Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Primer Semestre 2025

Curso : Probabilidad y Estadística
Sigla : EYP1113
Profesores : Ricardo Aravena C., Maritza Márquez N., Bladimir Morales T., Ricardo Olea O., Jose Ordoñez C. y Hernán Robledo A.

INTERROGACIÓN 4

Pregunta 1

Una empresa de servicios digitales utiliza un sistema automático de respuesta para atender las solicitudes de sus clientes. El tiempo total (en segundos) que un cliente espera desde que inicia la llamada, el sistema automático la acepta y recibe atención efectiva se modela mediante una distribución Exponencial(ν) desplazada en θ. Este modelo tiene función de densidad definida como

$$f_X(x) = \nu\, e^{-\nu\,(x-\theta)}, \quad x \geq \theta,$$

donde

- $\theta \in \mathbb{R}$ representa el tiempo mínimo técnico que toma procesar una llamada, y
- $\nu > 0$ representa la rapidez del sistema (a mayor ν, menor tiempo de espera promedio).

Suponiendo que se dispone de una muestra aleatoria de n tiempos totales independientes y que θ es conocido:

(a) **[2 Ptos.]** Encuentre el estimador de máxima verosimilitud de ν.

(b) **[2 Ptos.]** Determine un intervalo de confianza del 95 % para $g(\nu) = \log(\nu^2)$.

(c) **[2 Ptos.]** Suponga que θ es igual a 0.5 segundos, ν igual a 4, y que se toma una muestra de tamaño 30. Calcule la probabilidad aproximada de que $g(\hat{\nu}) = \log(\hat{\nu}^2)$ sea mayor a 2.5.

Solución

(a) Para θ conocido, la función de verosimilitud, y su logaritmo, se desarrollan a continuación.

$$L(\nu) = \prod_{i=1}^{n} \nu e^{-\nu(x_i-\theta)} = v^n \exp\left(-\nu \sum_{i=1}^{n} x_i + n\nu\theta\right)$$

$$\ell(\nu) = \log(L(\nu)) = n\log(\nu) - \nu \sum_{i=1}^{n} x_i + n\nu\theta$$

Maximizamos $\ell(\nu)$ derivando e igualando a cero.

$$\frac{d}{d\nu}\ell(\nu) = \frac{n}{\nu} - \sum_{i=1}^{n} x_i + n\theta \stackrel{!}{=} 0 \qquad \Longrightarrow \qquad \hat{\nu} = \frac{1}{(\bar{x} - \theta)}$$

Se verifica que $\ell(\nu)$ se maximiza en $\hat{\nu}$ pues $d^2/d\nu^2\ell(\hat{\nu}) = -n/\hat{\nu}^2 < 0$.

(b) Para encontrar el Intervalo de Confianza de $g(\nu)$ basado en el estimador de máxima verosiilitud (EMV) de ν se utiliza la normalidad asintótica de los EMV. Primero, por formulario se tiene que el EMV de $g(\nu)$ es $g(\hat{\nu})$, con $\hat{\nu}$ el estimador encontrado en (a). Luego, se tiene que $\text{Var}(g(\hat{\nu})) = \frac{[g'(\nu)]^2}{I_n(\nu)}$, donde $g'(\nu) = 2/\nu$.

$$I_n(\nu) = -\text{E}\left(\frac{d^2}{d\nu^2}\ell(\nu)\right) = -\text{E}\left(-\frac{n}{\nu^2}\right) = \frac{n}{\nu^2}$$

Así, $\text{Var}(g(\hat{\nu})) = \frac{[g'(\nu)]^2}{I_n(\nu)} = (4/\nu^2)/(n/\nu^2) = 4/n$. Finalmente, usando el pivote

$$Z_n = \left(g(\hat{\nu}) - g(\nu)\right) / \sqrt{\text{Var}(g(\hat{\nu}))} \sim \text{Normal}(0,1)$$

se tiene que el Intervalo de Confianza está dado por

$$< g(\nu) >_{1-\alpha/2} \in \left(g(\hat{\nu}) \pm z_{1-\alpha/2}\sqrt{\text{Var}(g(\hat{\nu}))}\right)$$

$$< \log(\nu^2) >_{1-\alpha/2} \in \left(\log(\hat{\nu}^2) \pm 1.96\sqrt{\frac{4}{n}}\right)$$

