Proyecto GOING
EYP1113 - Probabilidades y Estadística
2do Semestre, 2022
Michael Ramón (maramon@uc.cl)

linktr.ee/Going

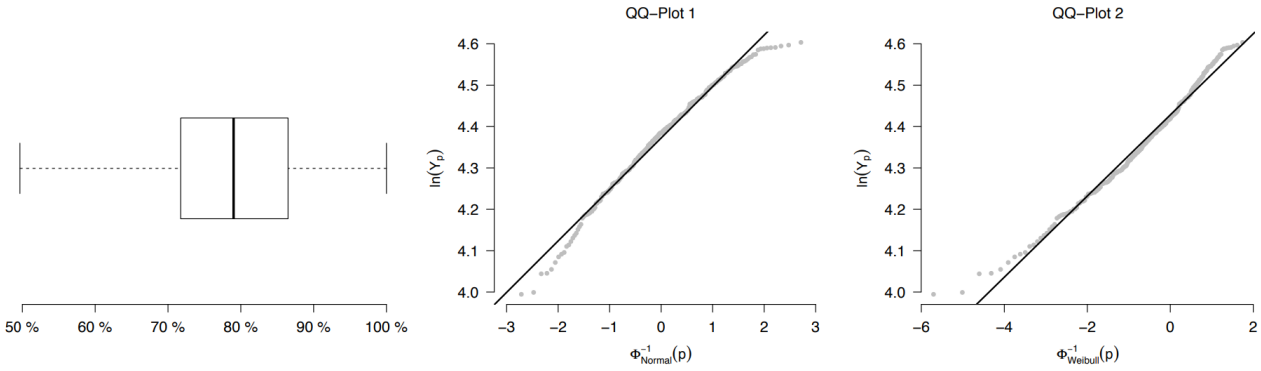
Ayudantía Masiva

Repaso Examen

EMV, Bondad de ajuste, Regresión Lineal, Test de Hipótesis

Pregunta 1 *QQ-Plot y test* χ^2

Un análisis de los % de humedad en la Región Metropolitana durante un cierto mes en los últimos 5 años que se registraron en distintas estaciones de monitoreo se presenta a continuación:



Mediante un ajuste por medio de una recta se obtiene la siguiente información sobre los gráficos de probabilidad:

```
             Intercepto   Pendiente
QQ-Plot 1:     4.372228  0.12435090
QQ-Plot 2:     4.428114  0.09805834
```

¿Entre los dos modelos ajustados por gráficos de probabilidad, cuál ajusta mejor? Realice una prueba de bondad de ajuste χ^2 a la siguiente tabla de frecuencia:

	<60 %	[60 %-70 %)	[70 %-80 %)	[80 %-90 %)	>=90 %
Frecuencia	8	40	98	107	47

Si fuese necesario colapsar (unir) intervalos, hágalo. Además, se le brinda la información de las probabilidades teóricas mediante la siguiente entrada de R:

```
> diff(plnorm(c(-Inf,60,70,80,90,Inf),meanlog=XXXX, sdlog=XXXX))
[1]  0.01271945  0.14714161  0.37154239  0.31614629  0.15245025

> diff(pweibull(c(-Inf,60,70,80,90,Inf),shape=XXXX, scale=XXXX))
[1]  0.03270061 0.11526849  0.31677093  0.41001393  0.12524603
```

Determine los valores restantes de los parámetros para cada distribución.

Solución:

Lo primero y más fácil de hacer es determinar la información faltante de los parámetros de las distribuciones, para el caso de la distribución Log-Normal, la relación lineal entre los percentiles teóricos y empíricos es la siguiente:

$$\ln(Y_p) = \lambda + \zeta \cdot \Phi^{-1}_{\text{Normal}}(p)$$

Esta recta es el QQ-Plot 1, de enunciado se tiene el valor del intercepto y pendiente, por lo que:

$$\texttt{meanlog} = \hat{\lambda} = 4.372228$$

$$\texttt{sdlog} = \hat{\zeta} = 0.12435090$$

En el caso de la distribución Weibull, la relación entre los percentiles teóricos y empíricos es:

$$\ln(Y_p) = \ln(\eta) + \frac{1}{\beta} \cdot \Phi^{-1}_{\text{Weibull}}(p)$$

Esta recta es el QQ-Plot 2 y se tiene el valor del intercepto y pendiente:

$$\ln(\eta) = 4.428114$$

$$\frac{1}{\beta} = 0.09805834$$

Solo falta despejar cada parámetro:

$$\texttt{shape} = \hat{\beta} = 10.19801$$

$$\texttt{scale} = \hat{\eta} = 83.77327$$

Ahora se puede hacer el test de bondad de ajuste pedido, primero se utilizará la distribución Log-Normal bajo las siguientes hipótesis:

$$H_0 : X \sim \text{Log-Normal} \qquad \text{vs} \qquad H_a : X \not\sim \text{Log-Normal}$$

