Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Segundo Semestre 2018

**Curso : Probabilidad y Estadística**
**Sigla : EYP1113**
**Profesores : Ricardo Aravena C. y Ricardo Olea O.**

# Pauta Examen

## Pregunta 1

Durante el último año, el perfil de los extranjeros que llegan ha cambiado. Específicamente, usted buscando demostrar lo anterior logra del Departamento de Extranjería y Migración la obtención de muestras aleatorias de extranjeros que arriban a Chile, una de Noviembre 2017 y otra de Noviembre 2018. Parte de los registros que para usted son relevantes son: Monto de dinero (en miles de US$) declarado al arribar (dado que vienen con visa de turistas) y número de viajes previos al extranjero (registro del pasaporte). Un breve resumen se presenta a continuación:

| | Nov 2017 | Nov 2018 |
|---|---|---|
| Número de casos | 25.00 | 17.00 |
| Monto promedio en miles US$ | 3.50 | 4.20 |
| Desv. Estándar en miles US$ | 1.20 | 0.75 |
| Promedio de viajes | 1.65 | 2.44 |

(a) **[4.0 Ptos.]** ¿Es posible afirmar que el monto promedio es superior ahora respecto a lo que traían hace un año? Sea explícito en relación al test a utilizar, validación de supuestos y conclusión. Utilice un nivel de significancia del 10 %.

(b) **[2.0 Ptos.]** Asumiendo que el número de viajes previos se puede modelar según un proceso Poisson, ¿existe evidencia que permita afirmar que difieren? Sea explícito respecto a test, supuestos y conclusión basada en valor-p para un nivel de significancia del 5 %.

### Solución

(a) Suponiendo Normalidad de los datos, se pide

$$\mathrm{H}_0 : \mu_{nov17} = \mu_{nov18} \qquad \text{vs} \qquad \mathrm{H}_a : \mu_{nov17} < \mu_{nov18} \qquad \textbf{[0.4 Ptos]}$$

Como

$$\textbf{[0.4 Ptos]} \qquad F_0 = \frac{S^2_{nov17}}{S^2_{no18}} = 2{,}56 \sim F(25{-}1,\, 17{-}1) \to F_{0{,}95}(24,\, 16) = 2{,}24 < 2{,}56 = F \qquad \textbf{[0.4 Ptos]}$$

utilizaremos un test de comparación de medias con varianzas desconocidas y distintas. **[0.4 Ptos]**

Bajo $\mathrm{H}_0$ se tiene que

$$\textbf{[0.4 Ptos]} \qquad T = \frac{\overline{X}_{nov17} - \overline{X}_{nov18}}{\sqrt{\dfrac{S^2_{nov17}}{25} + \dfrac{S^2_{nov18}}{17}}} \sim \text{t-Student}(\nu = 39{,}8) \approx \text{Normal}(0, 1), \qquad \textbf{[0.8 Ptos]}$$

Reemplazando,

**[0.4 Ptos]** $T_0 = -2{,}324 \rightarrow$ valor-p $\approx 0{,}0102 < 0{,}1 = \alpha$ (que es equivalente a que $T < -1{,}28 = k_{0,1}$) **[0.4 Ptos]**

Es decir, se rechaza fuertemente Ho y se puede afirmar que ahora traen, en promedio, un monto superior respecto a lo que traían hace un año. **[0.4 Ptos]**

(b) Se pide

$$\mathrm{H}_0 : \lambda_{nov17} = \lambda_{nov18} \qquad \text{vs} \qquad \mathrm{H}_a : \lambda_{nov17} \neq \lambda_{nov18} \qquad \textbf{[0.3 Ptos]}$$