(c) La probabilidad solicitada es $\text{Pr}\left(g(\hat{\nu}) > 2.5\right)$. Utilizando la normalidad asintótica de $g(\hat{\nu})$, es decir, que $Z_n = \left(g(\hat{\nu}) - g(\nu)\right) / \sqrt{\text{Var}(g(\hat{\nu}))} \sim \text{Normal}(0,1)$, desarrollamos la probabilidad.

$$\begin{aligned}\text{Pr}\left(g(\hat{\nu}) > 2.5\right) &= \text{Pr}\left(\frac{g(\hat{\nu}) - g(\nu)}{\sqrt{\text{Var}(g(\hat{\nu}))}} > \frac{2.5 - g(\nu)}{\sqrt{\text{Var}(g(\hat{\nu}))}}\right) \\ &= \text{Pr}\left(Z_n > \frac{2.5 - g(\nu)}{\sqrt{\text{Var}(g(\hat{\nu}))}}\right)\end{aligned}$$

Usando que $\nu = 4$ y $\theta = 0.5$, se tiene que $g(\nu) = \log(4^2) = 2.772$ y $\sqrt{\text{Var}(g(\hat{\nu}))} = \sqrt{4/n} = \sqrt{4/30} = 0.3651$. Así,

$$\begin{aligned}\text{Pr}\left(g(\hat{\nu}) > 2.5\right) &= \text{Pr}\left(Z_n > \frac{2.5 - 2.772}{0.3651}\right) \\ &= \text{Pr}\left(Z_n > -0.745\right) \\ &= \text{Pr}\left(Z_n < 0.745\right) \approx \Phi(0.74) = 0.7704\end{aligned}$$

Asignación de Puntaje:

(a)

Logro 1 [1.0 ptos]: por planteamiento de la función de verosimilitud o su versión en logaritmo.

- Asignar [0.5 pto] por indicar que $L(\nu)$ es el producto de las densidades o que $\ell(\nu)$ es la suma de las log-densidades.
- Asignar [0.5 pto] por obtener la versión desarrollada de $L(\nu)$ o $\ell(\nu)$.

Logro 2 [1.0 ptos]: por obtener que $\hat{\nu} = 1/(\bar{x} - \theta)$.

- Asignar [0.5 pto] por derivar la función e igualar a cero.
- Asignar [0.4 pto] por obtener el EMV de $\hat{\nu}$.
- Asignar [0.1 pto] por verificar que la verosimilitud es máxima en $\nu = \hat{\nu}$.

(b)

Logro 3 [1.0 ptos]: por indicar el uso del pivote Z_n o por indicar la fórmula del intervalo de confianza para una función del EMV.

- Asignar [0.4 pto] por indicar que el EMV de $g(\nu)$ es $g(\hat{\nu})$.
- Asignar [0.2 pto] por indicar indicar que Var$(g(\hat{\nu})) = g'(\nu)^2/I_n(\nu)$.
- Asignar [0.4 pto] por usar lo anterior para indicar la fórmula del Intervalo de Confianza.

Logro 4 [1.0 ptos]: por obtener la expresión analítica del intervalo de confianza.

- Asignar [0.1 pto] por obtener $g'(\nu)$.
- Asignar [0.5 pto] por obtener $I_n(\nu)$.
- Asignar [0.4 pto] por obtener la expresión correcta del intervalo de confianza.

(c)

Logro 5 [1.0 ptos]: por indicar el uso de la normalidad asintótica para el cálculo de la probabilidad solicitada (la asignación de puntaje es similar al ítem (b)).

- Asignar [0.4 pto] por indicar que el EMV de $g(\nu)$ es $g(\hat{\nu})$.
- Asignar [0.2 pto] por indicar indicar que Var$(g(\hat{\nu})) = g'(\nu)^2/I_n(\nu)$.
- Asignar [0.4 pto] por indicar la normalidad asintótica del EMV de $g(\nu)$.

Logro 6 [1.0 ptos]: por obtener que Pr$(g(\hat{\nu} > 2.5) = 0.6985$.

- Asignar [0.1 pto] por obtener $g'(\nu)$.
- Asignar [0.5 pto] por obtener $I_n(\nu)$.
- Asignar [0.4 pto] por obtener la probabilidad correcta haciendo uso de las propiedades de la distribución normal.

