CPE (SEGUNDO CURSO)

PRÁCTICA 18 **SOLUCIONES** (**Curso 2023–2024**)

1.– En la boya meteorológica cercana a Cabo Vilán (A Coruña) se miden las mayores olas que se producen en la costa española. Si nos fijamos en aquellas cuya altura iguale o supere los h m, su llegada a la boya puede representarse mediante un proceso de Poisson de parámetro θ olas/año. La altura X de estas olas viene definida por una distribución exponencial trasladada ($x \geq h$) con parámetro λ. Evaluar:

a) La altura máxima de ola Y registrada en 2 años.

b) La probabilidad de que $Y = h$. ¿Qué significa este resultado?

c) La probabilidad de que $Y > 30$. Coméntese este resultado y, en consecuencia, la validez del modelo.

Datos: $h = 5$, $\theta = 15$, $\lambda = 1/7$

Nota: La mayor ola registrada en las costas españolas fue de 27.81 m, en la boya de Villano-Sisargas (Cabo Vilán), el día 6 de enero de 2014.

Nota: Se recuerda que $e^a = \displaystyle\sum_{k=0}^{\infty} \frac{a^k}{k!}$.

——————————SOLUCIÓN——————————

a) Sea Y la altura máxima alcanzada en los dos años. Sea N el número de veces que el suceso ha ocurrido durante esos dos años. N sigue una distribución de Poisson de media 2θ (dado que la distribución de Poisson es regenerativa con respecto a la suma), y por el Teorema de la Probabilidad Total, para todo $y \geq 5$

$$P[Y \leq y] = \sum_{n=0}^{\infty} P[Y \leq y | N = n]\; P[N = n]$$

Sabemos que la distribución de X es una exponencial trasladada de forma que $x \geq 5$. Por tanto

$$f_X(x) = \lambda e^{-\lambda(x-5)}, \quad x \geq 5$$

Calculemos la distribución de la variable condicionada $Y|N = n$. Supuesto que ha sucedido n veces el fenómeno, el hecho de que el máximo nivel alcanzado sea menor o igual que y es la intersección de los sucesos consistentes en que cada una de las n veces el nivel haya sido menor o igual que y; asumiendo que los niveles alcanzados en distintas ocurrencias del fenómeno son independientes, y teniendo en cuenta que estos niveles siguen la misma distribución (exponencial trasladada de parámetro λ), se deduce $P[Y \leq y|N = n] = P[X \leq y]^n$ siendo X una exponencial trasladada de parámetro λ, es decir,

$$P[Y \leq y|N = n] = (1 - e^{-\lambda(y-5)})^n$$

y por lo tanto

$$P[Y \leq y] = \sum_{n=0}^{\infty} (1 - e^{-\lambda(y-5)})^n e^{-2\theta} \frac{(2\theta)^n}{n!} = e^{-2\theta} \sum_{n=0}^{\infty} \frac{(2\theta(1 - e^{-\lambda(y-5)}))^n}{n!}$$
$$= e^{-2\theta} e^{2\theta(1-e^{-\lambda(y-5)})} = e^{-2\theta e^{-\lambda(y-5)}}, \quad y \geq 5$$

La determinación completa de la distribución de Y es

$$F_Y(y) = P[Y \leq y] = \begin{cases} e^{-2\theta e^{-\lambda(y-5)}} & \text{si } y \geq 5 \\ 0 & \text{si } y < 5 \end{cases}$$

y sustituyendo los datos numéricos

$$F_Y(y) = P[Y \leq y] = \begin{cases} e^{-30 e^{-(y-5)/7}} & \text{si } y \geq 5 \\ 0 & \text{si } y < 5 \end{cases}$$