Bajo $\mathrm{H}_0$ se tiene que

$$\textbf{[0.3 Ptos]} \qquad Z_0 = \frac{\overline{Y}_{nov17} - \overline{Y}_{nov18}}{\sqrt{\frac{\hat{\lambda}}{25} + \frac{\hat{\lambda}}{17}}} \overset{\text{aprox}}{\sim} \text{Normal}(0,1), \qquad \textbf{[0.3 Ptos]}$$

con $\hat{\lambda} = \dfrac{25 \cdot \overline{Y}_{nov17} + 17 \cdot \overline{Y}_{nov18}}{25 + 17}$. **[0.3 Ptos]**

Reemplazando

**[0.3 Ptos]** $Z_0 = -1{,}79 \rightarrow$ valor-p $\approx 2 \cdot (1 - 0{,}9633) = 0{,}0734$ **[0.3 Ptos]**

Por lo tanto, no existe suficiente evidencia para rechazar $\mathrm{H}_0$, es decir, no se puede apoyar que el número de viajes previos entre estos años difieren. **[0.2 Ptos]**

**+ 1 Punto Base**

**Pregunta 2**

Durante el mes de noviembre (30 días) se ha observado un incremento en las interacciones en twitter (diversos motivos: banderazos, protestas o hechos externos como el partido de River vs Boca). Usted interesado en construir un modelo que permita explicar estas variaciones en las interacciones dispone de la siguiente información diaria: Interacciones (Y en unidad de miles), Actividad ($X_1$: 1 si existe actividad previamente programada, 0 en caso contrario), Actividades Externas ($X_2$: número de hechos diarios del exterior), T° max en Santiago ($X_3$), día de la semana ($X_4$: 1 si es laboral, 0 en caso contrario). Un rápido análisis y ajuste de varios modelos de regresión con **R** se presentan a continuación:

```
sum(Y) sum(Y^2)
  1200    48240

------------------------
Modelo   Variables    R2
------------------------
     1          X1 0.334
     2       X1+X2 0.500
     3       X1+X3 0.375
     4    X1+X2+X4 0.583
     5 X1+X2+X3+X4 0.708
------------------------
```

(a) **[3.0 Ptos.]** Reemplace los **XXXXX** por valores en la siguiente salida de regresión lineal simple:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   ------     ------   -----    -----
X1            ------     ------   XXXXX    XXXXX

Residual standard error: XXXXX on XXXXX degrees of freedom
Multiple R-squared:  XXXXX,Adjusted R-squared:  XXXXX
F-statistic:  XXXXX on XXXXX and XXXXX DF,  p-value: <0.001
```

(b) **[3.0 Ptos.]** ¿El aporte de $X_3$ es significativo para explicar linealmente las interacciones $Y$? Use $\alpha = 5\,\%$.

**Solución**

(a) A partir de la $SCT = (48240 - 30 \cdot (1200/30)^2 = 240$ y $R^2 = 0{,}334$, se tiene que

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   ------     ------   -----    -----
X1            ------     ------  +3.747   <0.001

Residual standard error: 2.389 on 28 degrees of freedom
Multiple R-squared: 0.334,Adjusted R-squared: 0.310
F-statistic: 14.04  on 1 and 28 DF,  p-value: <0.001
```

*+0.3 por cada XXXXXX, +0.2 SCT, +0.1 por el signo del estadístico t-Student, ya que si hay actividad programada la interacción debe aumentar* **[3.0 Ptos]**

(b) El aporte de $X_3$ se pude testear comparando:

*Modelos 1 vs Modelo 3*:

**[0.3 Ptos]** $SCE_{\text{mod1}} = SCT(1 - R^2) = 159{,}84$ y $SCE_{\text{mod3}} = 150$ **[0.3 Ptos]**

En presencia de $X_1$, la hipótesis a testear son:

$$\mathrm{H}_0 : \beta_3 = 0 \qquad \text{vs} \qquad \mathrm{H}_a : \beta_3 \neq 0$$

