Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Primer Semestre 2010

Curso	**:**	**Probabilidad y Estadística**
Sigla	**:**	**EYP1113**
Pauta	**:**	**Examen**
Profesores	**:**	**Ricardo Aravena (Sec 01 y 03) y Ricardo Olea (Sec 02 y 04)**
Ayudantes	**:**	**Tamara Fernández, Claudia Ortega, Constanza Quezada, Ignacia Vicuña.**

1. **[10 %]** Dos ríos, a y b, fluyen a través de un recinto en el cual se encuentra una fábrica de papel y las autoridades le han permitido a esta empresa depositar sus residuos en ambos ríos. El nivel de oxígeno disuelto en el agua, DO, río abajo es una indicador del grado de contaminación que presentan los ríos por causa de la fábrica.

   Sean los siguientes eventos:

   A: Río a tienen un nivel inaceptable de contaminación.

   B: Río b tienen un nivel inaceptable de contaminación.

   Considerando los registros del último año sobre los niveles de DO en los dos ríos analizados, se tiene que el 20 % y 30 % de los días los ríos a y b presentaron una contaminación inaceptable, respectivamente. Por otra parte, sólo el 10 % de los días del último año ambos ríos presentaron a la vez una contaminación inaceptable. A partir de esta información, calcule la probabilidad que sólo uno de los dos ríos tenga un nivel de contaminación inaceptable en un día cualquiera.

   Solución

   Del enunciado tenemos que

$$P(A) = 0{,}2; \quad P(B) = 0{,}3; \quad P(A \cap B) = 0{,}1$$

   Se pide

$$
\begin{aligned}
[\mathbf{3}\,\%] \quad P([A \cap \overline{B}] \cup [\overline{A} \cap B]) &= P(A \cap \overline{B}) + P(\overline{A} \cap B), \quad \text{por ser eventos disjuntos} \\
&= P(\overline{B} \mid A) \cdot P(A) + P(\overline{A} \mid B) \cdot P(B), \quad \text{por regla multiplicativa} \quad [\mathbf{2}\,\%] \\
&= [1 - P(B \mid A)] \cdot P(A) + [1 - P(A \mid B)] \cdot P(B), \quad \text{por ley del complemento} \\
&= \left[1 - \frac{P(B \cap A)}{P(A)}\right] \cdot P(A) + \left[1 - \frac{P(A \cap B)}{P(B)}\right] \cdot P(B), \quad \text{por definición de prob. cond.} \quad [\mathbf{2}\,\%] \\
&= \left[1 - \frac{0{,}1}{0{,}2}\right] \cdot 0{,}2 + \left[1 - \frac{0{,}1}{0{,}3}\right] \cdot 0{,}3 \quad [\mathbf{2}\,\%] \\
&= 0{,}3 \quad [\mathbf{1}\,\%]
\end{aligned}
$$

2. **[10 %]** En la fabricación de vigas y columnas de acero, hay variabilidad inevitables en las dimensiones (por ejemplo, la longitud). Supongamos que en una construcción, se requiere que las vigas y columnas tengan la dimensión especificada dentro de un rango de tolerancia de ± 5 mm con una confiabilidad del 99,74 %. Si la variabilidad presente en la fabricación de las vigas y columnas se comporta como una variable aleatoria con distribución Normal de media cero y desviación estándar σ, ¿cuál es la precisión requerida, σ, para el proceso de producción?

Solución

Sea E la tolerancia con respecto a la dimensión especifica en la fabricación.