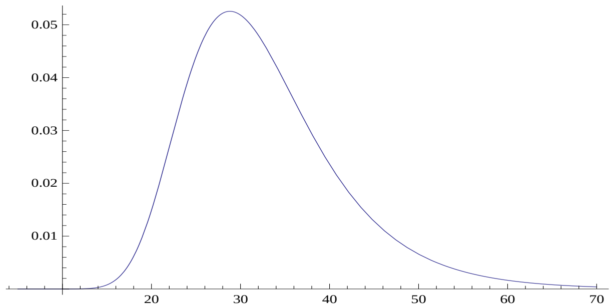
La función de densidad es fácilmente deducible, sin más que derivar la expresión anterior

$$f_Y(y) = \begin{cases} 2\lambda\theta e^{-\lambda(y-5) - 2\theta e^{-\lambda(y-5)}} & \text{si } y \geq 5 \\ 0 & \text{si } y < 5 \end{cases}$$

y sustituyendo los datos numéricos

$$f_Y(y) = \begin{cases} (30/7) e^{-(y-5)/7 - 30 e^{-(y-5)/7}} & \text{si } y \geq 5 \\ 0 & \text{si } y < 5 \end{cases}$$

Esta función de densidad puede verse en la siguiente figura



b) $P[Y = 5]$ es el salto de la función de distribución de Y en $y = 5$, es decir, la distancia entre los dos límites laterales de F_Y en ese punto.

$$F_Y(5^-) = 0, \quad F_Y(5^+) = F_Y(5) = e^{-30 e^{-\lambda \cdot 0}} = e^{-30} \Rightarrow P[Y = 5] = e^{-30} \approx 10^{-13}.$$

Notar que este valor coincide con el de la probabilidad de que se produzcan 0 llegadas en una Poisson de media 30: si el nivel máximo obtenido es cero podemos suponer (en términos de probabilidad) que el fenómeno no ha llegado a producirse ni una sola vez.

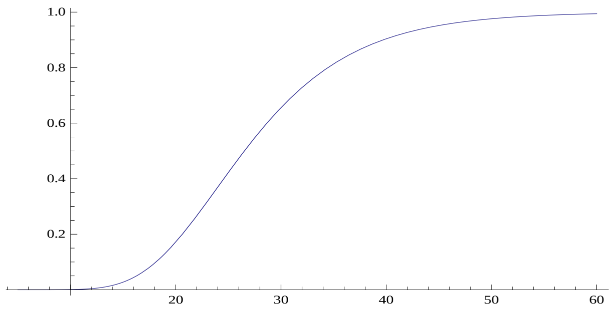
c) La probabilidad pedida es

$$P[Y > 30] = 1 - P[Y \leq 30] = 1 - F_Y(30) = 1 - e^{-30 e^{-(30-5)/7}} = 0.5698$$

Es decir, que de acuerdo con el modelo utilizado, la probabilidad de que en dos años se superen los treinta metros de ola es del orden del 57 %. Pero sabemos que la mayor ola conocida es de

27.81 m. Luego el modelo está mal. Y el error está, con toda probabilidad, en el modelado exponencial de la altura de ola.

Obsérvese que la función de distribución acumulada de Y es



lo que nos proporciona unas alturas de ola enormes con unas probabilidades bastante grandes, lo que no modela adecuadamente el fenómeno físico que estudiamos.

2.– Durante la fase de proyecto de la presa de Asuan (en Egipto) uno de los elementos críticos para el diseño fué el tamaño de los aliviaderos y, consecuentemente, de la capacidad de desagüe de la presa en un momento determinado. La presa se proyectó para una vida útil de m años. El número de avenidas del Nilo al año puede considerarse Poisson con parámetro ν (avenidas/año). Cada avenida alcanza un caudal X, aleatorio, que puede considerarse uniformemente distribuido entre 0 y b (en m^3/s).

a) Calcular la distribución del valor Y que han de ser capaces de desaguar los aliviaderos de la presa para que puedan resistir la avenida máxima que se produzca durante la vida útil de la presa.

b) ¿Cuánto vale la probabilidad de que $Y = 0$? Interpretar este resultado.

Datos: $m = 150$, $\nu = 3$, $b = 10000$. *Estos datos son informativos. No es necesario usarlos en la resolución del problema, excepto para interpretar el segundo apartado.*

Especificar claramente las hipótesis que se realicen.

