Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Segundo Semestre 2020

EYP1113 - Examen
Pauta

Pregunta 1 (I1 - IC) **[0.5 ptos]**

Determine el tamaño de muestra necesario para estimar la proporción de alumnos habilitados (que cumple los requisitos) que tomaran el curso EYP1113 en TAV, con un error de estimación no mayor a 4 puntos porcentuales y con una confianza del 90 %.

Considere como información para su cálculo, que en el TAV 2020, un 15 % de los alumnos que cumplían requisito, tomaron el ramo en ese TAV.

Nota: Como respuesta ingrese la parte entera + 1, por ejemplo [12.3] + 1 = 13; [12.6]+1 = 13

Solución

Para un 90 % de confianza el tamaño muestra sería

$$n = \left(\frac{k_{1-\alpha/2} \cdot \sqrt{\hat{p} \cdot (1-\hat{p})}}{\omega}\right)^2 = \left(\frac{k_{0,95} \cdot \sqrt{0,15 \cdot (1-0,15)}}{0,04}\right)^2 = 215,598 \to 216$$

Nota: Si el alumno ingresó como respuesta 215 o 217, asignar 0.5 ptos. Si el alumno ingreso 423 porqué utilizó criterio de varianza máxima o ingresó 131 porqué utilizó percentil 90 % en vez del 95 %, asignar 0.25 ptos. Cualquier otro caso asignar 0.0 ptos.

Pregunta 2 (I1 - I2 - I3 - Error Tipo II) **[0.5 ptos]**

Suponga que los puntos de golpes (HP) en los Pokemon no legendarios que usted analizó durante la I2 con la base `Pokedex.xlsx`, pueden ser modelados por una variable aleatoria X con distribución Gamma(k, $1/\nu$), donde k es un valor conocido mayor a cero.

A partir de un muestra aleatoria iid $(X_1, X_2, \ldots, X_n)$, se desea contrastar las siguientes hipótesis:

$$\mathrm{H}_0 : \nu = 0{,}1 \quad \text{v/s} \quad \mathrm{H}_a : \nu < 0{,}1$$

Utilizando la distribución aproximada del EMV de ν :

$$\hat{\nu} = \frac{1}{n\,k} \sum_{i=1}^{n} X_i$$

Obtenga la región de rechazo para un nivel de significancia del 5 % y calcule la probabilidad de cometer error tipo II si $\nu = 0{,}09$ y un tamaño muestra igual 20.

Nota: En el caso de necesitar el valor de k, utilice el valor estimado por máxima verosimilitud en base a la variable HP que se encuentra en `Pokedex.xlsx`

Solución

Como $\frac{\partial^2}{\partial\,\nu^2} \ln L = -\frac{n\,k}{\nu^2} \to CCR = \frac{\nu^2}{n\,k}$, entonces por propiedad de EMV se tiene que

$$\hat{\nu} \overset{\text{aprox}}{\sim} \text{Normal}\left(\nu,\, \frac{\nu}{\sqrt{n\,k}}\right) \to Z = \frac{\hat{\nu} - \nu}{\nu/\sqrt{n\,k}} \overset{\text{aprox}}{\sim} \text{Normal}(0,\, 1)$$

Se rechaza H_0 si:

$$Z_0 < k_\alpha \quad \to \quad \hat{\nu} < \nu_0 + k_\alpha \cdot \frac{\nu_0}{\sqrt{n\,k}} \quad \text{(Región de rechazo)}$$

con $\nu_0 = 0{,}10$.

La probabilidad de error tipo II está dada por

$$\begin{aligned}
\beta &= P(\text{No rechazar } \mathrm{H}_0 \,|\, \mathrm{H}_0 \text{ es falsa}) \\
&= 1 - P(\text{Rechazar } \mathrm{H}_0 \,|\, \mathrm{H}_0 \text{ es falsa}) \\
&= 1 - P\left(\hat{\nu} < \nu_0 + k_\alpha \cdot \frac{\nu_0}{\sqrt{n\,k}} \,\middle|\, \nu = 0{,}09\right) \\
&\approx 1 - \Phi\left(\frac{\nu_0 + k_\alpha \cdot \nu_0/\sqrt{n\,k} - \nu}{\nu/\sqrt{n\,k}}\right) \\
&= \texttt{1-pnorm(-0.4413223)} \\
&= \texttt{0.6705102}
\end{aligned}$$

para $n = 20$ y $k = 7.783321$.