Se tiene que

$$E \sim \text{Normal}(0, \sigma)$$

Se pide que

$$
\begin{aligned}
[\mathbf{3}\,\%] \quad P(-5 \leq E \leq 5) &= \Phi\left(\frac{5-0}{\sigma}\right) - \Phi\left(\frac{-5-0}{\sigma}\right) \quad [\mathbf{2}\,\%] \\
&= 2 \cdot \Phi\left(\frac{5}{\sigma}\right) - 1 \quad [\mathbf{2}\,\%] \\
&= 0{,}9974
\end{aligned}
$$

Es decir,

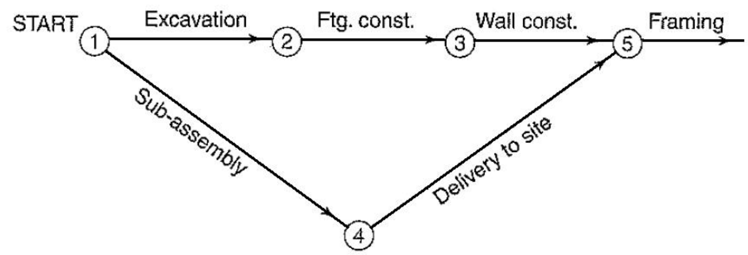
$$\frac{5}{\sigma} = \Phi^{-1}\left(\frac{0{,}9974 + 1}{2}\right) = \Phi^{-1}\left(0{,}9987\right) \approx 3{,}01 \quad [\mathbf{2}\,\%]$$

Por lo tanto, la precisión requerida es

$$\sigma = 1{,}661130 \quad [\mathbf{2}\,\%]$$

3. **[20 %]** La edificación de una casa prefabricada consiste en el ensamblando de distintas componentes que se construyen en una planta y que se ensamblan en el lugar escogido para la edificación. Simultáneamente, mientras se fabrican los componentes, el sitio escogido se prepara para la edificación con la excavación de los cimientos y otras actividades varias.

   Las diferentes actividades y la secuencia de ellas es la siguiente:

   La duración de cada trabajo necesario para las respectivas actividades se resumen en la siguiente tabla:

		Completion Time (days)	
Activity	Description	Mean	Std. Dev.
1-2	Excavation of foundation	2	1
2-3	Construction of footings	1	1/2
3-5	Construction of foundation walls	3	1
1-4	Fabrication and assembly of components	5	1
4-5	Delivery of assembled components to site	2	1/2

   Asumiendo que las duraciones de cada una de las actividades son estadísticamente independientes y se comportan como variables aleatorias con distribución Normal. Calcule la probabilidad que el armado (Framing) de la casa comience en menos de 8 días.

   Solución

   Sean X_1, X_2, X_3, X_4 y X_5 los tiempos de cada una de las actividades descritas los cuales son estadísticamente independientes entre sí.

   El evento de interés es
$$F = \{T_1 \leq 8\} \cap \{T_2 \leq 8\}$$
   donde
$$T_1 = X_1 + X_2 + X_3 \sim \text{Normal}\left(\mu = 2 + 1 + 3,\, \sigma = \sqrt{1^2 + 0{,}5^2 + 1^2}\right), \quad [\mathbf{1}\,\%]$$
$$T_2 = X_4 + X_5 \sim \text{Normal}\left(\mu = 5 + 2,\, \sigma = \sqrt{1^2 + 0{,}5^2}\right) \quad [\mathbf{1}\,\%]$$
   los cuales por construcción son variables aleatorias estadísticamente independientes.
   La probabilidad solicitada es
$$\begin{aligned}
P(F) &= P(\{T_1 \leq 8\} \cap \{T_2 \leq 8\}) \\
&= P(T_1 \leq 8) \cdot P(T_2 \leq 8), \quad \text{por independencia} \quad [\mathbf{5}\,\%] \\
&= \Phi\left(\frac{8-6}{1{,}5}\right) \cdot \Phi\left(\frac{8-7}{1{,}12}\right) \quad [\mathbf{4}\,\%] \\
&\approx \Phi(1{,}33) \cdot \Phi(0{,}89) \quad [\mathbf{4}\,\%] \\
&= 0{,}9082 \cdot 0{,}8133 \quad [\mathbf{2}\,\%] \\
&= 0{,}738639 \quad [\mathbf{3}\,\%]
\end{aligned}$$

4. **[20 %]** Las planchas de fibro-cemento (lisas u onduladas) deben cumplir rigurosos estándares en relación al contenido de asbesto. Como parte de un proceso de certificación, usted toma una muestra de 23 trozos de 3 mm de diámetro de planchas en el mismo proceso de producción, y evalúa el número de fibras de asbesto por trozo. Los resultados muestra un promedio de 24,9 con una desviación estándar de 5,5 fibras por trozo. Asumiendo que el número de fibras de asbesto por trozo sigue una distribución Normal.