SOLUCIÓN

a) Sea Y el caudal máximo alcanzado en m años. Sea N el número de veces que el suceso "avenida" ha ocurrido durante esos m años. N sigue una distribución de Poisson de media $m\nu$, y por el Teorema de la Probabilidad Total, para todo $y \geq 0$

$$P[Y \leq y] = \sum_{n=0}^{\infty} P[Y \leq y | N = n]\ P[N = n]$$

Calculemos la distribución de la variable condicionada $Y|N = n$. Supuesto que ha sucedido n veces el fenómeno, el hecho de que el máximo caudal alcanzado sea menor o igual que y es la intersección de los sucesos consistentes en que cada una de las n veces el caudal haya sido menor o igual que y; asumiendo que los caudales alcanzados en distintas ocurrencias del fenómeno son independientes, y teniendo en cuenta que estos niveles siguen la misma distribución, se deduce $P[Y \leq y|N = n] = P[X \leq y]^n$ siendo X una variable uniformemente distribuida con rango $[0, b]$), es decir,

$$P[Y \leq y|N = n] = (y/b)^n$$

y por lo tanto

$$P[Y \leq y] = F_Y(y) = \sum_{n=0}^{\infty} (y/b)^n e^{-m\nu} \frac{(m\nu)^n}{n!} = e^{-m\nu} \sum_{n=0}^{\infty} \frac{(m\nu(y/b))^n}{n!} = e^{-m\nu} e^{m\nu(y/b)} =$$

$$= e^{m\nu((y/b)-1)}, \quad 0 \leq y \leq b$$

La función de densidad es, obviamente, sin más que derivar,

$$f_Y(y) = (m\nu/b) e^{m\nu((y/b)-1)}, \quad 0 \leq y \leq b$$

b) $P[Y = 0]$ es el salto de la función de distribución de Y en $y = 0$, es decir, la distancia entre los dos límites laterales de F_Y en ese punto.

$$F_Y(0^-) = 0, \quad F_Y(0^+) = F_Y(0) = e^{m\nu(-1)} = e^{-m\nu} \Rightarrow P[Y = 0] = e^{-m\nu}.$$

Notar que este valor coincide con el de la probabilidad de que se produzcan 0 llegadas en una Poisson de media $m\nu$: si el caudal máximo obtenido es cero podemos suponer (en términos de probabilidad) que no ha habido ninguna avenida que necesitase desagüe en los m años de vida útil de la presa..

Independientemente, para los datos suministrados, esta probabilidad vale

$$P[Y = 0] = 3.7 \times 10^{-196}$$

es decir, cero.

3.– Se supone que los caudales anuales máximos de un río siguen una distribución de Gumbel. A lo largo de 10 años los caudales máximos fueron, en unidades arbitrarias, 10.1, 11.3, 17.8, 13.5, 16.1, 10.6, 14.2, 12.8, 13.2 y 14.5. Estimar los parámetros de la distribución por el método de los momentos y por el de máxima verosimilitud (aproximadamente, si es necesario).

——————————SOLUCIÓN——————————

El método de los momentos da resultados inmediatos. Sabemos que, en una distribución de Gumbel con parámetros α y μ se verifica

$$m = \mu + \frac{\gamma}{\alpha}, \qquad \sigma^2 = \frac{\pi^2}{6\alpha^2}, \qquad \gamma = 0.577$$

de donde obtenemos

$$\alpha = \frac{\pi}{\sigma\sqrt{6}}, \qquad \mu = m - \frac{\gamma\sigma\sqrt{6}}{\pi}$$

y como $\hat{m} = \bar{x}$ y $\hat{\sigma} = S$, donde $\bar{x}$ y S son la media y varianza muestrales, obtenemos

$$\boxed{\hat{\alpha} = \frac{\pi}{S\sqrt{6}}, \qquad \hat{\mu} = \bar{x} - \frac{\gamma S\sqrt{6}}{\pi}}$$

El cálculo de los estimadores mediante el método de máxima verosimilitud es más complejo. La función de densidad de una variable con distribución de Gumbel es

$$f_X(x) = \alpha e^{-\alpha(x-\mu)-e^{-\alpha(x-\mu)}}, \qquad R_X = (-\infty, \infty)$$