Pregunta 2

En la industria del salmón chileno, las autoridades sanitarias controlan la presencia de antibióticos en los filetes antes de su exportación. Si más del 20 % de los filetes de un lote contiene residuos sobre el límite permitido, el lote completo es rechazado. Además, para esos mismos filetes se mide el contenido de omega-3 (en $\mathrm{g/100g}$), el cual se asume que siguen una distribución Normal.

De un lote se extrajo una muestra aleatoria de 120 filetes de salmón, registrándose el promedio y la desviación estándar del contenido de omega-3 según si el filete presenta residuos por sobre el límite permitido o no:

	filetes con residuos sobre límite	filetes con residuo bajo límite
n	30.0	90.0
promedio	1.8	2.5
desv. est.	0.4	0.6

Responda las preguntas (a) y (b) indicando la hipótesis nula y alternativa con los parámetros involucrados, el valor del estadístico de prueba, el valor-p o el valor crítico, y la decisión basada en los resultados obtenidos.

(a) **[2 Ptos.]** ¿Existe suficiente evidencia para rechazar el lote por superar el 20 % de filetes con residuos de antibióticos, utilizando un nivel de significancia del 5 %?

(b) **[2 Ptos.]** ¿Existe suficiente evidencia para afirmar que el contenido de omega-3 en los filetes con residuos es inferior a $2\,\mathrm{g/100g}$, utilizando un nivel de significancia del 1 %?

(c) **[2 Ptos.]** Calcule el tamaño de muestra necesario para estimar la proporción de filetes con residuos con un error máximo de ± 0.05 y un nivel de confianza del 90 %. Para este cálculo, considere el caso más desfavorable, en el que la varianza de una variable de tipo Bernoulli es máxima, es decir, $p(1-p) = \frac{1}{4}$.

Solución

(a) Se pide contrastar las siguientes hipótesis:

$$H_0 : p = 0.20 \quad \text{vs.} \quad H_a : p > 0.20.$$

Calculamos la proporción muestral:

$$\hat{p} = \frac{30}{120} = 0.25.$$

Bajo H_0 se tiene que

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

Se rechaza H_0 si $Z_0 > Z_{1-\alpha} = Z_{0.95} = 1.645$. Para $n = 120$ y $\hat{p} = 0.25$ se tiene que

$$Z_0 = \frac{0.25 - 0.20}{0.03651} = 1.369.$$

Como $Z_0 = 1.369 < 1.645$, no se rechaza la hipótesis nula, concluyendo que no existe evidencia para rechazar el lote.

De manera equivalente, el valor-p se calcula como valor-p $= \text{Pr}(Z > Z_0) = \text{Pr}(Z > 1.369) = 1 - \Phi(1.369) = 1 - 0.9131 = 0.0869$. Como el valor-p es mayor a 0.05, no existe evidencia suficiente para rechazar la hipótesis nula, concluyendo que no existe evidencia para rechazar el lote.

(b) Para los filetes con residuos sobre el límite ($n = 30$), se pide contrastar las siguientes hipótesis:

$$H_0 : \mu = 2 \quad \text{vs.} \quad H_a : \mu < 2.$$

Como la varianza se calcula usando la muestra y se desconoce la varianza poblacional, se tiene que bajo H_0

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

Del enunciado, $n = 30$, $\overline{X} = 1.8$ y $S = 0.4$

$$T_0 = \frac{1.8 - 2}{0.4/\sqrt{30}} = \frac{-0.2}{0.07306} = -2.737.$$

Utilizando la tabla de percentiles t-student$(n - 1 = 29)$, se rechaza H_0 si $T_0 < t_{0.01}(29) = -2.462$. Como $T_0 = -2.737$, se rechaza H_0, concluyendo que sí existe evidencia para afirmar que el contenido de omega-3 en los filetes con residuos sobre el límite es inferior a 2 (g/100g).

Equivalentemente, de la tabla de percentiles de la distribución t-student se tiene que el valor-p es menor a 0.01, ya que $t_{0.99}(29) = 2.457 < -T_0 = 2.737$. Así, se rechaza H_0, concluyendo que sí existe evidencia para afirmar que el contenido de omega-3 en los filetes con residuos sobre el límite es inferior a 2 (g/100g).

