Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Segundo Semestre 2023

Curso : Probabilidad y Estadística
Sigla : EYP1113
Profesores : Ricardo Aravena C., Ricardo Olea O., Felipe Ossa M. y Mauricio Toro C.

PAUTA INTERROGACIÓN 4

Pregunta 1

Una compañía frutícola vende naranjas cosechadas en bolsas de 6 unidades cada una. El coeficiente de acidez de cada naranja se modela como una variable aleatoria Beta$(1, 2)$ (con soporte $(a,\, b) = (0,\, 1)$), donde un valor cercano a 0 indica poca acidez y un valor cercano a 1 indica acidez muy alta. Por medio de un escáner se puede medir el coeficiente de acidez en cada naranja dentro de las bolsas preparadas, y una bolsa se aceptará y enviará a venta sólo si la mayor acidez de sus naranjas encontrada está bajo 0.8.

Asumiendo independencia de la acidez entre naranjas cosechadas responda lo siguiente:

(a) **[3.0 Ptos.]** Calcule la probabilidad de que una bolsa de seis unidades de naranjas sea aceptada y por ende enviada a la venta

(b) **[3.0 Ptos.]** Si mil bolsas de 6 unidades fueron preparadas para la venta, calcule la probabilidad aproximada que más de 790 de ellas sean aceptadas y enviadas a venta finalmente.

Solución

(a) Sea X el coeficiente de acidez de una naranja y definamos como p a la probabilidad que una bolsa sea aceptada.

$$\begin{aligned} p &= P(\text{máx}\{X_1, \ldots, X_6\} < 0.8) = P(X_1 < 0.8,\, \ldots, X_6 < 0.8) \\ &\overset{\text{iid}}{=} [F_X(0.8)]^6 \\ &= \left[\int_0^{0.8} \frac{1}{B(1,2)} \cdot x^{1-1}\,(1-x)^{2-1}\,dx\right]^6 \\ &= \left[2\,x - x^2\Big|_0^{0.8}\right]^6 \\ &= 0.7827578 \qquad\qquad \approx 0.7828 \end{aligned}$$

(b) Sea Y el número de bolsas de 6 unidades aceptadas entre mil.

Por Teorema Central del Limite:

$$Y \sim \text{Binomial}(n = 1000,\, p = 0.7828) \overset{\text{aprox}}{\sim} \text{Normal}\left(n\,p,\, \sqrt{n\,p\,(1-p)}\right)$$

Se pide $P(Y > 790)$.

Aplicando corrección por continuidad:

$$P(Y > 790) = 1 - P(Y \leq 790) \approx 1 - \Phi\left(\frac{790.5 - n\,p}{\sqrt{n\,p\,(1-p)}}\right) = 1 - \Phi(0.59) = 0.2776$$

Nota: Como referencia el valor exacto de la probabilidad según modelo Binomial es 0.2779.

Asignación de Puntaje:

Logro 1: Asignar **[1.0 Ptos]** por reconocer que $p = P(\text{máx}\{X_1, \ldots, X_6\} < 0.8)$.

Logro 2: Asignar **[1.0 Ptos]** por indicar que $p = [F_X(0.8)]^6$ por iid.

Logro 3: Asignar **[1.0 Ptos]** por resolver la integral e indicar que $p = 0.7827578$.

Logro 4: Asignar **[1.0 Ptos]** por aplicar aproximación Normal $\left(n\,p,\ \sqrt{n\,p\,(1-p)}\right)$.

Logro 5: Asignar **[1.0 Ptos]** por aplicar corrección por continuidad al estandarizar. Si no realiza corrección al estandarizar asignar **[0.5 Ptos]**.

Logro 6: Asignar **[1.0 Ptos]** por responder que $P(Y > 790) = 0.2776$. Si no realiza la corrección por continuidad, la respuesta sería $1 - \Phi(0.55) = 0.2912$, no descontar puntaje.

+ 1 Punto Base

Pregunta 2

El crecimiento constante y sólido de la industria de cerezas en Chile, impulsado por la creciente demanda en China, se encuentra amenazado por las inclemencias climáticas que afectaron particularmente las regiones de O'Higgins y el Bío Bío. Uno de los principales efectos de las lluvias son grietas en la piel de la cereza, fenómeno conocido en la industria como cracking.