Por tanto, la función de verosimilitud para una muestra aleatoria de tamaño n es

$$L = \prod_{i=1}^{n} \alpha e^{-\alpha(x_i-\mu)-e^{-\alpha(x_i-\mu)}} = \alpha^n e^{\left[\sum -\alpha(x_i-\mu)-e^{-\alpha(x_i-\mu)}\right]}$$

Tomando logaritmos

$$\ln L = n \ln \alpha - \sum \left[\alpha(x_i-\mu)\right] - \sum \left[e^{-\alpha(x_i-\mu)}\right]$$

Derivando con respecto a μ y con respecto a α para hallar el máximo

$$\frac{\partial \ln L}{\partial \mu} = n\alpha - \alpha \sum e^{-\alpha(x_i-\mu)}$$

$$\frac{\partial \ln L}{\partial \alpha} = \frac{n}{\alpha} - \sum (x_i-\mu) + \sum (x_i-\mu) e^{-\alpha(x_i-\mu)}$$

y el sistema a resolver sería

$$\sum e^{-\hat{\alpha}(x_i-\hat{\mu})} = n$$

$$\frac{n}{\hat{\alpha}} - \sum (x_i-\hat{\mu}) + \sum (x_i-\hat{\mu}) e^{-\hat{\alpha}(x_i-\hat{\mu})} = 0$$

que es altamente no lineal y sólo resoluble mediante métodos numéricos en cada caso. Lo único razonablemente factible es aproximar las funciones exponenciales por aproximaciones lineales ($e^x \approx 1 + x$). Con esta simplificación (algo extrema) obtenemos

$$\sum \left[1 - \hat{\alpha}(x_i-\hat{\mu})\right] = n$$

$$\frac{n}{\hat{\alpha}} - \sum (x_i-\hat{\mu}) + \sum (x_i-\hat{\mu}) \left[1 - \hat{\alpha}(x_i-\hat{\mu})\right] = 0$$

De la primera ecuación obtenemos

$$n - \sum [-\hat{\alpha}(x_i-\hat{\mu})] = n; \quad \Rightarrow \quad \sum [-\hat{\alpha}(x_i-\hat{\mu})] = 0; \quad \Rightarrow \quad \boxed{\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}}$$

y de la segunda

$$\frac{n}{\hat{\alpha}} - \hat{\alpha} \sum (x_i-\hat{\mu})^2 = 0; \quad \Rightarrow \quad \frac{n}{\hat{\alpha}^2} = \sum (x_i-\bar{x})^2; \quad \Rightarrow \quad \hat{\alpha}^2 = \frac{1}{S^2} \quad \Rightarrow \quad \boxed{\hat{\alpha} = \frac{1}{S}}$$

Con los datos que tenemos resulta:

Método de los momentos: $\boxed{\hat{\mu} = 12.38362099,\ \hat{\alpha} = 0.5621705}$

Método de máxima verosimilitud: $\boxed{\hat{\mu} = 13.41,\ \hat{\alpha} = 0.438322541}$

4.– Sea un proceso de Bernouilli, en el que se realizan n pruebas independientes. Sea X el número de éxitos en esas n pruebas. Se proponen los dos siguientes estimadores de la probabilidad de éxito p.

$$\widehat{p}_o = \frac{X}{n}$$

$$\widehat{p}_1 = \frac{(X+1)}{(n+1)}$$

Contestar a las siguientes preguntas:

a) ¿Son dichos estimadores sesgados?

b) ¿Son asintóticamente insesgados?

c) ¿Son consistentes?

d) ¿Cuál es el de menor varianza?

e) Para un tamaño de muestra $n = 5$, ¿cuál tiene menor error cuadrático medio?

f) A la vista de los resultados anteriores, ¿cuál de los dos estimadores ha de utilizarse?