Marketing

La base `Marketing.xlsx` contiene el impacto de tres medios publicitarios (Youtube, Facebook y Periódico) en ventas de empresas. Los datos de esta base son las inversiones publicitarias en miles de dólares para cada medio y las ventas en miles de unidades.

Pregunta 3 (I1 - I3 - var.test - t.test) **[0.5 ptos]**

¿Existe evidencia para afirmar que las empresas con menos de 16 mil unidades vendidas presentan una menor inversión en publicidad en Periódicos que las empresas con 16 mil unidades o más vendidas? Asumiendo Normalidad en la inversión publicitario, realice la prueba de hipótesis correspondiente y entregue como respuesta el valor-p.

```
library(rio)
library(dplyr)
Base = import("Marketing.xlsx")
X = filter(Base, Ventas <  16)$Periódico
Y = filter(Base, Ventas >= 16)$Periódico
var.test(x = X, y = Y, alternative = "two.sided")$p.value
## 0.3578452
## H0: mux = muy. vs Ha: mux < muy (varianzas desconocidas e iguales)
t.test(x = X, y = Y, alternative = "less", var.equal = T)$p.value
## 0.02233606
```

Nota: Si el alumno ingresó como respuesta 0.02233606 (0.022) asignar 0.5 ptos. Si el alumno ingresó como respuesta 0.02295803 (0.023) que corresponde al test con varianzas desconocidas y distintas asignar 0.25 ptos. En cualquier otro caso 0.0 ptos.

Pregunta 4 (I3 - ks.test) **[0.5 ptos]**

¿La inversión publicitaria en Periódico distribuye Gamma? Realice una prueba de hipótesis KS utilizando los estimadores de momento y un nivel de significancia del 10 %:

```
library(rio)
library(dplyr)
library(fitdistrplus)
Base = import("Marketing.xlsx")
X = Base$Periódico
par = fitdist(X, distr = "gamma", method = "mme")$estimate
ks.test(X, "pgamma", shape = par[1], rate = par[2])
## Estadístico de prueba ks: 0.07895 (0.0790)
## Valor-p: 0.1652
## Conclusión: SI
```

Nota: valor-p correcto y responder SI a la pregunta 0.50 ptos, valor-p correcto y responder NO a la pregunta 0.25 ptos, cualquier otro caso 0.0 ptos.

Abalon

El consumo indiscriminado ha logrado, literalmente, acabar con "el exquisito loco", pero Chile tiene la oportunidad de revertirlo con un adecuado tratamiento de este nuevo producto el 'abalón rojo".

El objetivo es predecir el momento exacto de "cosecha", la cual puede estar relacionada con la edad del abalón (anillos) que solo puede obtenerse (postmorten) a partir de otras características más fáciles de evaluar (como peso total, largo o diámetro, que permitirá elegir los apropiados).

Pero más que la edad, el interés real es "maximizar" la carne a obtener (dado que se envía congelado o enlatado al sudeste asiático – China, Taiwan, Japón. Es decir, sin concha).

La base `Abalon.xlsx` tiene información recolectada en 4 centros a lo largo de Chile (Caldera, Coquimbo, Puerto Montt y Chiloé). Son 400 datos y es relevante lograr un buen modelo a fin de optimizar el manejo, cultivo y cosecha de este molusco. Las variables son: largo, diametro, alto, peso total, peso del cuerpo, peso de la concha, número de anillos.

A continuación responda las siguientes tres preguntas:

```
library(rio)
library(dplyr)
Base = import("Abalon.xlsx")
```

Pregunta 5 (I1 - I2 - I3 - chisq.test) **[0.5 ptos]**

¿Para el centro de Puerto Montt, el número de anillos distribuye Poisson? Realice un test χ^2 de bondad de ajuste considerando los siguientes intervalos:

$$< 9,\ 9 - 12,\ > 12$$

Considere un nivel de significancia del 10 %.

Solución

```
X = filter(Base, centro == "Puerto Montt")$anillos
lambda = mean(X)
O = c(sum(X < 9), sum(X <= 12) - sum(X < 9), sum(X > 12))
k = length(O)
n = sum(O)
p = diff(ppois(c(-Inf,8,12,Inf), lambda = lambda))
d = chisq.test(x = O, p = p)$statistic
valor.p = 1-pchisq(d, df = k-1-1)

Estadístico de Prueba = 0.176623
valor-p = 0.6742916
¿Distribuye Poisson? SI
```

¿Para el centro de Coquimbo, el número de anillos distribuye Poisson? Realice un test χ^2 de bondad de ajuste considerando los siguientes intervalos:

$$< 9,\ 9 - 14,\ > 14$$

Considere un nivel de significancia del 10 %.