Exhaustivos controles microbiológicos son llevados a cabo antes de exportar una partida (embarque). Para decidir, se toma una muestra de cerezas desde un pallet (caja) seleccionado al azar (entre miles), y si más del 15 % de las cerezas presenta cracking, la partida completa es rechazada. Además, para la muestra seleccionada se evalúa la concentración de azúcar, Brix, el cuál se expresa en grados y representa el porcentaje de sólidos solubles en el jugo de la fruta, principalmente azúcares como la glucosa y la fructosa. Asuma que el Brix presenta un comportamiento Gaussiano (Normal).

Un muestra aleatoria de 128 cerezas proveniente de un pallet, presentó los siguientes resultados:

		Cracking	
		SI	NO	TOTAL
	N	25	103	128
Brix	mean	18.2	19.5	19.2
	sd	3.4	4.5	4.0

(a) **[2.0 Ptos.]** ¿Existe suficiente evidencia para rechazar la partida completa? Considere un nivel de significancia del 5 %.

(b) **[1.0 Ptos.]** ¿Existe evidencia que permita afirmar que la concentración de azúcar de las cerezas con cracking es inferior a 20°? Considere un nivel de significancia del 1 %.

(c) **[2.0 Ptos.]** ¿Es posible afirmar la concentración de azúcar difiere entre las cerezas sin y con cracking? Considere un nivel de significancia del 10 %.

(d) **[1.0 Ptos.]** ¿Cuál debería ser el tamaño de una muestra para determinar la proporción de cerezas con cracking con un error no mayor a cinco puntos porcentuales, es decir ± 0.05, con una confianza del 90 %? Para su cálculo, considere el caso en que la variable Bernoulli alcanza máxima varianza, es decir, $1/4$.

Solución

(a) Se pide contrastar las siguientes hipótesis

$$H_0 : p = p_0 \quad \text{vs} \quad H_a : p > p_0,$$

con $p_0 = 0.15$.

Bajo H_0 se tiene que

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} \overset{\text{aprox}}{\sim} \text{Normal}(0,1).$$

Reemplazando

$$n = 128, \quad \hat{p} = \frac{25}{128} \rightarrow Z_0 = 1.435714$$

Como

$$\text{valor-p} \approx 1 - \Phi(1.44) = 1 - 0.9251 = 0.0749 > 0.05 = \alpha$$

o

$$Z_0 = 1.435714 < 1.645 = k_{0.95},$$

entonces no rechazamos H_0 y por lo tanto no apoyamos la hipótesis de rechazar la partida.

(b) Se pide contrastar las siguientes hipótesis

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu < \mu_0,$$

con $\mu_0 = 20$.

Bajo H_0 se tiene que

$$T_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim \text{t-Student}(n-1).$$

Reemplazando

$$n = 25, \quad \overline{X} = 18.2, \quad S = 3.4 \rightarrow T_0 = -2.647059$$

Utilizando tabla de percentiles t-Student(24) se deduce que

$$\text{valor-p} < 0.01 = \alpha$$

y también que

$$T_0 = -2.647059 < -2.492 = t_{0.01}(24).$$

Por lo tanto, rechazamos H_0 y apoyamos que la concentración de azúcar de las cerezas con cracking es inferior a 20°.

(c) Se pide contrastar las siguientes hipótesis

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2.$$

Previamente realizaremos la siguiente prueba de hipótesis con respecto a las varianzas:

$$H_0 : \sigma_1 = \sigma_2 \quad \text{vs} \quad H_a : \sigma_1 \neq \sigma_2.$$

Bajo H_0 se tiene que

$$F_0 = \frac{S_2^2}{S_1^2} = 1.75173 \sim F(102, 24) \approx F(30, 24).$$

Como $F_0 = 1.75 < 1.94 = F_{0.95}(30, 24)$, entonces no se rechaza la igualdad de varianza, por lo que se procede a realizar una prueba de hipótesis para la comparación de medias con varianzas desconocidas e iguales.