Bajo $\mathrm{H}_0$ se tiene que

**[0.3 Ptos]** $F_0 = \dfrac{(SCE_{\text{mod1}} - SCE_{\text{mod3}})/1}{SCE_{\text{mod3}}/(n-1-1-1)} \sim F(1,27) \rightarrow F_0 = 1{,}7712 < 4{,}21 = F_{0,95}(1,27)$ **[0.3 Ptos]**

Es decir, $X_3$ no aporta en presencia de $X_1$. **[0.3 Ptos]**

*Modelos 4 vs Modelo 5*:

**[0.3 Ptos]** $SCE_{\text{mod4}} = 100{,}08$ y $SCE_{\text{mod5}} = 70{,}08$ **[0.3 Ptos]**

En presencia de $X_1$, $X_2$ y $X_4$, la hipótesis a testear son:

$$\mathrm{H}_0 : \beta_3 = 0 \qquad \text{vs} \qquad \mathrm{H}_a : \beta_3 \neq 0$$

Bajo $\mathrm{H}_0$ se tiene que

**[0.3 Ptos]** $F_0 = \dfrac{(SCE_{\text{mod4}} - SCE_{\text{mod5}})/1}{SCE_{\text{mod5}}/(n-3-1-1)} \sim F(1,25) \rightarrow F_0 = 10{,}70205 > 4{,}24 = F_{0,95}(1,25)$ **[0.3 Ptos]**

Es decir, $X_3$ aporta en presencia de $X_1$, $X2$ y $X4$. **[0.3 Ptos]**

*Nota: Los estadísticos Fisher también se podría calcular utilizando los $R^2$, sin calcular SCT. Asignar todo el puntaje*

**+ 1 Punto Base**

**Pregunta 3**

Para las recientes elecciones de la FEUC, un grupo de estudiantes deseaba predecir el resultado de primera vuelta. Para ello, previo a la elección, entrevistó a 160 alumnos respecto a su intención de voto. El resultado de esta encuesta es la siguiente:

```
1A NAU Surgencia Solidaridad N-Nulo NS/NR
46  32        31          30     15     6
```

En base a estos resultados:

(a) **[2.0 Ptos.]** Obtenga un intervalo de confianza al 90 % para la estimación de la proporción de votos que obtendría la lista 1a. ¿Existía la posibilidad de ganar en primera vuelta?

(b) **[2.0 Ptos.]** Al dividir los apoyos en derecha (1A + Solidaridad) e izquierda (Surgencia + NAU + N-Nulo), ¿existe evidencia suficiente que permita afirmar que el apoyo 1A obtenido en su sector es superior al apoyo obtenido por el NAU en su sector? Utilice un nivel de significancia del 1 %.

(c) **[2.0 Ptos.]** Usando la información previa, obtenga el tamaño de muestra necesario que permita estimar el apoyo al NAU con un error no mayor a 5 % con un 95 % de confianza.

**Solución**

(a) Definamos como $p_{1A}$ a la proporción que apoya 1A en la encuesta.

$$< p_{1A} >_{90\,\%} \in \hat{p}_{1A} \pm k_{0,95} \cdot \sqrt{\frac{\hat{p}_{1A}\,(1-\hat{p}_{1A})}{160}} \qquad \textbf{[0.5 Ptos]}$$

con $\hat{p}_{1A} = \frac{46}{160}$. **[0.5 Ptos]**

Reemplazando

$$< p_{1A} >_{90\,\%} \in (0{,}2286404 \quad - \quad 0{,}3463596) \qquad \textbf{[0.5 Ptos]}$$

Es decir, no había posibilidad que ganaran en 1ra vuelta. **[0.5 Ptos]**

(b) Definamos como $p_{1A}$ y $p_{NAU}$ a la proporción que apoya 1A y NAU en sus respectivos sectores.