   (a) ¿Existe evidencia estadística ($\alpha = 5\,\%$) que permita afirmar que el nivel medio del número de fibras asbesto supera la norma fijada en 23?.

   (b) Obtenga un Intervalo de Confianza al 95 % para el nivel medio del número de fibras de asbesto.

   Solución

   (a) Tenemos las siguientes hipótesis

   $$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs.} \quad \mathrm{H}_a : \mu > \mu_0 \quad [\mathbf{3}\,\%]$$

   con $\mu_0 = 23$.

   Bajo H_0 el estadístico de prueba

   $$T_n = \frac{\overline{X}_n - \mu_0}{S_X/\sqrt{n}} \sim \text{t-student}(n-1)$$

   Alternativa 1

   Se rechaza H_0 si $T_n > t_{p=1-\alpha}(n-1)$.

   Reemplazando tenemos que

   $$[\mathbf{3}\,\%] \quad T_n = \frac{24{,}9 - 23}{5{,}5/\sqrt{23}} = 1{,}657 \not> 1{,}717 = t_{0{,}95}(22) \quad [\mathbf{3}\,\%]$$

   Es decir, no existe suficiente evidencia para afirmar que el nivel medio del número de fibras de asbesto supera la norma fijada. **[3 %]**

   Alternativa 2

   Se define

   $$valor - p = P(T > T_n), \quad \text{con } T \sim \text{t-student}(n-1).$$

   Se rechaza la hipótesis nula si valor-p $< \alpha$.

   Reemplazando

   $$T_n = \frac{24{,}9 - 23}{5{,}5/\sqrt{23}} = 1{,}657 \quad [\mathbf{3}\,\%]$$

   y

   $$\text{valor-p} = P(T > 1{,}657) \Rightarrow 5\,\% < \text{ valor-p} < 10\,\% \quad [\mathbf{3}\,\%]$$

   Es decir, no existe suficiente evidencia para afirmar que el nivel medio del número de fibras de asbesto supera la norma fijada. **[3 %]**

(b) Un intervalo de confianza para μ esta dado por

$$\overline{X}_n \pm \frac{S_X}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) \quad [\mathbf{3}\,\%]$$

$$24{,}9 \pm \frac{5{,}5}{\sqrt{23}} \cdot 2{,}074, \quad \text{para } \alpha = 0{,}05 \text{ y } n = 23 \quad [\mathbf{3}\,\%]$$

$$24{,}9 \pm 2{,}3785$$

Por lo tanto

$$< \mu >_{0{,}95} = (22{,}5;\ 27{,}3) \quad [\mathbf{2}\,\%]$$

5. **[20 %]** Para una muestra de 320 observaciones de tensiones residuales de vigas de acero de ala ancha se desea ajustar una distribución Normal y Log-Normal. A partir de los datos se estiman los valores de $\mu = 0{,}3561$ y $\sigma = 0{,}1927$.

Interval Res. Str./Yield Str.	$< 0{,}111$	$0{,}111 - 0{,}222$	$0{,}222 - 0{,}333$	$0{,}333 - 0{,}444$	$0{,}444 - 0{,}555$
Observed Frequency	15	72	88	54	38

Interval Res. Str./Yield Str.	$0{,}555 - 0{,}666$	$0{,}666 - 0{,}777$	$0{,}777 - 0{,}888$	$> 0{,}888$
Observed Frequency	31	16	3	3

¿Cual de las dos distribuciones propuesta es más apropiada? Sea explícito y fundamente su respuesta.