Bajo H_0 se tiene que

$$T_0 = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{25} + \frac{1}{103}}} \sim \text{t-Student}(25 + 103 - 1) \approx \text{Normal}(0, 1).$$

Reemplazando

$$n = 25, \quad m = 103, \quad \overline{X} = 18.2, \quad \overline{Y} = 19.5 \quad \text{y} \quad S_p = \sqrt{\frac{24 \cdot 3.4^2 + 102 \cdot 4.5^2}{25 + 103 - 2}} = 4.312164,$$

se tiene que

$$|T_0| = |-1.35| < 1.645 = k_{0.95} \quad \text{y} \quad \text{valor-p} = 2 \cdot [1 - \Phi(|-1.35|)] = 0.177 > 0.10 = \alpha$$

Por lo tanto, no rechazamos H_0 y no apoyamos la hipótesis la concentración de azúcar difiere entre las cerezas sin y con cracking.

(d) Como $1-\alpha = 0.90$, entonces el tamaño muestral por criterio de varianza máxima está dado por:

$$n = \left(\frac{k_{0.95}}{2 \cdot \text{e.e.}}\right)^2 = \left(\frac{1.645}{2 \cdot 0.05}\right)^2 = 270.6025 \rightarrow n = 271.$$

Asignación de Puntaje:

Logro 1: Asignar **[0.5 Ptos]** por $Ha : p > p_0$ y **[0.5 Ptos]** por $Z_0 = 1.435714$.

Logro 2: Asignar **[0.5 Ptos]** por por cálculo valor-p o determinación del valor críticoo y **[0.5 Ptos]** por conclusión.

Logro 3: Asignar **[0.3 Ptos]** por $T_0 = -2.647059$, **[0.4 Ptos]** por cálculo valor-p o determinación del valor crítico y **[0.3 Ptos]** por conclusión.

Logro 4: Asignar **[1.0 Ptos]** por concluir que las varianzas desconocidas pueden considerarse iguales.

Logro 5: Asignar **[1.0 Ptos]** Asignar **[0.5 Ptos]** por por cálculo valor-p o determinación del valor críticoo y **[0.5 Ptos]** por conclusión con respecto a la diferencia de medias.

Logro 6: Asignar **[1.0 Ptos]** por determinar que $n = 271$. No descontar si responde $n = 270$.

+ 1 Punto Base

Pregunta 3

Un especialista postula que el precio de los arriendos (en UF) en la Región Metropolitana depende de la superficie, cercanía al metro (1 = Si, 0 = No), número de dormitorios, número de baños, número de bodegas, número de estacionamientos y existencia de piscina (1 = Si, 0 = No), entre otros atributos. A partir de información extraída desde portales que ofrecen propiedades en arriendo, se obtuvo una muestra aleatoria para una comuna del sector oriente con la cual se ajustaron dos modelos de regresión lineal.

(a) **[3.0 Ptos.]** Complete la información que aparece con XXXX en la siguiente salida de R:

```
var(Data$Precio)
[1] 51.48817

lm(formula = Precio ~ Superficie, data = Data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.88839    1.84645  11.854  < 2e-16
Superficie   0.11354    0.01453    XXXX     XXXX

Residual standard error: 5.807 on 113 degrees of freedom
Multiple R-squared:  XXXX,Adjusted R-squared:  XXXX
F-statistic: XXXX on 1 and 113 DF,  p-value: 3.105e-12
```

¿Hay regresión entre precio y superficie?, ¿qué tan buena es esta regresión?.

Uno de los supuestos es que los residuos (error) del modelo se ajustan a una distribución Normal$(0, \sigma^2)$. Se aclaro durante la prueba que el σ^2 es la varianza del error del modelo.