(c) La fórmula del tamaño muestral está dado por $n = \left(k_{1-\alpha/2}\sigma/w\right)^2$. Del enunciado se pide una confianza del 90 %, estableciendo una significancia $\alpha = 0.1$; un error máximo de $w = 0.05$; y una varianza $\sigma^2 = p(1-p) = 1/4$.

$$n = \left(\frac{k_{0.95} \cdot \sigma}{w}\right)^2 = \frac{1.645^2 \cdot 0.25}{0.05^2} = \frac{2.706 \cdot 0.25}{0.0025} \approx 270.6.$$

Asignación de Puntaje:

(a)

Logro 1 [0.5 ptos]: por indicar la hipótesis nula y alternativa correctamente.

Logro 2 [0.5 ptos]: por seleccionar el estadístico de prueba correcto.

Logro 3 [0.5 ptos]: por obtener el valor-p o la región de rechazo junto al estadístico para tomar la decisión.

Logro 4 [0.5 ptos]: Por tomar la decisión correcta respecto a la hipótesis nula, justificando con base en el contraste de hipótesis. Este logro podrá ser evaluado positivamente si la hipótesis nula y alternativa están correctamente formuladas y se ha identificado adecuadamente el estadístico de prueba, incluso si existen errores en el cálculo del valor-p o la determinación de la región de rechazo.

(b)

Logro 5 [0.5 ptos]: equivalente a Logro 1.

Logro 6 [0.5 ptos]: equivalente a Logro 2.

Logro 7 [0.5 ptos]: equivalente a Logro 3.

Logro 8 [0.5 ptos]: equivalente a Logro 4.

(c)

Logro 9 [1.0 ptos]: por indicar la fórmula de tamaño muestral a utilizar.

Logro 10 [1.0 ptos]: por obtener el tamaño muestral correcto.

- Asignar [0.5 pto] por usar que $\sigma = \sqrt{1/4}$.
- Asignar [0.3 pto] por obtener $k_{0.95} = 1.645$.
- Asignar [0.2 pto] por obtener que $n = 270.6$, o, si redondea, $n = 271$.

Pregunta 3

El salario al inicio de la vida laboral es un factor muy relevante; por ello, cada vez existe más publicidad relacionada con los ingresos durante el primer año tras el egreso en distintas carreras. Usted está interesado en conocer los ingresos de los egresados de Ingeniería, por lo que decide enviar una encuesta a una lista de 166 egresados del año 2024 (94 de la UC y 72 de otras universidades). Al cabo de una semana, recibe 30 respuestas sobre el ingreso percibido en el último mes (medido en U.F.), cuyas estadísticas, según la universidad, se presentan a continuación:

	UC	Otras Universidades	Total
N° Egresados en lista	94	72	166.0
N° Respuestas obtenidas	18	12	30.0
Media de Ingresos	45	41	43.4
Desv. Est. de Ingresos	6	10	7.8

Asumiendo una distribución normal para los ingresos en U.F., y considerando un nivel de significancia del 10 %, responda las siguientes preguntas indicando la hipótesis nula y alternativa con los parámetros involucrados, el valor del estadístico de prueba, el valor-p o el valor crítico, y la decisión basada en los resultados obtenidos.

(a) **[2 Ptos.]** ¿Existe evidencia que permita afirmar que las tasas de respuesta difieren entre los dos grupos?.

(b) **[2 Ptos.]** ¿Existe evidencia para afirmar que las varianzas de los ingresos de los egresados UC y de las otras universidades difieren?

(c) **[2 Ptos.]** ¿Existe evidencia para afirmar que los egresados de la UC muestran un ingreso medio superior a los egresados de otras universidades? Asuma que las varianzas de los ingresos no difieren.

Solución

(a) Que un encuestado responda o no a la encuesta se puede modelar como una variable aleatoria Bernoulli. En este caso se dispone de dos poblaciones: los encuestados de la UC y los encuestados de otras universidades. Así, la hipótesis a evaluar corresponde a

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_a : p_1 \neq p_2,$$

donde p_1 es la tasa de respuestas bajo UC, y p_2 es la tasa de respuesta de otras universidades. Del formulario se extrae el estadístico de prueba dado por

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \sim \text{Normal}(0, 1)$$

Bajo H_0, $p_1 = p_2$, lo que implica que se asume una única población de cuya tasa de respuesta es $p = p_1 = p_2$. Considerando esta única población, se tiene que este p se estima por $\hat{p} = (18 + 12)/(94 + 72) = 0.18$. Así,

$$Z = \frac{(0.191 - 0.166) - 0}{\sqrt{0.18 \cdot (1 - 0.18)\left(\frac{1}{94} + \frac{1}{72}\right)}} = 0.4155.$$

Finalmente, se rechaza H_0 si $|Z| > k_{1-\alpha/2} = 1.645$. Como $Z = 0.4155$, no existe suficiente para rechazar H_0, por lo que no se puede afirmar que las tasas de respuestas difieren entre los dos grupos.