——————————SOLUCIÓN——————————

La variable aleatoria X es obviamente binomial con parámetros n y p.

a) Comprobemos si son sesgados:

$$E[\widehat{p}_o] = E\left[\frac{X}{n}\right] = \frac{E[X]}{n} = \frac{np}{n} = p \implies \text{Insesgado}$$

$$E[\widehat{p}_1] = E\left[\frac{(X+1)}{(n+1)}\right] = \frac{E[X]+1}{n+1} = \frac{np+1}{n+1} \neq p \implies \text{Sesgado}$$

b) Comprobemos si son asintóticamente insesgados:

$\widehat{p}_o$ es asintóticamente insesgado porque es insesgado. En cuanto a $\widehat{p}_1$

$$\lim_{n\to\infty} E[\widehat{p}_1] = \lim_{n\to\infty} \frac{np+1}{n+1} = p \implies \text{Asintóticamente insesgado}$$

c) Para que sean consistentes han de ser asintóticamente insesgados y de varianza tendente a cero cuando el tamaño de la muestra tiende a infinito. Calculemos, pues, las varianzas:

$$Var[\widehat{p}_o] = Var\left[\frac{X}{n}\right] = \frac{Var[X]}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$Var[\widehat{p}_1] = Var\left[\frac{X+1}{n+1}\right] = \frac{Var[X]}{(n+1)^2} = \frac{np(1-p)}{(n+1)^2}$$

Y en ambos casos $\lim_{n\to\infty} Var[\widehat{p}_o] = \lim_{n\to\infty} Var[\widehat{p}_1] = 0$, y como ambos estimadores son asintóticamente insesgados, ambos son consistentes.

d) Veamos qué varianza es menor:

$$p(1-p) = nVar[\widehat{p}_o]; \quad Var[\widehat{p}_1] = \frac{n^2}{(n+1)^2} Var[\widehat{p}_o]$$

Luego $\widehat{p}_1$ tiene menor varianza.

e) El error cuadrático medio se define como $ECM = Var + \text{Sesgo}^2$. Como el sesgo de $\widehat{p}_o$ es nulo, $ECM(\widehat{p}_o) = Var[\widehat{p}_o] = \frac{p(1-p)}{n}$. En el caso de $\widehat{p}_1$ el sesgo es

$$\text{Sesgo}(\widehat{p}_1) = E[\widehat{p}_1] - p = \frac{np+1}{n+1} - p = \frac{1-p}{n+1}$$

El ECM de $\widehat{p}_1$ es

$$ECM(\widehat{p}_1) = Var[\widehat{p}_1] + \text{Sesgo}^2(\widehat{p}_1) = \frac{np(1-p)}{(n+1)^2} + \frac{(1-p)^2}{(n+1)^2} = \frac{np(1-p)+(1-p)^2}{(n+1)^2}$$

Particularizando para $n = 5$ y llamando κ al cociente de ambos errores

$$\kappa = \frac{ECM(\widehat{p}_1)}{ECM(\widehat{p}_o)} = \frac{5p(1-p)+(1-p)^2}{\frac{36p(1-p)}{5}} = \frac{5p+(1-p)}{\frac{36p}{5}} = \frac{20p+5}{36p}$$

Este valor κ puede ser mayor o menor que uno. Valdrá uno si $36p = 20p + 5$, es decir, si $p = 5/16 = 0.3125$. Si $p < 0.3125$, κ será mayor que uno, y si $p > 0.3125$, κ será menor que uno. Como el valor de p depende de cada problema, no puede deducirse "a priori" cuál de los dos estimadores tiene menor error cuadrático.

e) Si bien es cierto que la varianza de $\widehat{p}_1$ es menor que la de $\widehat{p}_o$, no puede asegurarse que su error cuadrático medio sea menor en todos los casos. Por otra parte, $\widehat{p}_o$ es insesgado para cualquier valor de n, siendo además, como es sabido, el estimador de máxima verosimilitud. Por tanto, es preferible utilizar $\widehat{p}_o$.

