCICIO 3

El índice de resistencia de rotura, expresado en kg, de un determinado tipo de acero sigue una distribución normal con desviación típica 15.6 kg. Con una muestra de 5 de estos aceros, seleccionados al azar, se obtuvieron los siguientes índices: 280, 240, 270, 285, 270.

Los estadísticos muestrales son:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1345}{5} = 269\ kg$$

$$\hat{s} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = 17.4642\ kg$$

Se trata de un problema de medias aritméticas de pequeñas muestras (n = 5) de una población (hay una serie estadística únicamente). El modelo de probabilidad a utilizar es la t de Student.

(**A**) Calcular un intervalo de confianza para la media del índice de resistencia a la rotura de este tipo de acero, utilizando un nivel de confianza del 95%.

El estimador insesgado de varianza mínima (máxima información y robusto) es $\hat{\mu} = \overline{x}$. El error probable (error estándar o desviación típica de la estimación) viene dada por:

$$\sigma_{\hat{\mu}} = \frac{\hat{s}}{\sqrt{n}} = \frac{17.4642}{\sqrt{5}} = 7.8102\ kg$$

El intervalo de confianza viene definido por:

$$[\mathrm{l}, \mathrm{L}] = \left[\overline{x} - t_{\alpha,\nu}\sigma_{\hat{\mu}}, \overline{x} + t_{\alpha,\nu}\sigma_{\hat{\mu}}\right] = [247.32\ kg, 290.69\ kg]$$

donde $t_{\alpha,\nu} = t_{95\,\%,\,\nu=n-1=4\ gdl} \equiv t_{97.5\,\%,\,4\ gdl} = \pm 2.776445105$

(**B**) Con el mismo nivel de confianza, ¿cuál sería el mínimo tamaño muestral para obtener un error máximo de 5 kg en la estimación de la media?, ¿será suficiente con elegir una muestra de 30 cuerdas?

El error viene dado por

$$t_{\alpha,\nu}\sigma_{\hat{\mu}} = t_{\alpha,\nu}\frac{\hat{s}}{\sqrt{n}} \le e_{\max} = 5\ kg \Leftrightarrow n \ge \left(t_{\alpha,\nu}\frac{\hat{s}}{e_{\max}}\right)^2 = \left(\frac{2.776445105 \times 17.4642}{5}\right)^2 = 94.05$$

con lo que se deduce que $n_{\min} = 95$, al menos. En consecuencia, con n = 30 se obtendría un error aún mayor, con lo que no sería suficiente.

EJERCICIO 4

Se trata de establecer si dos métodos de medición de la temperatura son numéricamente equivalentes, salvo por la imprecisión estadística. En otras palabras, se reconoce que las dos técnicas no pueden suministrar valores similares aún disponiendo de especímenes idénticos, pero lo que se trata de establecer es "si en promedio" se obtendrán valores cercanos de temperatura para sujetos aproximadamente idénticos. Para ello se han tomado las siguientes series estadísticas de medidas (en º C):

Dato	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Método A	338	243	267	195	203	262	225	214	292	218
Método B	327	248	246	192	222	261	223			

Los estadísticos muestrales para ambas series estadísticas son:

	MÉTODO A	**MÉTODO B**
n	10	7
$\sum_{i=1}^{n} x_i$	2457 ºC	1719 ºC
$\sum_{i=1}^{n} x_i^2$	621669 (ºC)^2	432947 (ºC)^2
$\overline{x}$	245.7 ºC	245.57 ºC
$\hat{s}$	44.70 ºC	42.45ºC

(**A**) ¿Puede aceptarse que el método A es menos preciso que el método B con un nivel del significación $\alpha =$ 1%? Razona la respuesta que des indicando las condiciones que son necesarias tener en cuenta para poder plantear el contraste.