Solución

Test de Bondad de Ajuste χ^2 para una distribución Normal y Log-Normal. A partir de la muestra de tamaño $n = 302$ se tiene los siguientes momentos muestrales:

$$\overline{X}_n = 0{,}3561 \quad \text{y} \quad S_X = 0{,}1927$$

Sean las siguientes hipótesis:

$$\text{H}_0 : \text{Datos} \sim \text{Normal}(\mu,\, \sigma) \text{ vs. } \text{H}_a : \text{Datos} \not\sim \text{Normal}(\mu,\, \sigma) \tag{1}$$

$$\text{H}_0 : \text{Datos} \sim \text{LogNormal}(\lambda,\, \zeta) \text{ vs. } \text{H}_a : \text{Datos} \not\sim \text{LogNormal}(\lambda,\, \zeta) \tag{2}$$

Para (1) tenemos que:

$$\mu = \overline{X}_n = 0{,}3561 \quad \text{y} \quad \sigma = S_X = 0{,}1927 \quad [\mathbf{2}\,\%]$$

Para (2) tenemos que:

$$\zeta^2 = \log\left((1 + \left[\frac{S_X}{\overline{X}_n}\right]^2\right) \Rightarrow \zeta = \sqrt{\log(1 + 0{,}1927^2/0{,}3561^2)} = 0{,}5067896 \quad [\mathbf{1}\,\%]$$

$$\lambda = \log \overline{X}_n - \frac{1}{2}\zeta^2 \Rightarrow \lambda = \log 0{,}3561 - \frac{1}{2} \cdot 0{,}5067896^2 = -1{,}160962 \quad [\mathbf{1}\,\%]$$

Para el calculo de las probabilidades teóricas utilizamos que

$$\text{Si } X \sim \text{Normal}(\mu,\, \sigma) \Rightarrow P(X \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

$$\text{Si } X \sim \text{Log-Normal}(\lambda,\, \zeta) \Rightarrow P(X \leq x) = \Phi\left(\frac{\ln x - \lambda}{\zeta}\right)$$

Interval	Observed	Theoretical Frequencies		$\sum (o_i - e_i)^2/e_i$	
Res. Str/Yield Str.	Frequency	Normal	Log-Normal	Normal	Log-Normal
$< 0{,}111$	15	32,54	6,51	9,46	11,07
0,111 - 0,222	72	45,29	73,03	15,75	0,01
0,222 - 0,333	88	66,89	95,88	6,66	0,65
0,333 - 0,444	54	71,54	66,02	4,30	2,19
0,444 - 0,555	38	55,41	37,14	5,47	0,02
0,555 - 0,666	31	31,07	19,58	0,00	6,67
0,666 - 0,777	16	12,62	10,17	0,91	3,34
0,777 - 0,888	3	3,71	5,32	0,13	1,01
$> 0{,}888$	3	0,92	6,36	4,66	1,77
$\sum$	320	320,00	320,00	47,34	26,74
		[2.5 %]	**[2.5 %]**	**[2.5 %]**	**[2.5 %]**

En ambos casos al comparar con una distribución $\chi^2_{(9-1)-2} = \chi^2_6$ **[2 %]** se tiene que el valor-p es inferior a 0.5 % **[2 %]**, esto implica que ambas distribuciones propuestas no serían adecuadas. Entre las dos, el mejor ajuste se logra asumiendo una distribución Log-Normal. **[2 %]**

6. **[20 %]** Con el fin de predecir la concentración media de oxígeno disuelto (DO) en función de la temperatura promedio (T) de una piscina, se realizan 11 mediciones. Se postula que un modelo exponencial puede ser apropiado para representar la relación; es decir:

$$DO = \alpha\, e^{\beta\, T}$$

que al tomar logaritmo natural a ambos lados, se tiene:

$$\ln DO = \ln \alpha + \beta\, T$$

Un resumen de los datos es el siguiente:

$$\sum_{i=1}^{n} x_i = 279{,}7; \quad \sum_{i=1}^{n} x_i^2 = 7149{,}0; \quad \sum_{i=1}^{n} \ln y_i = 11{,}68; \quad \sum_{i=1}^{n} (\ln y_i)^2 = 13{,}15;$$

$$\sum_{i=1}^{n} x_i\, \ln y_i = 292{,}62; \quad \sum_{i=1}^{n} (\ln y_i - \ln {y_i}^{'})^2 = 0{,}2271$$

con $X = T$ e $Y = DO$.