De enunciado se tienen las probabilidades teóricas, por lo que se construye la tabla:

Intervalo	**Observado**	**Prob. teo. (p_i)**	**Esperado (np_i)**	**X^2**
$< 60\,\%$	8	0.01271945	3.815836	4.58804385
$[60\,\% - 70\,\%)$	40	0.14714161	44.142484	0.38874507
$[70\,\% - 80\,\%)$	98	0.37154239	111.462717	1.62605706
$[80\,\% - 90\,\%)$	107	0.31614629	94.843888	1.55804527
$\geq 90\,\%$	47	0.15245025	45.735076	0.03498483
Total	$n = 300$	$p_T = 1$	$n = 300$	$X^2 = 8.195876$

como se estimó 2 parámetros de la distribución, entonces:

$$X^2 = 8.195876 \sim \chi^2(5 - 1 - 2)$$

Se recomienda en este test que la cantidad de intervalos sea mayor a 5 o que el valor esperado sea mayor a 5, como se tiene un intervalo con valor esperado menor a 5 se realiza un colapso, en este caso, entre el intervalo 1 y 2, por lo que se tiene lo siguiente:

Intervalo	**Observado**	**Prob. teo. (p_i)**	**Esperado (np_i)**	**X^2**
$< 70\,\%$	48	0.1598611	47.95832	3.622301×10^{-5}
$[70\,\% - 80\,\%)$	98	0.37154239	111.462717	1.62605706
$[80\,\% - 90\,\%)$	107	0.31614629	94.843888	1.55804527
$\geq 90\,\%$	47	0.15245025	45.735076	0.03498483
Total	$n = 300$	$p_T = 1$	$n = 300$	$X^2 = 3.219123$

donde:

$$X^2 = 3.219123 \sim \chi^2(4 - 1 - 2)$$

Ahora se realiza el test para la distribución Weibull, para esto se proponen las siguientes dos hipótesis:

$$H_0 : X \sim \text{Weibull} \qquad \text{vs} \qquad H_a : X \not\sim \text{Weibull}$$

De enunciado se tienen las probabilidades teóricas, por lo que se construye la tabla:

Intervalo	**Observado**	**Prob. teo. (p_i)**	**Esperado (np_i)**	**X^2**
$< 60\,\%$	8	0.03270061	9.810184	0.33401678
$[60\,\% - 70\,\%)$	40	0.11526849	34.580548	0.84933457
$[70\,\% - 80\,\%)$	98	0.31677093	95.031280	0.09274104
$[80\,\% - 90\,\%)$	107	0.41001393	123.004180	2.08231757
$\geq 90\,\%$	47	0.12524603	37.573808	2.36476141
Total	$n = 300$	$p_T = 1$	$n = 300$	$X^2 = 5.723171$

Como el número de intervalos es mayor o igual a 5 y los valores esperados son mayores o iguales a 5, entonces no se debe colapsar intervalos. Como se estimó dos parámetros, entonces:

$$X^2 = 5.723171 \sim \chi^2(5 - 1 - 2)$$

Hay dos formas de concluir con este test de bondad de ajuste, una mediante comparar los estadísticos y otra con el valor-p:

- ***Comparación de estadísticos de prueba***: Si se está haciendo un test de bondad de ajuste entre dos o más distribuciones para saber cual ajusta mejor los datos, entonces la distribución que tenga el estadístico con el menor valor posible entre los demás estadísticos será la que mejor ajusta a los datos. EN este caso se tienen 3 estadísticos:
  - Considerando el colapso a 4 intervalos para el modelo Log-Normal, entonces se tiene que:

    $$3.219123 < 5.723171$$

    El modelo Log-Normal ajusta mejor que el modelo Weibull
  - Considerando los 5 intervalos originales para el modelo Log-Normal, entonces se tiene que:

    $$8.195876 > 5.723171$$

    El modelo Weibull ajusta mejor que el modelo Log-Normal
- ***Comparando Valores-p***: Mediante la tabla chi cuadrado es posible obtener intervalos para los valores-p correspondientes, esto son:

  $$1\,\% < \text{valor-p} < 2.5\,\% \quad \text{(Log-Normal 5 intervalos)}$$
  $$5\,\% < \text{valor-p} < 10\,\% \quad \text{(Log-Normal 4 intervalos)}$$
  $$5\,\% < \text{valor-p} < 10\,\% \quad \text{(Weibull)}$$

  - Considerando los 4 intervalos para el modelo Log-Normal, al 5 % de significancia ambas distribuciones ajustan bien los datos, pero el modelo Log-Normal ajusta mejor que el modelo Weibull
  - Considerando los 5 intervalos originales para el modelo Log-Normal, al 5 % de significancia el modelo Weibull tiene mejor ajuste respecto al modelo Log-Normal.