Se trata de un contraste de hipótesis de desviación típica (dos poblaciones porque hay dos m.a.s.). Suponiendo que las dos poblaciones se distribuyen normalmente el modelo de probabilidad que interviene es la distribución de probabilidad F de Fisher-Snedecor. Por otra parte, se trata de un contraste unilateral de cola superior (derecha), siendo las hipótesis correspondientes:

$$\begin{cases} H_0: & \sigma_{A,0} = \sigma_{B,0} \\ H_a: & \sigma_A > \sigma_B \end{cases}$$

El estadístico del contraste es:

$$F = \frac{\chi_A^2}{\chi_B^2} = \frac{s_A^2}{s_B^2}\frac{\sigma_B^2}{\sigma_A^2} = \frac{s_A^2}{s_B^2} = \frac{44.70^2}{42.45^2} = 1.1091$$

El valor frontera región de aceptación/región crítica viene dado por:

$$F_1 = F_{\alpha, \nu_A, \nu_B} = F_{99\,\%, \nu_A = 9\,gdl, \nu_B = 6\,gdl} = 7.97612$$

Se deduce, que a partir de las muestras del enunciado no existe evidencia estadística que el método A sea menos preciso que el método B con un nivel de significación $\alpha =$ 1 %.

(**B**) Suponiendo que las temperaturas se distribuyen normalmente y que las varianzas poblacionales son iguales, razonar si se puede aceptar que los dos métodos de medición de la temperatura tienen la misma media con un nivel de significación $\alpha =$ 5 %. ¿Y si el nivel de significación fuera $\alpha =$ 1 %?

Justifica las respuestas.

Dado que se trata de un contraste de medias de dos poblaciones de muestras pequeñas (se aplicará el modelo t de Student), y que es un contraste bilateral (de dos colas):

$$\begin{cases} H_0: & \mu_{A,0} = \mu_{B,0} \\ H_a: & \mu_A \neq \mu_B \end{cases}$$

El estadístico del contraste es:

$$t = \frac{\left(\overline{x}_A - \overline{x}_B\right) - \left(\mu_{A,0} - \mu_{B,0}\right)}{\sqrt{\frac{\left(n_A - 1\right)\hat{s}_A^2 + \left(n_B - 1\right)\hat{s}_B^2}{n_A + n_B - 2}}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{\left(n_A - 1\right)\hat{s}_A^2 + \left(n_B - 1\right)\hat{s}_B^2}{n_A + n_B - 2}}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} =$$

$$= \frac{0.128571429}{\sqrt{1919.587619}\sqrt{0.242857143}} = \frac{0.128571429}{\sqrt{1919.587619}\sqrt{0.242857143}} = \frac{0.128571429}{43.8131 \times 0.492805} = 0.005954$$

El valor frontera región de aceptación/región crítica viene dado por:

$$t_{\alpha,\nu} = t_{\alpha,\,\nu = n_A + n_B - 2} = t_{5\,\%,\,\nu = n_A + n_B - 2 = 15\,gdl} \equiv t_{97.5\,\%,\,\nu = 15\,gdl} = 2.13145$$

Se deduce, que a partir de las muestras del enunciado no existe evidencia estadística de que la media del método sea diferente de la del método B con un nivel de significación $\alpha =$ 5 %. Si se reduce la región crítica, se seguirá satisfaciendo con mayor seguridad la hipótesis nula.

EJERCICIO 5

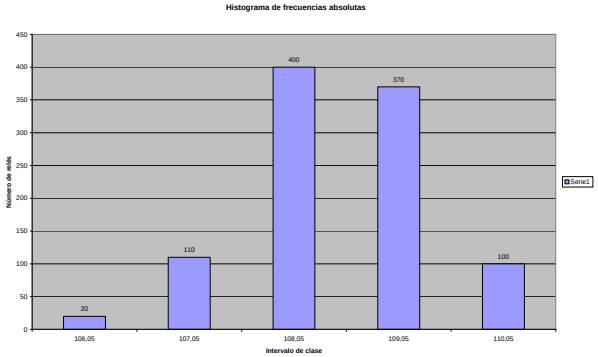
Se ha medido el punto de fusión (en ºC) de 1000 muestras de un determinado tipo de relés térmicos

Clase			
Límite inferior (º C)	Límite superior (º C)	**Marca de clase** (º C)	**Número de relés**
105.55	106.55	106.05	20
106.55	107.55	107.05	110
107.55	108.55	108.05	400
108.55	109.55	109.05	370
109.55	110.55	110.05	100

(A) Plantea un contraste de hipótesis adecuado que establezca si hay evidencia para rechazar que los datos de la muestra tomada siguen una

distribución normal. Resuélvelo con los niveles de significación $\alpha = 0.05$ y $\alpha = 0.01$.