(b) Se solicita realizar una prueba de comparación de varianzas entre dos poblaciones dadas por las hipótesis

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_a : \sigma_1^2 \neq \sigma_2^2,$$

donde σ_1^2 y σ_2^2 son las varianzas de los encuestados de la UC y de otras universidades, respectivamente. y Del formulario se tiene que si S_1 y σ_1^2 son las varianzas muestrales y poblacionales de ambos grupos, entonces

$$F_0 = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$$

Bajo H_0, $\sigma_1^2 = \sigma_2^2$. Luego, $F_0 = 6^2/10^2 = 0.36$. Se rechaza H_0 si $F_0 > F_{0.95}(17, 11) = 2.69$ o si $F_0 < F_{0.05}(17, 11) = 1/F_{0.95}(11, 17) = 1/2.41 = 0.414$. Como $F_0 = 0.36 < 0.414$, existe evidencia suficiente para indicar que las varianzas de los ingresos entre los egresados UC y de las otras universidades difieren.

(c) Se solicita realizar una prueba de comparación de medias entre dos poblaciones asumiendo igualdad de varianzas. Esto es, evaluar las hipótesis

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vs} \quad H_a : \mu_1 > \mu_2,$$

donde μ_1 y μ_2 son las medias de los ingresos de los egresados de la UC y de otras universidades, respectivamente. Del formulario, al asumir igualdad de varianzas, el estadístico de prueba está dado por

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

donde s_p^2 es la varianza combinada de ambas muestras:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Con los datos disponibles:

- $\bar{x}_1 = 45,\ s_1 = 6,\ n_1 = 18$ (UC)
- $\bar{x}_2 = 41,\ s_2 = 10,\ n_2 = 12$ (No UC)

se obtiene:

$$s_p^2 = \frac{(17)(36) + (11)(100)}{28} = \frac{612 + 1100}{28} = \frac{1712}{28} \approx 61.14, \quad s_p \approx 7.82.$$

Luego, el estadístico t es:

$$t = \frac{45 - 41}{7.82 \cdot \sqrt{\frac{1}{18} + \frac{1}{12}}} \approx \frac{4}{7.82 \cdot 0.372} \approx \frac{4}{2.91} \approx 1.374.$$

Se rechaza H_0 si $t > t_{0.10}(28) = 1.311$, donde este último valor se obtiene de la tabla de percentiles de la distribución t-student dada en el formulario. Como $t = 1.374 > 1.311$, se rechaza H_0. Por lo tanto, al 10 % de significancia, existe evidencia suficiente para afirmar que el ingreso medio de los egresados UC es mayor al de los egresados de otras universidades.

Alternativamente, se puede dar un intervalo para acotar el valor-p, pues según el formulario $1.313 = t_{0.90}(28) < \text{Pr}(t(28) > 1.374) t_{0.90}(28) < t_{0.95}(28) = 1.701$. Así, el valor-p se espera que esté entre 0.5 y 0.10, por lo que se rechaza la hipótesis nula a una significancia del 10 %.

Asignación de Puntaje:

(a)

Logro 1 [0.5 ptos]: por indicar la hipótesis nula y alternativa correctamente.

Logro 2 [0.5 ptos]: por seleccionar el estadístico de prueba correcto.

Logro 3 [0.5 ptos]: por obtener el valor-p o la región de rechazo junto al estadístico para tomar la decisión.

Logro 4 [0.5 ptos]: Por tomar la decisión correcta respecto a la hipótesis nula, justificando con base en el contraste de hipótesis. Este logro podrá ser evaluado positivamente si la hipótesis nula y alternativa están correctamente formuladas y se ha identificado adecuadamente el estadístico de prueba, incluso si existen errores en el cálculo del valor-p o la determinación de la región de rechazo.

(b)

Logro 5 [0.5 ptos]: equivalente a Logro 1.

Logro 6 [0.5 ptos]: equivalente a Logro 2.

Logro 7 [0.5 ptos]: equivalente a Logro 3.

Logro 8 [0.5 ptos]: equivalente a Logro 4.

(c)

Logro 9 [0.5 ptos]: equivalente a Logro 1.

