Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Temporada Académica de Verano 2018

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

Pauta Examen

Problema 1

Un alumno de ingeniería que espera titularse este año decide hacer una encuesta entre los titulados del año pasado para ver cómo está el mercado laboral. Para ello envía 240 encuestas, de las cuales vuelven poco más de un centenar y que clasifica según sexo del entrevistado. Interesa si trabajan y cuál es su ingreso. Cabe mencionar que las razones para no trabajar son variadas y destacan: continuidad de estudio, viajes (sabático), familia, entre otras.

A continuación se presenta el resultados de las encuestas recibidas:

	Mujer	Hombre
Número de encuestas recibidas	32	74
Número de entrevistados trabajando	20	24

Al considerar solo los que trabajan:

	Mujer	Hombre
Ingreso promedio (en miles de $)	1275	1420
Desviación estándar (en miles de $)	270	180

Suponga que el ingreso se comporta de acuerdo a una distribución normal.

(a) ¿Existe evidencia que permita afirmar que los que trabajan son minoría entre los titulados del año pasado? Determine valor-p y concluya para un nivel de significancia del 5 %.

(b) ¿Existe evidencia que permita validar la afirmación: "en el 1er año de titulación se puede observar que los hombres reciben mejores salarios que las mujeres"? Asuma varianzas iguales, obtenga valor-p y concluya para un nivel de significancia del 1 %.

(c) Basado en la información previa, para un estudio posterior se desea estimar el salario medio de las mujeres con un error no mayor a cincuenta mil, con un 90 % de confianza, ¿cuántas encuestas deberían volver?

Solución

(a) Sean X_1, X_2, ..., X_n variables aleatorias iid Bernoulli(p), donde p es la proporción de titulados del año pasado que contestaron la encuesta y que reportaron estar trabajando.

Se pide testear

$$\mathrm{H}_0 : p = p_0 \quad \text{vs} \quad \mathrm{H}_a : p < p_0 \quad \textbf{[0.5 Ptos]}$$

Si H_0 es correcta, entones

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} \overset{\text{aprox}}{\sim} \text{Normal}(0,\,1) \quad \textbf{[0.5 Ptos]}$$

con $n = 106$, $\hat{p} = \frac{44}{106}$ y $p_0 = \frac{1}{2}$.

Reemplazando

$$Z_0 = -1{,}748315 \to \text{valor-p} = \Phi(Z_0) \approx \Phi(-1{,}75) = 1 - \Phi(1{,}75) = 0{,}0401 < 0{,}05 = \alpha \quad \textbf{[0.5 Ptos]}$$

Por lo tanto, existe suficiente evidencia para rechazar H_0 y apoyar así la hipótesis que los que trabajan son minoría. **[0.5 Ptos]**

(b) Se pide comparación de medias, con sigmas desconocido pero iguales muestra Normal.

$$H_0 : \mu_H = \mu_M \quad \text{vs} \quad H_a : \mu_H > \mu_M \quad \textbf{[0.5 Ptos]}$$

Si H_0 es correcta, entones

$$T_0 = \frac{\overline{X}_H - \overline{X}_M}{S_p\sqrt{\frac{1}{20} + \frac{1}{24}}} \overset{\text{aprox}}{\sim} \text{t-Student}(20 + 24 - 2) \quad \textbf{[0.5 Ptos]}$$

con $\overline{X}_H = 1420$, $\overline{X}_M = 1275$, $S_H = 180$, $S_M = 270$ y $S_p^2 = \dfrac{(20-1)\,S_H^2 + (24-1)\,S_M^2}{20 + 24 - 2} = 50721{,}4$.