Se pide contrastar

$$H_0 : p_{1A} = p_{NAU} \qquad \text{vs} \qquad H_a : p_{1A} > p_{NAU} \qquad \textbf{[0.3 Ptos]}$$

**[0.3 Ptos]** $\hat{p}_{1A} = \dfrac{46}{46+30}$ y $\hat{p}_{NAU} = \dfrac{32}{32+31+15}$ **[0.3 Ptos]**

Bajo $H_0$ se tiene que

$$Z_0 = \frac{\hat{p}_{1A} - \hat{p}_{NAU}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{46+30} + \frac{1}{32+31+15}}} \sim \text{Normal}(0,1), \qquad \textbf{[0.3 Ptos]}$$

con

$$\hat{p} = \frac{(46+30)\cdot \hat{p}_{1A} + (32+31+15)\cdot \hat{p}_{NAU}}{46+30+32+31+15} \qquad \textbf{[0.3 Ptos]}$$

Reemplazando

**[0.2 Ptos]** $Z_0 = 2{,}41997 \rightarrow$ valor-p $\approx 1 - \Phi(2{,}42) = 0{,}0078 < 0{,}01 = \alpha$ **[0.2 Ptos]**

Es decir, existe suficiente evidencia para rechazar $H_0$ y aporya la hipótesis que el apoyo 1A obtenido en su sector es superior al apoyo obtenido por el NAU en su sector. **[0.1 Ptos]**

*Alternativamente cómo $Z_0 > k_{0,99}$ existe suficiente evidencia para rechazar $H_0$ y aporya la hipótesis que el apoyo 1A obtenido en su sector es superior al apoyo obtenido por el NAU en su sector. Asignar puntaje completo*

(c) Tenemos como información previa $\hat{p}_{\text{NAU}} = \dfrac{32}{160} = 0{,}2$. **[0.5 Ptos]**

Por tanto

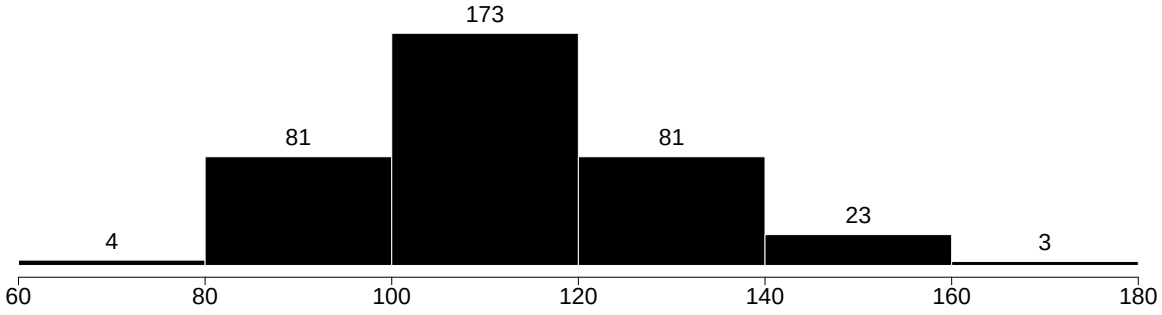
**[1.0 Ptos]** $$n = \left( \frac{1{,}96 \times \sqrt{0{,}2 \times 0{,}8}}{0{,}05} \right)^2 = 245{,}8$$ **[0.3 Ptos]**

Se necesitan al menos 246 casos. **[0.2 Ptos]**

**+ 1 Punto Base**

**Pregunta 4**

Existen varias aplicaciones que miden la velocidad de conexión a internet que proveen los distintas empresas del rubro. Un usuario realiza contante mediciones con *Speedtest* para ir verificando si lo que ofrece la empresa se cumple. La siguiente figura muestra el comportamiento de las mediciones diarias, en Mbps, realizadas el último año por el cliente:




```
quantile(x, c(0.25,0.50))
     25%      50%
100.5765 111.3941
```

Realice ahora un test $\chi^2$ de bondad de ajuste Log-Normal y Log-Logístico para los intervalos de la figura. Si fuese necesario colapse (una) intervalos. ¿Cuál de los dos modelos ajusta mejor según este test? (En sus cálculos redondee al 4to decimal y si lo requiere un nivel de significancia del 5 %)