Logro 10 [0.5 ptos]: equivalente a Logro 2.

Logro 11 [0.5 ptos]: equivalente a Logro 3.

Logro 12 [0.5 ptos]: equivalente a Logro 4.

Pregunta 4

Durante el verano, el consumo de agua potable por habitante tiende a incrementarse debido a factores ambientales como temperatura, humedad y radiación solar. Para analizar este fenómeno, se solicita a una empresa sanitaria de la Región de Valparaíso información diaria sobre el consumo promedio de agua potable por persona (en litros), en 45 días seleccionados aleatoriamente entre los años 2015 y 2023. Además, se obtuvieron datos provenientes de estaciones meteorológicas cercanas para los días seleccionados, obteniendo una tabla con los siguientes datos.

- `waterconsumption`: Consumo diario de agua potable por persona (litros).
- `temp.max`: Temperatura máxima del día (°C).
- `humidity.avg`: Humedad relativa promedio del día (proporción entre 0 y 1).
- `solarrad`: Radiación solar promedio diaria (W/m^2).

(a) **[2 ptos]** Para los siguientes tres modelos de regresión simple, complete la información faltante:

```
summary(waterconsumption)
-------------------------------------------------------
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     sd.
  152.3   178.9   186.9   186.9   195.0   213.8   13.8
-------------------------------------------------------
```

```
-------------------------------------------------------------
Modelo 1: lm(waterconsumption ~ temp_max)

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.9984     7.3601   15.08  < 2e-16
temp_max      2.6311     0.2523   10.43 2.39e-13

Residual standard error: 7.456 on 43 degrees of freedom
Multiple R-squared:  X.XXXX, Adjusted R-squared:   X.XXXX
F-statistic: 108.7 on 1 and 43 DF,  p-value: 2.389e-13
-------------------------------------------------------------
```

```
-------------------------------------------------------------
Modelo 2: lm(waterconsumption ~ humidity_avg)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)    207.818      6.629  31.350  < 2e-16
humidity_avg  -35.938     10.909   X.XXX  0.00198

Residual standard error: 12.52 on 43 degrees of freedom
Multiple R-squared:  0.2015, Adjusted R-squared:  0.183
F-statistic: XX.XX on 1 and 43 DF,  p-value: 0.00198
-------------------------------------------------------------
```

```
Modelo 3: lm(waterconsumption ~ solarrad)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 171.32312   12.92764  13.252   <2e-16
solarrad      0.02750    0.02259   1.217   X.XXXX

Residual standard error: 13.77 on 43 degrees of freedom
Multiple R-squared:  0.03332, Adjusted R-squared:  0.01084
F-statistic: 1.482 on 1 and 43 DF,  p-value: X.XXXX
-------------------------------------------------------------
```

(b) **[1 Ptos]** ¿Son los Modelos 1 y 2 significativos al 10 % de significancia? Justifique.

(c) **[1 Ptos]** Entre el Modelo 2 y el Modelo 3, ¿cuál presenta un mayor porcentaje de variabilidad explicada? Justifique.

(d) **[2 Ptos]** Uno de los supuestos del modelo de regresión simple, es que el error del modelo distribuye Normal$(0, \sigma)$. Para verificar este supuesto, usted realiza una prueba de bondad de ajuste χ^2 sobre los residuos del mejor modelo ajustado en el item (a), considerando el `Residual standard error` como estimador de σ y la siguiente tabla:

```
-----------------------------------------
          Observado Probabilidad Esperado
-----------------------------------------
(-Inf,-9]         6       0.1137   5.1165
(-9,-3]           9       0.2300  10.3500
(-3,3]           12       0.3126  XX.XXXX
(3,9]            14       X.XXXX  XX.XXXX
(9, Inf]          4       0.1137   5.1165
-----------------------------------------
```

Complete la información faltante, obtenga el estadístico de prueba y calcule el valor-p. ¿Existe evidencia al 5 % de significancia para rechazar el supuesto de normalidad de los errores?