Reemplazando

$$T_0 = 2{,}126506 \approx 2{,}13 \to \text{valor-p} \approx 1 - \Phi(2{,}13) = 1 - 0{,}9834 = 2\cdot(1 - 0{,}9826) = 0{,}0166\,(1{,}66\,\%) \quad \textbf{[0.5 Ptos]}$$

Por tanto, no existe suficiente evidencia al 1 % de significancia para rechazar H_0, es decir, no es válida la afirmación en el 1er año de titulación se puede observar que los hombres reciben mejores salarios que las mujeres". **[0.5 Ptos]**

(c) A partir de la información podemos asumir que $\sigma_M = 270$. **[0.5 Ptos]**

Al 90 % de confianza el tamaño muestral está dado por

$$n = \left(\frac{1{,}645 \times 270}{50}\right)^2 = 78{,}9, \quad \textbf{[0.5 Ptos]}$$

pero como este tamaño muestra corresponde a 30.18868 % de las encuestas recibidas **[0.5 Ptos]**, entonces necesitamos tener en total de vuelta igual a $78{,}9/0{,}3018868 = 261{,}3562 \to 262$. **[0.5 Ptos]**

Notar que el número de encuestas enviadas deberían ser $261{,}3562/0{,}4416667 = 591{,}7498 \to 592$ encuestas, ya que solo llegan de vuelta el 44.16667 %.

+ 1 Punto Base

Problema 2

Producto de diversas faltas cometidas en la UC, a fines del 2016 se oficializó la firma del código de honor. Un estudioso del área indica que el comportamiento de las faltas denunciadas semanalmente por escuelas, post firma, se comportan variables aleatorias Poisson. Una muestra 36 semanas, entrega que en promedio se hacen dos denuncias semanales. ¿Esta información, es suficiente al 5 % de significancia para afirmar que, en una semana cualquiera, la probabilidad que se denuncien casos sea mayor a 0,7?

Solución

Tenemos X_1, X_2, ..., X_n variables aleatorias iid Poisson(λ), lo que implica que

$$P(X > 0) = 1 - P(X = 0) = 1 - e^{-\lambda} = g(\lambda) \quad \textbf{[1.0 Ptos]}$$

Alternativa 1: Se pide testear

$$\mathrm{H}_0 : g(\lambda) = g(\lambda_0) \quad \text{vs} \quad \mathrm{H}_a : g(\lambda) > g(\lambda_0) \quad \textbf{[0.5 Ptos]}$$

con $g(\lambda_0) = 0{,}7 \to \lambda_0 = -\ln(0{,}3)$. **[0.5 Ptos]**

Como el estimador máximo verosímil de λ es:

$$\hat{\lambda} \overset{\text{aprox}}{\sim} \text{Normal}\left(\lambda, \sqrt{\frac{\lambda}{n}}\right) \quad \textbf{[0.5 Ptos]}$$

entonces el estimador máximo verosímil de $g(\lambda)$ está dado por:

$$\widehat{g(\lambda)} = g(\hat{\lambda}) \overset{\text{aprox}}{\sim} \text{Normal}\left(g(\lambda), \sqrt{[g'(\lambda)]^2 \cdot \frac{\lambda}{n}}\right) \quad \textbf{[0.5 Ptos]}$$

$$\overset{\text{aprox}}{\sim} \text{Normal}\left(g(\lambda), \sqrt{\frac{\lambda\, e^{-2\,\lambda}}{n}}\right) \quad \textbf{[1.0 Ptos]}$$

Si H_0 es correcta, entones

$$Z_0 = \frac{g(\hat{\lambda}) - g(\lambda_0)}{\sqrt{\dfrac{\lambda_0\, e^{-2\,\lambda_0}}{n}}} \overset{\text{aprox}}{\sim} \text{Normal}(0,\, 1) \quad \textbf{[0.5 Ptos]}$$

Evaluando en $\hat{\lambda} = \overline{X}_n = 2$ y $n = 36$, se tiene que $Z_0 = 3{,}001388$. **[0.5 Ptos]**

Opción 1: Como $Z_0 = 3{,}001388 > 1{,}645 = k_{0{,}95}$, existe suficiente evidencia para rechazar H_0 y afirmar que la probabilidad de denuncias es mayor a 0.7. **[1.0 Ptos]**

Opción 2: Como valor-p$= 1 - \Phi(Z_0) = 1 - \Phi(3{,}001388) \approx 1 - \Phi(3{,}001388) = 0{,}0013 < 0{,}05 = \alpha$, existe suficiente evidencia para rechazar H_0 y afirmar que la probabilidad de denuncias sea mayor a 0.7. **[1.0 Ptos]**