Una representación gráfica de dicha serie estadística es



Y la tabla de cálculos intermedios

Clase					
Extremo inferior	**Extremo superior**	**Marca de clase**	**Número de relés**	MC NR	MC^2 NR
105,55	106,55	106,05	20	2121,0000	224932,0500
106,55	107,55	107,05	110	11775,5000	1260567,2750
107,55	108,55	108,05	400	43220,0000	4669921,0000
108,55	109,55	109,05	370	40348,5000	4400003,9250
109,55	110,55	110,05	100	11005,0000	1211100,2500
			1000	108470,0000	11766524,5000

n	1000
Media aritmética	108,4700
Varianza	0,7836
Estimación de la varianza	0,7844
Desviación estándar	0,8852
Estimación desv. estánd.	0,8857

INTERVALO DE CLASE (X)		INTERVALO DE CLASE (Z)		CÁLCULO DE PROBABILIDADES				
Extremo inferior	**Extremo superior**	**Extremo inferior**	**Extremo superior**	**Extremo inferior**	**Extremo superior**	**Frecuencia observada**	**Frecuencia teórica**	Contribución a chi cuadrada
105,55	106,55	-3,2986	-2,1690	0,0000	0,0150	20	15,0424	1,6339
106,55	107,55	-2,1690	-1,0393	0,0150	0,1493	110	134,2904	4,3936
107,55	108,55	-1,0393	0,0904	0,1493	0,5360	400	386,6721	0,4594
108,55	109,55	0,0904	1,2200	0,5360	0,8888	370	352,7715	0,8414
109,55	110,55	1,2200	2,3497	0,8888	1,0000	100	111,2235	1,1326
						1000	1000,0000	8,4609

para calcular el estadístico del contraste

$$\chi^2 = \sum_{i=1}^{NC=5} \frac{\left(n_i^t - n_i^o\right)^2}{n_i^t} = 8.4609$$

porque se trata de un problema de bondad de ajuste, contraste unilateral de cola superior siendo las hipótesis de trabajo

$$\begin{cases} H_0: & \mathrm{N}\left(\mu = 108.47^{o}C, \sigma = 0.8852^{o}C\right) \\ H_a: & \cancel{\mathrm{N}\left(\mu = 108.47^{o}C, \sigma = 0.8852^{o}C\right)} \end{cases}$$

En este problema de ajuste se observa que el número de grados de libertad será:

$$\nu = n - 1 - c = 5 - 1 - 2 = 2$$

siendo c el número de parámetros que ha sido necesario calcular (dos en este caso, la media aritmética y la varianza).

Por otra parte, de las tablas de la distribución chi cuadrada se obtiene que

$$\chi^2_{95\%,\,2} = 5.99$$
$$\chi^2_{99\%,\,2} = 9.21$$

de donde se deduce por comparación con el valor calculado para la prueba chi cuadrada (es decir, $\chi^2 = 8.4610 > \chi^2_{95\%,\,3} = 5.99$) que para $\alpha = 5\,\%$ no hay evidencia para aceptar la hipótesis nula (es decir, los datos de la muestra tomada no siguen una distribución normal). Sin embargo, para $\alpha = 1\,\%$ sí existe tal evidencia ya que $\chi^2 = 8.4610 < \chi^2_{99\%,\,3} = 9.21$; si bien es una situación muy exigente (es decir, un 99 % de confianza).

(B) Calcula el número de observaciones teóricamente esperadas entre [106.33 ºC, 108.01 ºC] suponiendo que la distribución es normal.

Recordando $\bar{x} = 108.47$ y $s_x = 0.8852$. Por otro lado la puntuación tipificada se calcula a partir de

$$Z = \frac{X - \bar{x}}{s_x}$$

se deduce que

$$X = 106.33 \Rightarrow Z = -2.4157$$
$$X = 108.01 \Rightarrow Z = -0.5196$$

o términos de probabilidades asociadas

$$\text{Ï}\,(106.33 \le X \le 108.01) = \text{Ï}\,(-2.4157 \le Z \le -0.5196) = 0.3017 - 0.0078 = 0.29384013$$