(b) **[3.0 Ptos.]** Realice una prueba de bondad de ajuste χ^2 para testear si los residuos del modelo que considera entre sus variables la superficie, cercanía al metro y piscina, cumple con este supuesto en base a la siguiente información:

```
(-Inf,-4]    (-4,0]     (0,4]  (4, Inf]
       28        35        26        26

Residual standard error: 4.83
```

Solución

(a) Tenemos que

$$\begin{aligned}
\texttt{t value} &= \frac{0.11354}{0.01453} = 7.814178 \\
\texttt{F-statistic} &= 7.814178^2 = 61.06137 \\
\texttt{Pr(>|t|)} &= 3.105e-12 \\
\texttt{SCE} &= (\texttt{Residual standard error})^2 \cdot 113 = 3810.501 \\
\texttt{SCT} &= 51.48817 \cdot 114 = 5869.651 \\
\texttt{Multiple R-squared} &= 1 - \frac{3810.501}{5869.651} = 0.350813 \\
\texttt{Adjusted R-squared} &= 1 - \frac{5.807^2}{51.48817} = 0.345068
\end{aligned}$$

Como valor-p <0.001, entonces la regresión es significativa y explica el 35 % de la variabilidad presente en los datos.

Otra relación que se puede utilizar es la siguiente:

$$\texttt{Multiple R-squared} = \frac{\texttt{SCR}}{\texttt{SCT}} = \frac{\texttt{SCR}}{\texttt{SCR}+\texttt{SCE}} = \frac{\frac{\texttt{SCR}}{\texttt{SCE}}}{\frac{\texttt{SCR}}{\texttt{SCE}}+1} = \frac{\frac{\texttt{SCR}}{\texttt{SCE}}(n-2)}{\frac{\texttt{SCR}}{\texttt{SCE}}(n-2)+(n-2)} = \frac{\texttt{F-statistic}}{\texttt{F-statistic}+n-2}$$

(b) Tenemos que $\sigma = 4.83$ y a partir de

$$\Phi(-4/\sigma) = 1 - \Phi(+4/\sigma) \approx 1 - \Phi(0.83) = 0.2033$$

se completa la siguiente tabla

```
           O      p       E
--------------------------
(-Inf,-4] 28 0.2033 23.3795
(-4,0]    35 0.2967 34.1205
(0,4]     26 0.2967 34.1205
(4, Inf]  26 0.2033 23.3795
--------------------------
```

El estadístico de prueba está dado por

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 3.162178 \sim \chi^2(4-1-1)$$

y a partir de la tabla de percentiles $\chi^2(2)$ del formulario se deduce que

$$20\,\% < \text{valor-p} < 30\,\%.$$

Por lo tanto, el supuesto no es rechazado.

Nota: Si el alumno considera como σ el valor 4.83^2, entonces el valor del estadístico de prueba χ^2 es $X^2 = 159.2494$ y el supuesto es rechazado. No descontar puntaje. La tabla con los valores esperados en este caso sería la siguiente:

```
--------------------------
           O      p       E
--------------------------
(-Inf,-4] 28 0.4325 49.7375
(-4,0]    35 0.0675  7.7625
(0,4]     26 0.0675  7.7625
(4, Inf]  26 0.4325 49.7375
--------------------------
```

Asignación de Puntaje:

Logro 1: Asignar **[0.5 Ptos]** por `t-value` $= 7.814178$ y **[0.5 Ptos]** `F-statistic` $= 61.06137$.

Logro 2: Asignar **[0.2 Ptos]** por `Pr(>|t|)` $= 3.105e-12$, **[0.4 Ptos]** por `Multiple R-squared` $= 0.350813$ y **[0.4 Ptos]** por `Adjusted R-squared` $= 0.345068$.

Logro 3: Asignar **[0.5 Ptos]** por indicar que si hay regresión y **[0.5 Ptos]** por comentar que esta solo logra explicar el 35 % de la variabilidad presente en los datos.

Logro 4: Asignar **[1.0 Ptos]** por $X^2 = 3.162178$. No descontar si responde $X^2 = 159.2494$

Logro 5: Asignar **[1.0 Ptos]** por rango donde se encuentra el valor-p en base a una $\chi^2(2)$ según X^2 utilizado.

Logro 6: Asignar **[1.0 Ptos]** por concluir correctamente según valor-p.

+ 1 Punto Base

Pregunta 4

Considere una muestra aleatoria de tamaño n proveniente de una distribución Uniforme$(0,\, \theta)$.