Alternativa 2: Se pide testear

$$\mathrm{H}_0 : g(\lambda) = g(\lambda_0) \quad \text{vs} \quad \mathrm{H}_a : g(\lambda) > g(\lambda_0) \quad \textbf{[0.5 Ptos]}$$

con $g(\lambda_0) = 0{,}7 \to \lambda_0 = -\ln(0{,}3)$. **[0.5 Ptos]**

Esto es equivalente a testear

$$\mathrm{H}_0 : \lambda = -\ln(0{,}3) \quad \text{vs} \quad \mathrm{H}_a : \lambda > -\ln(0{,}3) \quad \textbf{[1.5 Ptos]}$$

El estimador máximo verosímil de λ es:

$$\hat{\lambda} \overset{\text{aprox}}{\sim} \text{Normal}\left(\lambda, \sqrt{\frac{\lambda}{n}}\right) \quad \textbf{[0.5 Ptos]}$$

Si H_0 es correcta, entones

$$Z_0 = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \overset{\text{aprox}}{\sim} \text{Normal}(0, 1) \quad \textbf{[0.5 Ptos]}$$

Evaluando en $\hat{\lambda} = \overline{X}_n = 2$ y $n = 36$, se tiene que $Z_0 = 4{,}352821$. **[0.5 Ptos]**

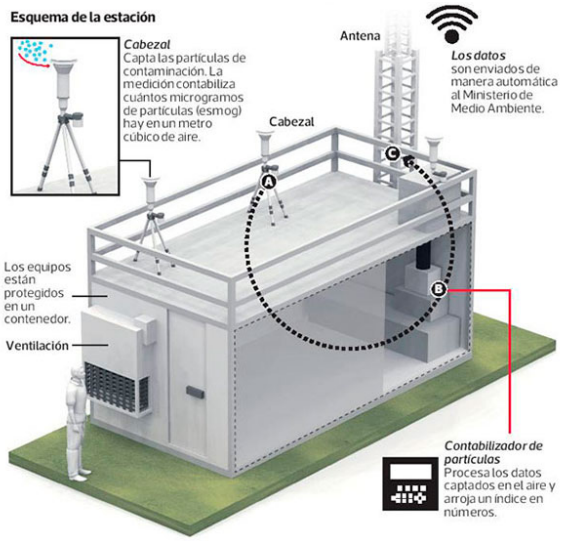
Opción 1: Como $Z_0 = 4{,}352821 > 1{,}645 = k_{0,95}$, existe suficiente evidencia para rechazar H_0 y afirmar que la probabilidad de denuncias es mayor a 0.7. **[1.0 Ptos]**

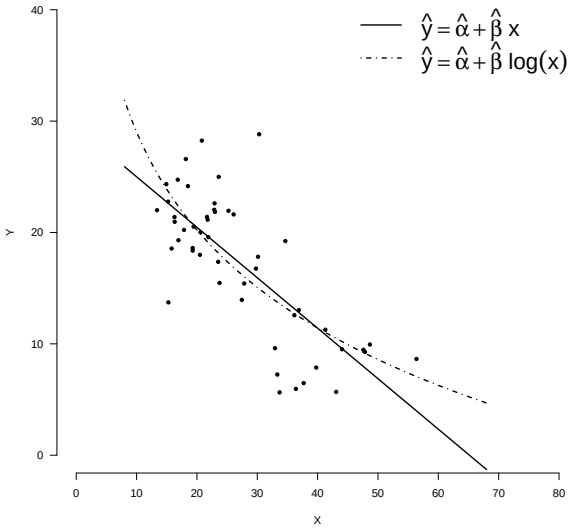
Opción 2: Como valor-p$= 1 - \Phi(Z_0) = 1 - \Phi(4{,}352821) \approx 0 < 0{,}05 = \alpha$, existe suficiente evidencia para rechazar H_0 y afirmar que la probabilidad de denuncias sea mayor a 0.7. **[1.0 Ptos]**

+ 1 Punto Base

Problema 3

Desde el 01 de abril de 1997, la estación meteorológica de Cerrillos, ver figura izquierda, se encuentra realizando constantemente mediciones con el objetivo de contribuir de manera activa a la protección de la salud de la población, promoviendo la difusión de información oportuna y confiable acerca de la calidad del aire. La figura a la derecha muestra la relación entre los niveles promedios mensuales Y de ozono (O_3) y X de partículas en suspensión de menos de 2,5 micras ($PM_{2,5}$) en esta zona.