(a) Estime los coeficientes del modelo propuesto y entregue los estadísticos asociados (utilice 3 decimales).

(b) Obtenga el intervalo de confianza al 95 % para el $\ln DO$ cuando la temperatura alcanza los 25° y los 28°. ¿Cuál es más preciso? ¿Por qué? (justifique en una línea)

Solución

(a) Para el modelo

$$\ln DO = \gamma + \beta\, T$$

los estimadores de mínimos cuadrados son:

$$\begin{aligned}
\hat{\beta} &= \frac{\displaystyle\sum_{i=1}^{n} x_i \cdot \ln y_i - n \cdot \overline{x} \cdot \overline{\ln y}}{\displaystyle\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \\
&= \frac{296{,}62 - 11 \cdot 25{,}427 \cdot 1{,}062}{7149 - 11 \cdot 25{,}427^2} \\
&= -0{,}119 \quad [\mathbf{2}\,\%] \\
\hat{\gamma} &= \overline{\ln y} - \hat{\beta} \cdot \overline{x} \\
&= 1{,}062 - 0{,}119 \cdot 25{,}427 \\
&= 4{,}088 \to \hat{\alpha} = 1{,}408 \quad [\mathbf{2}\,\%]
\end{aligned}$$

con $X = T$, $Y = DO$.

Los estadísticos asociados son:

$$
\begin{aligned}
S^2_{\ln Y} &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\ln y_i)^2 - n\cdot\overline{\ln y}^2\right]\\
&= \frac{1}{11-1}\left[13{,}15 - 11\cdot 1{,}062^2\right]\\
&= 0{,}074 \quad [\mathbf{1}\,\%]\\
S^2_{\ln Y\,|\,X} &= \frac{1}{n-2}\left[\sum_{i=1}^{n}(\ln y_i - \ln y_i')^2\right]\\
&= \frac{0{,}2271}{11-2}\\
&= 0{,}025 \quad [\mathbf{1}\,\%]\\
r^2 &= 1 - \frac{S^2_{Y\,|\,X}}{S^2_Y}\\
&= 1 - \frac{0{,}025}{0{,}074}\\
&= 0{,}662 \quad [\mathbf{2}\,\%]
\end{aligned}
$$

El modelo propuesto queda como

$$DO' = 1{,}408\cdot e^{-0{,}119\,T} \quad [\mathbf{2}\,\%]$$

(b) Tenemos que un intervalo de confianza al $(1-\alpha)\times 100\,\%$ para el valor esperado del $\ln DO$ condicionada a una temperatura T esta dado por

$$< \mu_{\ln y_i} >_{1-\alpha} = \overline{\ln y_i} + t_{(1-\alpha/2),\,n-2}\times S_{\ln Y\,|\,X}\times\sqrt{\frac{1}{n}+\frac{(x_i-\overline{x})^2}{\displaystyle\sum_{j=1}^{n}(x_j-\overline{x})^2}}$$

con $X = T$, $Y = DO$ y $\overline{\ln y} = \hat{\gamma} + \hat{\beta}\cdot x$.

Para un 95 % de confianza se tiene que

$$
\begin{aligned}
T = 25^\circ \qquad < \mu_{\ln DO} >_{95\,\%} &= 1{,}110 \pm 2{,}262\times 0{,}159\times\sqrt{\frac{1}{11}+\frac{(25{,}0-25{,}43)^2}{(7149-11\times 25{,}43^2)}}\\
&= (1{,}00;\ 1{,}22) \quad [\mathbf{3}\,\%]\\
T = 28^\circ \qquad < \mu_{\ln DO} >_{95\,\%} &= 0{,}780 \pm 2{,}262\times 0{,}159\times\sqrt{\frac{1}{11}+\frac{(28{,}0-25{,}43)^2}{(7149-11\times 25{,}43^2)}}\\
&= (0{,}59;\ 0{,}97) \quad [\mathbf{3}\,\%]
\end{aligned}
$$

Para $T = 25$ se obtiene un intervalo más pequeño por que es un valor más cercano al promedio de las temperaturas. [**2 %**]