(a) **[3.0 Ptos.]** Obtenga el estimador de momento para θ y calcule su error cuadrático medio.

(b) **[3.0 Ptos.]** Se propone como estimador para θ el máximo valor observado en la muestra, el cual distribuye Beta$(n,\, 1)$ con soporte el intervalo $(0,\, \theta)$. ¿Este estimador es más eficiente que el obtenido en (a)?

Solución

(a) El estimador de momentos de θ esta dado por $\hat{\theta} = 2\,\overline{X}$.

Del formulario se tiene que

$$\mathsf{E}(\hat{\theta}) = \theta \quad \text{y} \quad \mathsf{Var}(\hat{\theta}) = \frac{\theta^2}{3\,n}.$$

Por lo tanto

$$\mathsf{ECM}(\hat{\theta}) = \frac{\theta^2}{3\,n}$$

(b) Sea $\widetilde{\theta} = \text{máx}\{X_1, \ldots, X_n\} \sim$ Beta$(n,\, 1)$ con soporte $(0,\, \theta)$.

Del formulario se tiene que

$$\mathsf{E}(\widetilde{\theta}) = \frac{n\,\theta}{(n+1)} \quad \text{y} \quad \mathsf{Var}(\widetilde{\theta}) = \frac{n\,\theta^2}{(n+1)^2\,(n+2)}.$$

Esto implica que

$$\mathsf{ECM}(\widetilde{\theta}) = \frac{n\,\theta^2}{(n+1)^2\,(n+2)} + \left[\frac{n\,\theta}{(n+1)} - \theta\right]^2 = \frac{2\,\theta^2}{(n+1)\,(n+2)}.$$

Como

$$\frac{\mathsf{ECM}(\widetilde{\theta})}{\mathsf{ECM}(\hat{\theta})} = \frac{6\,n}{(n+1)\,(n+2)} \to 0,$$

cuanto $n \to \infty$, por lo que efectivamente $\widetilde{\theta}$ es más eficiente para estimar θ que $\hat{\theta}$.

Alternativamente como $\mathsf{E}(\widetilde{\theta}) = \frac{n\,\theta}{(n+1)} \to \theta$, es decir, es un estimador asintóticamente insesgado, bastaría compara las varianzas para determinar cual estimador es más eficiente:

$$\frac{\mathsf{Var}(\widetilde{\theta})}{\mathsf{Var}(\hat{\theta})} = \frac{3\,n^2}{(n+1)^2\,(n+2)} \to 0,$$

cuando $n \to \infty$, por lo que efectivamente $\widetilde{\theta}$ es más eficiente para estimar θ que $\hat{\theta}$.

Asignación de Puntaje:

Logro 1: Asignar **[1.0 Ptos]** por $\hat{\theta} = 2\,\overline{X}$.

Logro 2: Asignar **[0.5 Ptos]** por $\mathsf{E}(\hat{\theta}) = \theta$ y **[0.5 Ptos]** por $\mathsf{Var}(\hat{\theta}) = \frac{\theta^2}{3\,n}$.

Logro 3: Asignar **[1.0 Ptos]** por $\mathsf{ECM}(\hat{\theta}) = \frac{\theta^2}{3\,n}$.

Logro 4: Asignar **[0.5 Ptos]** por $\mathrm{E}(\widetilde{\theta}) = \dfrac{n\,\theta}{(n+1)}$ y **[0.5 Ptos]** por $\mathrm{Var}(\widetilde{\theta}) = \dfrac{n\,\theta^2}{(n+1)^2\,(n+2)}$.

Logro 5: Asignar **[1.0 Ptos]** por $\mathrm{ECM}(\widetilde{\theta}) = \dfrac{2\,\theta^2}{(n+1)\,(n+2)}$. Si indica que utilizará $\mathrm{Var}(\widetilde{\theta})$ en vez de ECM, por ser asintóticamente insesgado, asignar puntaje.

Logro 6: Asignar **[1.0 Ptos]** por mostrar que estimador propuesto en (b) es más eficiente.

+ 1 Punto Base