5.– Se ha decidido modelar el tiempo entre pedidos, R, de un cierto servicio mediante una distribución gamma. Para estimar la media y varianza de la distribución se ha tomado una muestra de extensión $n = 200$ de dichos tiempos. Se estimó la media con $\hat{m}_r = \bar{r} = 35\ dias$ y la varianza con $\hat{\sigma}_r^2 = S_r^2 = 1\ (dia)^2$. Hallar el intervalo de confianza al nivel del 90 % para la media m_r. Especificar y justificar las hipótesis que se realicen. ¿Qué hipótesis deberíamos añadir para poder calcular un intervalo de confianza sobre la varianza? En nuestro caso, ¿es admisible la realización de esta hipótesis?

——————SOLUCIÓN——————

Tomamos una muestra de tamaño $n = 200$ de la que obtenemos $\bar{r} = 35$ días y $S = 1$ día. Dado el tamaño de la muestra se puede considerar que $\bar{r}$ es normal. Entonces, en general,

$$\frac{(\bar{r}-m)}{S^*/\sqrt{n}} \equiv T_{n-1} \quad \text{(de Student)}$$

Pero el tamaño de la muestra es $n = 200$, por lo que podemos considerar que $\sigma \approx S$ y por lo tanto

$$\frac{(\bar{r}-m)}{S/\sqrt{n}} \equiv N(0,1)$$

es decir que el intervalo de confianza sobre la muestra de la población es

$$[\bar{r} - \frac{1.645S}{\sqrt{n}}, \bar{r} + \frac{1.645S}{\sqrt{n}}]$$

que operando resulta

$$\boxed{IC_{90\,\%} = [34.884, 35.116]}$$

Para poder calcular intervalos de confianza sobre la varianza debemos suponer que R es normal. Pero nos dicen que es Gamma. Veamos cuánto valen sus parámetros:

$$\hat{m} = \frac{\hat{k}}{\hat{\lambda}} = \bar{r}, \quad \hat{\sigma^2} = \frac{\hat{k}}{\hat{\lambda}^2} = S^2$$

de donde obtenemos

$$\hat{\lambda} = 35, \quad \hat{k} = 35^2 = 1225$$

Por consiguiente, k parece ser muy grande, por lo que la hipótesis de normalidad sería adecuada y podríamos calcular intervalos de confianza sobre la varianza.

6.– Una empresa que fabrica losas anclables para la construcción de muros de tierra armada considera cambiar el sistema de retención del anclaje para tratar de disminuir los fallos que está obteniendo con el sistema antiguo. Según un estudio realizado por la empresa, de 4000 anclajes realizados con la técnica antigua, 200 fueron defectuosos, y de 3500 realizados con la técnica nueva resultaron 140 con defectos. Determinar un intervalo de confianza del 95 % para la diferencia real de las proporciones de anclajes defectuosos que resultan de utilizar los dos sistemas de retención distintos. A la vista de este resultado, y sin realizar nuevos cálculos, ¿puede deducirse que la técnica nueva es mejor que la antigua?

SOLUCIÓN

Sea p_1 la proporción de elementos defectuosos con el sistema antiguo, y p_2 con el nuevo. Como $n_1 = 4000$ y $n_2 = 3500$ son grandes, podemos aproximar la distribución de los estimadores por

$$\hat{p}_1 \equiv N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right), \quad \hat{p}_2 \equiv N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

Por lo tanto

$$\hat{p}_1 - \hat{p}_2 \equiv N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Como los tamaño de muestra son muy grandes, podemos aproximar la varianza como

$$Var[\hat{p}_1 - \hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \approx \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

Por consiguiente, el intervalo de confianza del 95 % centrado sobre $p_1 - p_2$ es

$$IC_{90\,\%} : \left[(\hat{p}_1 - \hat{p}_2) - u_{0.975}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) + u_{0.975}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

Teniendo en cuenta que $\hat{p}_1 = \frac{200}{4000} = 0.05$, $\hat{p}_2 = \frac{140}{3500} = 0.04$, $n_1 = 4000$, $n_2 = 3500$ y $u_{0.975} = 1.96$ obtenemos

$$\boxed{IC_{90\,\%} : [0.0006349, 0.019365]}$$

El intervalo de confianza no contiene el punto $p_1 - p_2 = 0$. Luego a ese nivel de confianza podemos asegurar que el porcentaje de anclajes defectuosos cambia. Por otra parte, todos los puntos del intervalo de confianza son positivos, es decir, podemos asegurar a ese nivel de confianza que $p_1 > p_2$. Luego el nuevo sistema mejora la calidad de los anclajes.