Pregunta 2 *Regresión Lineal Simple*

Durante un cierto mes de 30 días se ha observado un incremento en las interacciones en twitter. Usted está interesado en construir un modelo que permita explicar estas variaciones en las interacciones, para eso dispone de la siguiente información **diaria**:

- *Interacción*: Y en unidad de miles
- *Actividad*: X_1, 1 si existe actividad y 0 en caso contrario
- *Actividades Externas*: X_2, número de hechos diarios del exterior
- *T° en Santiago*: X_3
- *Día de la semana*: X_4, 1 si es laboral y 0 en caso contrario

Un rápido análisis y ajuste de varios modelos de regresión con R se presentan a continuación:

```
> sum(Y)
[1]  1200
> var(Y)
[1]  8.27586
```

Modelo	Variables	R2
1	X_1	0.334
2	$X_1 + X_2$	0.500
3	$X_1 + X_3$	0.375
4	$X_1 + X_2 + X_4$	0.583
5	$X_1 + X_2 + X_3 + X_4$	0.708

A continuación se muestra la salida de R para el primer modelo:

```
> lm(formula = Y ~ X1)

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

Residual standard error: XXXXX on XXXX degrees of freedom
Multiple R-squared: XXXX, Adjusted R-squared: XXXX

F-statistic: XXXXX on X and XXXX DF,  p-value: 0.02
```

Obtenga los valores faltantes y considerando $\alpha = 5\,\%$, ¿es posible afirmar que la actividad explica el incremento de las interacciones en twitter?

Solución:

Como estamos en un modelo de regresión lineal simple, entonces un valor a determinar fácil es el valor-p para el intercepto, ya que el valor-p para el estadístico F_0 es igual que el valor-p para el valor T_0 de la pendiente (solo para el de la pendiente, **no** intercepto):

$$\texttt{p value = 0.02 = Pr(>|t|)}_{\texttt{X1}}$$

el valor de `Multiple R-squared` se encuentra en la tabla por enunciado para el modelo 1:

$$\texttt{Multiple R-squared} = 0.334$$

La cantidad de datos que se tienen son 30 ya que por enunciado se menciona que se dispone de la información diaria de las interacciones y que se analizó en un cierto mes de 30 días, por lo que

$$\texttt{n} = 30$$

A partir de la cantidad de datos se tienen los grados de libertad para el `Residual standard error`:

$$\texttt{on n-2 degrees of freedom} = \texttt{on 28 degrees of freedom}$$

Para el `F-statistic` se tienen dos grados de libertad, `df1 and df2 DF`, en regresión lineal simple se tiene que los grados de libertad son:

$$\texttt{1 and n-2 DF} = \texttt{1 and 28 DF}$$

Como se tiene la cantidad de datos n y el valor de R^2, es posible encontrar $r^2 =$ `Adjusted R-squared` a partir de la siguiente relación:

$$R^2 = 1 - (1 - r^2)\frac{n-2}{n-1}$$

despejando r^2 se tiene:

$$r^2 = \texttt{Adjusted R-squared} = 0.310214$$

Para determinar el valor de `Residual standard error` se utiliza la ecuación de r^2:

$$r^2 = 1 - \frac{S^2_{Y|x}}{S^2_Y}$$

se despeja $S^2_{Y|x}$ y reemplazando con el valor de r^2 y la varianza de Y que por enunciado es $S^2_Y =$ `var(Y) = 8.27586` se obtiene:

$$0.310214 = 1 - \frac{S^2_{Y|x}}{8.27586}$$
$$S^2_{Y|x} = 5.70857$$

Entonces:

$$\texttt{Residual standard error} = S_{Y|x} = \sqrt{S^2_{Y|x}} = 2.38926$$

Se puede obtener el `t value` de la pendiente mediante el `Pr(>|t|)`:

$$\begin{aligned}
2 \cdot P(T > |\texttt{t value}|) &= \texttt{Pr(>|t|)} \\
P(T > |\texttt{t value}|) &= \texttt{Pr(>|t|)}/2 \\
1 - P(T \leq |\texttt{t value}|) &= \texttt{Pr(>|t|)}/2 \\
P(T \leq |\texttt{t value}|) &= 1 - \texttt{Pr(>|t|)}/2 \\
|\texttt{t value}| &= t_{1-\texttt{Pr(>|t|)}/2}(n-2) \\
|\texttt{t value}| &= t_{0.99}(28)
\end{aligned}$$