Fuente: http://sinca.mma.gob.cl

Se ajustaron dos modelos de regresión simple que explican el nivel de ozono en función de las partículas en suspensión, pero la información de las salidas de **R** es parcial.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.53207    XXXXXXX  17.683  < 2e-16
X           XXXXXXXX    0.05697  XXXXXX 2.57e-10

Residual standard error: 4.248 on XX degrees of freedom
Multiple R-squared:  0.5687,Adjusted R-squared:  XXXXXX
F-statistic:  63.3 on 1 and XX DF,  p-value: XXXXXXX

 mean(Y) mean(Y^2)
17.13416  333.7507
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.259      5.440  10.709 2.56e-14
log(X)       XXXXXXX      1.669  -7.609 XXXXXXXX

Residual standard error: 4.355 on 48 degrees of freedom
Multiple R-squared:  XXXXX,Adjusted R-squared:  0.5373
F-statistic: XXXXX on 1 and 48 DF,  p-value: 8.64e-10
```

Complete la información faltante y comente.

Solución

Tenemos que

$$S^2_{Y\,|\,X} = \frac{SCE}{n-2} = (\text{Residual standard error})^2 \quad \text{y} \quad S^2_Y = \frac{n}{n-1}\left(\overline{Y^2} - (\overline{Y})^2\right)$$

Los coeficientes de determinación están dados por

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{(n-1)}{(n-2)} \cdot \frac{S^2_{Y\,|\,X}}{S^2_Y} \quad \text{y} \quad r^2 = 1 - \frac{S^2_{Y\,|\,X}}{S^2_Y}$$

Los estadísticos t y F por:

$$T_{\hat{\alpha}} = \frac{\hat{\alpha}}{\sqrt{\widehat{\mathrm{Var}(\hat{\alpha})}}}, \quad T_{\hat{\beta}} = \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}}, \quad F = T_{\hat{\beta}}^2 = \frac{SCR/1}{SCE/(n-2)}$$

con $SCT = SCR + SCE = n \cdot (\overline{Y^2} - (\overline{Y})^2)$.

Luego, la salida de **R** queda como sigue:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.53207    1.67008  17.683  < 2e-16 ***
X           -0.45328    0.05697  -7.956 2.57e-10 ***

Residual standard error: 4.248 on 48 degrees of freedom
Multiple R-squared:  0.5687,Adjusted R-squared:  0.5597
F-statistic:  63.3 on 1 and 48 DF,  p-value: 2.57e-10
```

[2.0 Ptos]

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.259      5.440  10.709 2.56e-14 ***
log(X)       -12.696      1.669  -7.609 8.64e-10 ***

Residual standard error: 4.355 on 48 degrees of freedom
Multiple R-squared:  0.5467,Adjusted R-squared:  0.5373
F-statistic: 57.89 on 1 and 48 DF,  p-value: 8.64e-10
```

[2.0 Ptos]

Notemos que el mejor ajuste se logra en el modelo Y vs X, ya que el R^2 y r^2 son mayores **[1.0 Ptos]**, valores p más pequeño en el estadístico t **[0.5 Ptos]** y F. **[0.5 Ptos]**

+ 1 Punto Base

Problema 4

El mejor modelo de regresión simple del problema 3 presentan errores (residuos) con mediana igual $-0{,}048$ y desviación estándar igual a 4,2. Para la siguiente distribución de frecuencia:

Intervalo	[-10, -5)	[-5, 0)	[0, +5)	[+5, +10)	[+10, +15]
Frecuencia	5	21	19	4	1

¿Qué modelo ajusta mejor: Normal o Logístico? Si usted encuentra recomendable colapsar la tabla, hágalo.

Solución

Test de Bondad de Ajuste Normal:

$$\mathrm{H}_0 : X \sim \text{Normal}(\mu,\, \sigma) \quad \text{vs} \quad \mathrm{H}_a : X \not\sim \text{Normal}(\mu,\, \sigma)$$

Los parámetros estimados son:

[0.2 Ptos] $\hat{\mu} = -0{,}048$ y $\hat{\sigma} = 4{,}2$ **[0.2 Ptos]**

Las probabilidades estimadas para cada intervalo se obtienen restando funciones de probabilidad acumulada:

$$P(a < X \leq b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right) \quad \textbf{[0.4 Ptos]}$$

Luego