7.– En un determinado cauce de un río se ha instalado una granja experimental de turbinas de generación eléctrica renovable. Con el fin de aumentar el caudal del río en la granja se realizaron una serie de actuaciones sobre la cuenca hidrográfica.

Para analizar el efecto de las actuaciones, se recogieron los caudales máximos anuales en los 15 años anteriores a las actuaciones:

Antes de las actuaciones: 174 146 206 153 159 169 142 133 61 138 166 200 229 149 131

Y los caudales máximos anuales de los 5 años posteriores a las actuaciones

Después de las actuaciones: 249 308 252 336 402

Se pide:

a) ¿Puede decirse que las actuaciones han provocado el aumento del caudal deseado? ¿A qué nivel p?

b) ¿Podemos considerar que el caudal tras las actuaciones es inferior a 300 m^3/s con un nivel de significación del 10 %?

———————SOLUCIÓN———————

Sea X el caudal previo a las actuaciones e Y el caudal posterior a las actuaciones. De las dos muestras obtenemos:

$n_x = 15$, $\bar{x} = 157.07$, $S_x^2 = 1413.80$, $S_x^{*2} = 1514.78$

$n_y = 5$, $\bar{y} = 309.4$, $S_y^2 = 3245.44$, $S_y^{*2} = 4056.8$

a) En este caso se nos plantean el siguiente contraste de hipótesis

$H_0 : m_x \leq m_y$

$H_1 : m_x > m_y$

Nos encontramos ante un Contraste de Hipótesis sobre la media de dos conjuntos de datos con varianzas desconocidas, también conocido como Test de Welch.

$$\eta = \frac{\left(\frac{S_x^{*2}}{n_x} + \frac{S_y^{*2}}{n_y}\right)^2}{\frac{S_x^{*4}}{n_x^2(n_x-1)} + \frac{S_y^{*4}}{n_y^2(n_y-1)}} = 5.035 \approx 5$$

Nuestro estadístico será en la Hipótesis Primaria:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{S_x^{*2}}{n_x} + \frac{S_y^{*2}}{n_y}}}$$

que tiene una distribución t de Student con 5 grados de libertad. El estadístico muestral vale

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{S_x^{*2}}{n_x} + \frac{S_y^{*2}}{n_y}}} = -5.043$$

En este Contraste la Región crítica se encuentra a la derecha. Si calculamos el nivel p planteamos

$$1 - p = P[t \leq -5.0433]$$

De donde obtenemos que $p = 99.8\,\%$ lo cual es un valor muy elevado y podemos afirmar que el caudal ha aumentado.

b) En este caso se nos plantean el siguiente contraste de hipótesis

$H_0 : m_y \leq 300$

$H_1 : m_y > 300$

En este caso nos encontramos en un Contraste de Hipótesis sobre la media con varianza desconocida.

Nuestro estadístico será en H_o

$$t = \frac{\bar{y} - 300}{S_y^*/\sqrt{n_y}}$$

y tiene una distribución t-Student con $n - 1 = 4$ grados de libertad.

La Región Crítica se encuentra a la derecha y por tanto la Región de Aceptación es $R_A = (-\infty, x_c]$

Donde el valor x_c se puede obtener de la siguiente manera

$$1 - \alpha = P[t \leq x_c | H_o] = 0.90 \Longrightarrow x_c = 1.5332$$

Dada la muestra nuestro estadístico muestral resulta:

$$t = \frac{309.4 - 300}{63.69/\sqrt{5}} = 0.33$$

Como $t \in R_A$ aceptamos la hipótesis primaria con un nivel de confianza del $90\,\%$

El nivel p se puede calcular de la siguiente forma

$$1 - p = P[t \leq 0.33] = 0.6210$$

De donde obtenemos que $p = 37.9\,\%$ y podemos afirmar a los niveles de significación habituales que el caudal es inferior a 325 m^3/s.