Solución

(a) Tenemos que

Para el Modelo 1:

- Calculamos el coeficiente de determinación R^2, `Multiple R-squared`, con la fórmula

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n-2}{n-1}\frac{s^2_{Y|x}}{s^2_Y}$$

s^2_Y es la varianza muestral, por ello, de la primera tabla de estadística se tiene que $s^2_Y = 13.8^2 = 190.44$. $s^2_{Y|x}$ corresponde a $s^2_{Y|x} = (\texttt{Residual standard error})^2 = (7.456)^2 = 55.591$. Así,

$$R^2 = 1 - \frac{43}{44} \cdot \frac{55.591}{190.44} = 0.7147$$

- Calculamos el coeficiente de determinación ajustado r^2, `Adjusted R-squared`, con la fórmula

$$r^2 = 1 - \frac{s^2_{Y|x}}{s^2_Y} = 1 - \frac{55.591}{190.44} = 0.708$$

Para el Modelo 2:

- Valor-t incógnito: $\texttt{t-value} = \texttt{Estimate}/\texttt{Std. Error} = -35.938/10.909 = -3.2943$.
- Estadístico F incógnito: $\texttt{F-statistic} = \texttt{t value}^2 = (-3.294)^2 = 10.8527$.

Para el Modelo 3:

- Valor-p incógnito de prueba t: $\texttt{Pr(>|t|)} = \Pr(|t_{n-2}| > 1.217) = 2(1 - \Pr(t_{43} < 1.217)) \approx 2(1 - \Pr(Z < 1.217)) = 2(1 - 0.8869) = 0.2262$.
- El Valor-p incógnito de la prueba F es igual al valor-p de la prueba t anterior. $\texttt{p-value} = 0.2262$.

(b) Para verificar si un modelo es significativo se revisa el resultado de la prueba F de significancia global del modelo. Para el Modelo 1, se tiene que su valor-p = $2.389 \cdot 10^{-13} < 0.10$, por lo que se rechaza la hipótesis de que el modelo no sea significativo. Por otra parte, el modelo 2 también es significativo pues su valor-p asociado a la prueba F es valor-p = $0.00198 < 0.1$.

(c) El porcentaje de variabilidad explicada por un modelo está indicado por el `Multiple R-squared` que aparece en la salida. El Modelo 2 explica una mayor proporción de la variabilidad que el Modelo 3: un 20.15 % frente a un 3.3 %.

(d) La columna `Esperado` se calcula como $\texttt{Observado} \times \texttt{Probabilidad}$. Tras completar la tercera fila, se completa el valor de probabilidad incógnito de manera que la columna sume 1 (por ser probabilidades), y con esto se calcula el último valor incógnito.

	Observado	Probabilidad	Esperado
(-Inf,-9]	6	0.1137	5.1165
(-9,-3]	9	0.2300	10.3500
(-3,3]	12	0.3126	14.0670
(3,9]	14	0.2300	10.3500
(9, Inf]	4	0.1137	5.1165

Por formulario, calculamos el estadístico de prueba χ^2, el cual sigue una distribución χ^2 con 3 grados de libertad, ya que se suman cinco intervalos ($k = 5$), y se posee un parámetro libre estimado (el σ del modelo Normal) ($\nu = 1$). Así,

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = 2.16 \sim \chi^2(3)$$

De la tabla de percentiles χ^2 del formulario, se tiene que $c_{0.4}(3) < 2.16 < c_{0.6}(3)$. Así, se tiene que el valor-p = Pr$(\chi^2(3) > 2.16) \in (0.4, 0.6)$. Por ello, no existe evidencia estadística para rechazar el supuesto de normalidad de los errores.

Asignación de Puntaje:

(a)

Logro 1 [2 ptos]: por obtener los seis valores incógnito.

- Asignar [0.35 pto] por cada valor calculado.

Si todos los valores incógnitos son correctos, el puntaje de este logro se redondeará a 2 puntos. En caso contrario, el puntaje obtenido se redondeará al primer decimal.

(b)

Logro 2 [1 ptos]: por interpretar correctamente la salida de la prueba F de cada modelo.

- [0.5 pto] por cada interpretación correcta.

(c)

Logro 3 [1 ptos]: por interpretar correctamente los coeficientes de determinación de cada modelo.

- [0.4 pto] por cada interpretación correcta.
- [0.2 pto] por indicar cuál modelo explica un mayor porcentaje de variabilidad.

Si se utiliza el `Adjusted R-squared` en vez del `Multiple R-squared`, descontar 0.1 ptos por cada vez que este sea utilizado.

(d)

Logro 4 [0.5 ptos]: por rellenar la tabla.

Logro 5 [1.5 ptos]: por calcular el estadístico χ^2 y concluir correctamente.

Tiempo: 120 minutos