```
-----------------------------------------
Intervalo Observado Probabilidad Esperado
-----------------------------------------
[-10, -5)         5       0.1190    5.950
[ -5,  0)        21       0.3850   19.250
[  0, +5)        19       0.3809   19.045
[ +5,+10)         4       0.1067    5.335
[+10,+15]         1       0.0084    0.420
-----------------------------------------
```

[0.4 Ptos]

Notemos que el 5to intervalo presenta un valor esperado inferior a 5, por esta razón se colapsan el 4to y 5to intervalo.

```
-----------------------------------------------------
          Observado Probabilidad Esperado    (O-E)^2/E
-----------------------------------------------------
[-10, -5)         5       0.1190    5.950 0.1516806723
[ -5,  0)        21       0.3850   19.250 0.1590909091
[  0, +5)        19       0.3809   19.045 0.0001063271
[ +5,+15)         5       0.1151    5.755 0.0990486533
-----------------------------------------------------
Total            50       1.0000   50.000 0.4099265618
-----------------------------------------------------
```

[0.5 Ptos]

El estadístico de prueba $X^2 = 0{,}4099265618 \sim \chi^2(4-1-2)$ **[0.4 Ptos]**, es decir,

$$40\,\% < \text{valor-p} < 60\,\% \quad \textbf{[0.4 Ptos]}$$

Test de Bondad de Ajuste Logístico:

$$\mathrm{H}_0 : X \sim \text{Logistica}(\mu,\, \sigma) \quad \text{vs} \quad \mathrm{H}_a : X \not\sim \text{Logística}(\mu,\, \sigma)$$

Los parámetros estimados son:

[0.2 Ptos] $\hat{\mu} = -0{,}048$ y $\hat{\sigma} = \dfrac{4{,}2 \cdot \sqrt{3}}{\pi} = 2{,}316756$ **[0.2 Ptos]**

Las probabilidades estimadas para cada intervalo se obtienen restando funciones de probabilidad acumulada:

$$P(a < X \leq b) = F_X(b) - F_X(a) = \frac{\exp\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)}{1+\exp\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)} - \frac{\exp\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{1+\exp\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)} \quad \textbf{[0.4 Ptos]}$$

Luego

	Observado	Probabilidad	Esperado
[-10, -5)	5	0.10550777	5.2753883
[-5, 0)	21	0.39967171	19.9835853
[0, +5)	19	0.39316000	19.6580000
[+5,+10)	4	0.08875486	4.4377431
[+10,+15]	1	0.01290567	0.6452833

[0.4 Ptos]

Notemos que el 5to intervalo presenta un valor esperado inferior a 5, por esta razón se colapsan el 4to y 5to intervalo.

	Observado	Probabilidad	Esperado	(O-E)^2/E
[-10, -5)	5	0.1055078	5.275388	0.014375947
[-5, 0)	21	0.3996717	19.983585	0.051697372
[0, +5)	19	0.3931600	19.658000	0.022024825
[+5,+15)	5	0.1016605	5.083026	0.001356158
Total	50	1.0000000	50.000000	0.089454302

[0.5 Ptos]

El estadístico de prueba $X^2 = 0{,}089454302 \sim \chi^2(4-1-2)$ **[0.4 Ptos]**, es decir,

$$70\,\% < \text{valor-p} < 80\,\% \quad \textbf{[0.4 Ptos]}$$

Por lo tanto, aunque ambos modelos ajustan bien, el modelo Logístico por presentar un valor-p es mayor logra un mejor ajuste. **[1.0 Ptos]**

Nota: Si el alumno concluye sin obtener los valores-p, pero indica que lo hace por que ambos estadísticos distribuyen igual y el modelo que ajusta mejor es el que tiene un estadístico más cercano a cero, asignar los [1.8 Ptos]

+ 1 Punto Base