**Solución**

*Test Log-Normal*

$$\mathrm{H}_0 : X \sim \text{Log-Normal}(\lambda,\, \zeta) \qquad \text{vs} \qquad \mathrm{H}_a : X \not\sim \text{Log-Normal}(\lambda,\, \zeta) \qquad \textbf{[0.4 Ptos]}$$

**[0.4 Ptos]** $\hat{\lambda} = \ln(111{,}3941) = 4{,}713074$ y $\hat{\zeta} = -\dfrac{\ln(100{,}5765) - \hat{\lambda}}{0{,}675} = 0{,}1513418$ **[0.4 Ptos]**

| Intervalo | Observado | Probabilidad | Esperado |
|---|---|---|---|
| 1 | 4 | 0.0144 | 5.2560 |
| 2 | 81 | 0.2236 | 81.6140 |
| 3 | 173 | 0.4506 | 164.4690 |
| 4 | 81 | 0.2460 | 89.7900 |
| 5 | 23 | 0.0571 | 20.8415 |
| 6 | 3 | 0.0084 | 3.0660 |

Como el 6to valor esperado es menor a 5, se procede a colapsar los intervalos 5 y 6

| Intervalo | Observado | Probabilidad | Esperado | (O-E)^2/E |
|---|---|---|---|---|
| 1 | 4 | 0.0144 | 5.2560 | 0.300140030 |
| 2 | 81 | 0.2236 | 81.6140 | 0.004619257 |
| 3 | 173 | 0.4506 | 164.4690 | 0.442502605 |
| 4 | 81 | 0.2460 | 89.7900 | 0.860497828 |
| 5 | 26 | 0.0654 | 23.9075 | 0.183145718 |
| TOTAL | 365 | 1.0000 | 365.0365 | 1.790905438 |

**[0.8 Ptos]**

Como $X^2 = 1{,}790905438 \sim \chi^2(5-1-2)$ **[0.4 Ptos]** $\rightarrow 40\,\% <$ valor-p $< 60\,\%$. **[0.4 Ptos]**

*Test Log-Logística*

$$\mathrm{H}_0 : X \sim \text{Log-Logística}(\mu,\, \sigma) \qquad \text{vs} \qquad \mathrm{H}_a : X \not\sim \text{Log-Logística}(\mu,\, \sigma) \qquad \textbf{[0.4 Ptos]}$$

**[0.4 Ptos]** $\hat{\mu} = \ln(111{,}3941) = 4{,}713074$ y $\hat{\sigma} = \dfrac{\ln(100{,}5765) - \hat{\mu}}{\ln(0{,}25/0{,}75)} = 0{,}09298615$ **[0.4 Ptos]**

| Intervalo | Observado | Probabilidad | Esperado | (O-E)^2/E |
|---|---|---|---|---|
| 1 | 4 | 0.0276 | 10.0740 | 3.6622470 |
| 2 | 81 | 0.2109 | 76.9785 | 0.2100906 |
| 3 | 173 | 0.4515 | 164.7975 | 0.4082647 |
| 4 | 81 | 0.2311 | 84.3515 | 0.1331636 |
| 5 | 23 | 0.0589 | 21.4985 | 0.1048679 |
| 6 | 3 | 0.0200 | 7.3000 | 2.5328767 |
| TOTAL | 365 | 1.0000 | 365.0000 | 7.0515106 |

**[0.8 Ptos]**

Como $X^2 = 7{,}0515106 \sim \chi^2(6-1-2)$ **[0.4 Ptos]** $\rightarrow 5\,\% <$ valor-p $< 10\,\%$. **[0.4 Ptos]**

A un $5\,\%$ de significancia, ambos modelos ajustan, pero el mejor ajuste se logra con la distribución Log-Normal, ya que tiene valor-p más grande. **[0.4 Ptos]**

**+ 1 Punto Base**