8.– En una intersección de carreteras en la que se producen frecuentes accidentes graves se realiza un señalización extensiva, obteniéndose de datos anteriores (16 meses) y posteriores (14 meses) a la señalización los datos que se indican en la tabla. Si se supone que el número de accidentes graves al mes está normalmente distribuido, tanto antes como después de la señalización,

a) ¿Puede decirse, con un nivel de significación del $5\,\%$, que la señalización ha mejorado el funcionamiento de la intersección en lo que a accidentes graves se refiere?

b) ¿Puede decirse, con el mismo nivel de significación, que el número de accidentes no ha variado?

c) ¿Puede determinarse cual ha sido el resultado de la señalización: bueno, indiferente o malo?

ACCIDENTES/MES																
Inicial (X)	2	1	3	2	1	3	3	2	1	3	5	4	3	2	1	3
Tras señalizar (Y)	1	2	1	3	2	3	2	3	1	2	5	3	4	4		

———————————SOLUCIÓN———————————

Diremos que ha mejorado la señalización si el número medio de accidentes en el plazo estudiado disminuye. Llamemos X al número de accidentes al mes antes de la señalización, e Y al número de accidentes al mes después de la señalización.

a) En este caso el contraste de hipótesis será

$H_0 : m_y \leq m_x$

$H_1 : m_y > m_x$

Como solo sabemos que las distribuciones pueden considerarse normales, utilizaremos un test de Welch. Los datos son

$$\bar{x} = 2.4375, \quad \bar{y} = 2.571428571, \quad S_x^2 = 1.24609375, \quad S_y^2 = 1.387755102,$$

$$S_x^{*^2} = 1.329166667, \quad S_y^{*^2} = 1.494505495$$

Utilizaremos el estadístico

$$\theta = \frac{(\bar{y} - \bar{x}) - (m_y - m_x)}{W} \qquad \text{donde} \qquad W = \sqrt{\frac{S_y^{*2}}{n_y} + \frac{S_x^{*2}}{n_x}}$$

Obviamente en H_0

$$\theta = \frac{(\bar{y} - \bar{x})}{W}$$

Sabemos que con las hipótesis expresadas en el enunciado, θ tiene una distribución t de Student con η grados de libertad, siendo η

$$\eta = \frac{\left(\frac{S_x^{*^2}}{n_x} + \frac{S_y^{*^2}}{n_y}\right)^2}{\frac{S_x^{*^4}}{n_x^2(n_x-1)} + \frac{S_y^{*^4}}{n_y^2(n_y-1)}}$$

En nuestro caso, $\eta = 26.96$. Es evidente, por la forma del test, que la región crítica estará a la derecha de la distribución. Dado que en nuestro caso, y para esta muestra, $W = 0.435687169$, resulta nuestro estadístico

$$\theta = \frac{(\bar{y} - \bar{x})}{W} = \frac{(2.571428571 - 2.4375)}{0.435687169} = 0.30739618$$

con lo que el valor p es $p = 0.3826 > 0.05$, es decir que puede admitirse la hipótesis primaria al 5 %de significación.

b) En este caso el contraste de hipótesis será

$H_0 : m_y = m_x$

$H_1 : m_y \neq m_x$

La región crítica está a ambos lados de la distribución de θ. Como los datos son los mismos, el nivel p es ahora $p = 0.7651 > 0.05$, luego también aceptaremos la hipótesis primaria a ese nivel de significación.

c) Para completar todas las situaciones posibles, solo falta el siguiente contraste:

$H_0 : m_y \geq m_x$

$H_1 : m_y < m_x$

En este caso, la región crítica está a la izquierda de la distribución del estadístico θ. Como los datos no cambian, el nivel p es ahora $p = 1 - 0.3826 = 0.6174$. Luego teniendo en cuenta los tres contrastes, la señalización, con los datos que tenemos, no ha mejorado el funcionamiento de la intersección.

En la siguiente figura se muestran esquemáticamente los niveles p para los tres contrastes de hipótesis planteados