Solución

```
X = filter(Base, centro == "Coquimbo")$anillos
lambda = mean(X)
O = c(sum(X < 9), sum(X <= 14) - sum(X < 9), sum(X > 14))
k = length(O)
n = sum(O)
p = diff(ppois(c(-Inf,8,14,Inf), lambda = lambda))
d = chisq.test(x = O, p = p)$statistic
valor.p = 1-pchisq(d, df = k-1-1)

Estadístico de Prueba = 0.0628175
valor-p = 0.8020969
¿Distribuye Poisson? SI
```

¿Para el centro de Chiloé, el número de anillos distribuye Poisson? Realice un test χ^2 de bondad de ajuste considerando los siguientes intervalos:

$$< 8,\ 8 - 12,\ > 12$$

Considere un nivel de significancia del 10 %.

Solución

```
X = filter(Base, centro == "Chiloé")$anillos
lambda = mean(X)
O = c(sum(X < 8), sum(X <= 12) - sum(X < 8), sum(X > 12))
k = length(O)
n = sum(O)
p = diff(ppois(c(-Inf,7,12,Inf), lambda = lambda))
d = chisq.test(x = O, p = p)$statistic
valor.p = 1-pchisq(d, df = k-1-1)

Estadístico de Prueba = 0.05127183
valor-p = 0.8208648
¿Distribuye Poisson? SI
```

¿Para el centro de Caldera, el número de anillos distribuye Poisson? Realice un test χ^2 de bondad de ajuste considerando los siguientes intervalos:

$$< 8,\ 8 - 13,\ > 13$$

Considere un nivel de significancia del 10 %.

Solución

```
X = filter(Base, centro == "Caldera")$anillos
lambda = mean(X)
O = c(sum(X < 8), sum(X <= 13) - sum(X < 8), sum(X > 13))
k = length(O)
n = sum(O)
p = diff(ppois(c(-Inf,7,13,Inf), lambda = lambda))
d = chisq.test(x = O, p = p)$statistic
valor.p = 1-pchisq(d, df = k-1-1)

Estadístico de Prueba = 1.224256
valor-p = 0.2685269
¿Distribuye Poisson? SI
```

Nota: valor-p correcto y conclusión correcta 0.5 ptos. Solo valor-p correcto 0.25 ptos. En cualquier otro caso 0.0 ptos.

Pregunta 6 (Regresión simple) **[0.5 ptos]**

¿Qué variable explica mejor linealmente el peso cuerpo (pesocu): largo, diametro o alto? Para el mejor modelo complete la información

Solución

```
fit1 = lm(pesocu ~ largo, data = Base)
fit2 = lm(pesocu ~ diametro, data = Base)
fit3 = lm(pesocu ~ alto, data = Base)
summary(fit1)$r.squared
##0.8286976
summary(fit2)$r.squared
##0.8234833
summary(fit3)$r.squared
##0.7841707
fit1$coef[2]*5
##77.34106

Multiple R-squared = 0.8286976

¿En cuántas unidades aumenta o decrece el peso cuerpo, por cada 5 unidades que aumenta el
mejor regresor? 77.34106
```

Nota: Multiple R-squared (o Adjusted R-squared) correcto + $5 \cdot \beta_1$ *correcto asignar 0.5 puntos, solo uno de ellos correcto asignar 0.25 y en cualquier otro caso 0.0 ptos.*

Pregunta 7 (Regresión múltiple) **[0.5 ptos]**

¿El diámetro en presencia de el largo y alto, es significativo para explicar linealmente el peso cuerpo (pesocu) de un Abalon ? Considere un nivel de significancia del 5 %.

Solución

```
fit1 = lm(pesocu ~ largo+alto, data = Base)
fit2 = lm(pesocu ~ largo+diametro+alto, data = Base)
anova(fit1, fit2)
##Analysis of Variance Table
##
##Model 1: pesocu ~ largo + alto
##Model 2: pesocu ~ largo + diametro + alto
##  Res.Df    RSS Df Sum of Sq      F Pr(>F)
##1    397 142674
##2    396 142664  1    10.275 0.0285  0.866

F-statistic: 0.0285
Conclusión: NO
```

Nota: El estadístico F también se puede obtener "a mano" a partir de SCE o R^2*, ver script ensayo 7.*

Si el estadístico F es correcto y conclusión NO asignar 0.5 ptos. Solo F correcto asignar 0.25 ptos, en cualquier otro caso 0l0 ptos.