que es la probabilidad pedida. En consecuencia el número de observaciones

teóricamente esperadas es

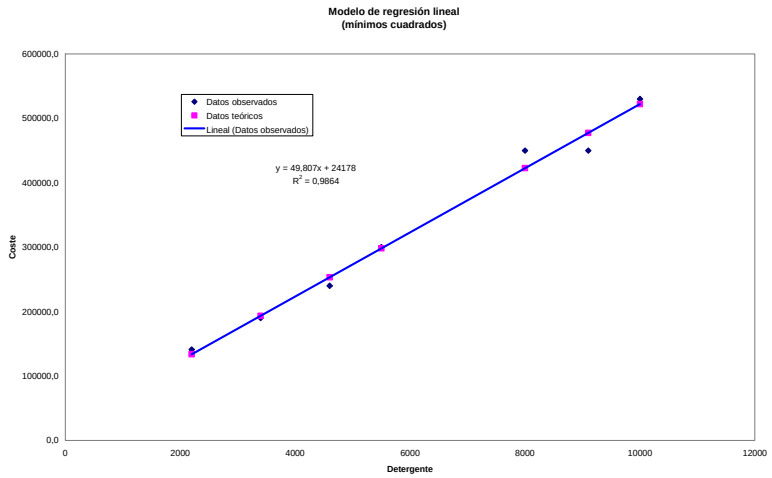
$$N_{solicitado} = 0.29384013 \times 10^3 = 294 \ \ reles$$

EJERCICIO 6

Una empresa química conoce que el coste de fabricar un determinado detergente responde a una ley lineal del tipo $y = a + b.x$, donde y = Coste en €, x = Detergente fabricado en m^3, a = Coste fijo en €, b = Coste adicional de la producción en €/m^3. Los datos recogidos en 7 plantas de la empresa en el último ejercicio son los siguientes:

Planta	1	2	3	4	5	6	7
x (m^3)	2200	3400	4600	5500	8000	9100	10000
y (€)	141000	190000	240000	300000	450000	450000	530000

El diagrama de dispersión es



Y la tabla de frecuencias

		Detergente	Coste					
	n	X	$Y_{observada}$	X_i^2	X_iY_i	Y_i^2	$Y_{teórica}$	$Residuo^2$
	1	2200	141000,0	4840000	310200000	19881000000	133754,3320	52499705,0009
	2	3400	190000,0	11560000	646000000	36100000000	193523,0769	12412071,0059
	3	4600	240000,0	21160000	1104000000	57600000000	253291,8219	176672528,4204
	4	5500	300000,0	30250000	1650000000	90000000000	298118,3806	3540491,6914
	5	8000	450000,0	64000000	3600000000	202500000000,000	422636,5992	748755703,8732
	6	9100	450000,0	82810000	4095000000	202500000000,000	477424,6154	752109528,9941
	7	10000	530000,0	100000000	5300000000	280900000000,000	522251,1741	60044302,9979
Sumas	7	42800	2301000	314620000	16705200000	889481000000,000		1806034332

β_1	49,8072874
β_0	24178,2996
σ^2	361206866
σ	19005,4431

SS_{XX}	52928571,43
SS_{YY}	133109428571,43
SS_{XY}	2636228571,43
SS_E	1806034331,98
SS_R	131303394239,45

$\sigma^2(\beta_0)$	306728486,7
$\sigma^2(\beta_1)$	6,824421227
r	0,99319281
r^2	0,986431958

(A) Obtener, a partir de estos datos, la recta de regresión.

$$Y = 49.80728745\mathrm{X}+24178.2996$$

(B) De acuerdo con el resultado obtenido en (A) decir cuál es la estimación del coste fijo en €, así como el coste adicional en €/m^3, de la producción.

Se sigue que el coste fijo es 24178.2996 € y el coste adicional de la producción se eleva a 49.81 €/m^3.

(C) Calcular el coeficiente de correlación de Pearson entre las variables x e y.

$$r = \frac{s_{XY}}{s_X s_Y} = 0.99319281$$

(D) Estimar cuál será el gasto en € de una planta que pretenda producir 6000 m^3 de detergente.

$$\hat{y}_i^t = 49.80728745\,\mathrm{X}\big|_{\mathrm{x_i}=6000} +24178.2996=323022.0\ €$$

Para un nivel de confianza α = 99 % el intervalo de confianza de una interpolación viene dado por

$$[\mathrm{I}, \mathrm{L}] = \hat{y}_i^t \pm t_1 \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_i-\overline{x})^2}{SS_{XX}}}$$

siendo $t_1 = t_{\alpha,\, v=n-2} = t_{99\,\%,\, v=5\ gdl} \equiv t_{99.5\,\%,\, v=5\ gdl} = \pm 4.773340605$. Entonces, se deduce que el intervalo buscado es: $[\mathrm{I}, \mathrm{L}] = [226028.38\ €, 420015.67\ €]$.