Pregunta 8 (I1) **[0.5 ptos]**

Una alternativa al modelo Exponencial es una función de densidad definida por

$$f_X(x) = \beta^2\,(1+\beta\,x)^{-(\beta+1)}$$

con $x > 0$ y $\beta > 0$.

El percentil `[p]` % teórico de este modelo si $\beta =$ `[beta]` es :

Solución

Definamos como x_p al percentil $p \times 100\,\%$ Tenemos que

$$F_X\,(x_p) = \int_0^{x_p} \beta^2\,(1+\beta\,x)^{-(\beta+1)}dx = 1 - (1+\beta x_p)^{-\beta} = p$$

Despejando

$$x_p = \frac{(1-p)^{-1/\beta} - 1}{\beta}$$

Pregunta 9 (I1) **[0.5 ptos]**

Una alternativa al modelo Gamma es una función de densidad definida por

$$f_X(x) = 2\,\alpha\,x\,e^{-\alpha\,x^2}$$

con $x > 0$ y $\alpha > 0$.

Si $\alpha =$ `[alpha]`, entonces la moda de este modelo es:

Solución

Tenemos que

$$f'(x) = 2\,\alpha\,e^{-\alpha\,x^2} - 4\,\alpha^2\,x^2\,e^{-\alpha\,x^2} = 0 \to x = \frac{1}{\sqrt{2\,\alpha}}$$

y como

$$f''(1/\sqrt{2\,\alpha}) = \frac{\alpha^2}{\sqrt{2\,\alpha}} \cdot e^{-1/2} \cdot (-12 + 4\,e^{-1/2}) < 0$$

entonces $1/\sqrt{2\,\alpha}$ es la moda.

Pregunta 10 (I3) **[0.5 ptos]**

Considere una muestra aleatoria independiente de tamaño 10 con distribución Normal (μ, σ), con μ desconocido. ¿Cuál sería la probabilidad exacta que el estimador máximo verosímil de σ^2 sea mayor a $\sigma^2/2$?

Solución

Tenemos que

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \to \frac{n\,\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

Se pide

$$P(\hat{\sigma}^2 > \sigma^2/2) = P\left(\frac{n\,\hat{\sigma}^2}{\sigma^2} > \frac{n}{2}\right) = \texttt{1-pchisq(10/2, df = 10-1)} = \texttt{0.8343083}$$

Nota: El alumno que contesto 0.8343083 (0.8343) asignar 0.5 ptos, al alumnos que utilizo una $\chi^2(10)$ y contesto 0.8912 asignar 0.25 ptos. Cualquier otro caso 0.0 ptos.

Pregunta 11 (I3) **[0.5 ptos]**

Sea $X \sim \text{Normal}(1,\, 1/2)$ e $Y \sim \text{Normal}(2,\, 1)$, dos variables aleatorias independientes, determine $\text{Cov}\left(e^X + e^Y, e^X - e^Y\right)$

Solución

Tenemos que

$$e^X \sim \text{Log-Normal}(1,\, 1/2) \qquad \text{y} \qquad e^Y \sim \text{Log-Normal}(2,\, 1)$$

y por independencia

$$\begin{aligned}
\text{Cov}\left(e^X + e^Y, e^X - e^Y\right) &= \text{Var}(e^X) - \text{Var}(e^Y) \\
&= \left(e^{1+0{,}5^2/2}\right)^2\left(e^{0{,}5^2} - 1\right) - \left(e^{2+1^2/2}\right)^2\left(e^{1^2} - 1\right) \\
&= -252{,}3209
\end{aligned}$$

Pregunta 12 (I2) **[0.5 ptos]**

Este fin de semana Santiago vuelve a cuarentena y se estima que las solicitudes de permisos por minuto en la comisaría virtual se comporten como una variable aleatoria Poisson. Si el número esperado de solicitudes cada 10 minutos es de 150, calcule la probabilidad que el tiempo entre 10 solicitudes sea superior a un minuto.

Solución

Tenemos que $T \sim \text{Gamma}(k = 9,\, \nu = 15)$ y se pide

$$P(T > 1) = \texttt{1-pgamma(1, shape = 9, rate = 15)} = \texttt{0.03744649}$$

Nota: Si el alumno alumno responde correctamente 0.03744649 (0.0374) asignar 0.5 ptos y si responde 0.06985366 (0.0699) por utilizar una Gamma(10, 15) asignar 0.25 ptos. En cualquier otro caso asignar 0.0 ptos