Buscando en la tabla t-Student fijando $\nu = 28$ se tiene que el percentil 99 % es:

$$\texttt{|t value|} = 2.467$$

El signo de `t value` es el mismo que para el del `Estimate` de la pendiente y de enunciado se habla de un incremento, por lo que la pendiente es positiva y por lo tanto el `t value` también lo será:

$$\texttt{t value} = +2.467$$

El valor restante es `F-statistic`, se podría determinar mediante el `p value` pero no hay tabla para percentiles 98 %, pero se puede obtener mediante el `t value` de la pendiente:

$$\texttt{F-statistic} = \texttt{t value}^2$$

```
F-statistic = 6.08609
```

Esto **solo** es válido para **modelo de regresión lineal simple**. A continuación se muestra la salida de R completa:

```
Coefficients:

             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  ---------   ---------   ------    -----
X1           ---------   ---------   +2.467    0.02

Residual standard error: 2.389 on 28 degrees of freedom
Multiple R-squared: 0.334, Adjusted R-squared: 0.310

F-statistic: 6.086 on 1 and 28 DF,  p-value: 0.02
```

Con un 5 % de significancia se puede afirmar que la actividad explica el incremento de las interacciones en twitter, es decir, existe regresión.

$$\texttt{p value} < \alpha$$

Pregunta 3 *Estimación de parámetros por QQ-Plot*

El archivo BandaAncha.xlsx contiene información sobre la velocidad media de subida (en Mbps) durante periodos de alto tráfico en 50 ciudades de Chile y 4 operadores: M, E, W y C.

```
Base  <-  rio::import("BandaAncha.xlsx")
```

Para cada ciudad se promedia las velocidades de subida de las cuatro operadoras y se ajusta mediante un gráfico de probabilidad una distribución Exponencial, a continuación se muestra el proceso y resultado en R:

```
> X = aggregate(Velocidad~Ciudad, data = Base, FUN = mean)[,2]
> xp = sort(X)
> N = length(X)
> p = 1:N/(N+1)
> par = lm(xp~(-log(1-p)))$coef
> par
[1]  (Intercept)   (-log(1-p))
       5.0658975     25.345879
```

De igual forma se realiza un gráfico de probabilidad de una distribución Log-Logística, los resultados son los siguientes:

```
> par = lm(log(xp)~(log(p/(1-p))))$coef
> par
[1]  (Intercept)   (log(p/(1-p)))
       2.596741         0.5798554
```

Obtenga los parámetros estimados de la distribución Exponencial y Log-Logística, ¿en cuanto se ha desplazado la distribución exponencial?

Solución:

Empecemos con la distribución Exponencial, la relación entre los percentiles teóricos y empíricos es la siguiente:

$$x_p = \alpha + \frac{1}{\nu} \cdot [-\ln(1-p)]$$

Del QQ-Plot para la distribución exponencial se obtiene el valor de la pendiente que es utilizado para determinar el valor estimador de ν:

$$\frac{1}{\nu} = 25.345879$$

$$\hat{\nu} = 0.0394541$$

Para contestar la pregunta de cuanto se ha desplazado la distribución Exponencial se utiliza el valor del intercepto, este es el valor del desplazamiento:

$$\hat{\alpha} = 5.0658975$$

Para la distribución Log-Logística se tiene la siguiente relación entre los percentiles teóricos y empíricos:

$$\ln(x_p) = \mu + \sigma \cdot \ln\left(\frac{p}{1-p}\right)$$

El valor estimado de μ y σ son el intercepto y la pendiente del QQ-Plot 2 respectivamente:

$$\hat{\mu} = 2.596741$$

$$\hat{\sigma} = 0.5798554$$

Pregunta 4 *Comparación de Poblaciones*

Continuando con el contexto de la pregunta anterior. Para los promedio por ciudad, calculados anteriormente, ¿existe evidencia para afirmar que la velocidad media es mayor a 10 Mbps? Suponga que los datos distribuyen Gamma$(4, \nu)$, realice la prueba de hipótesis correspondiente y responda la pregunta en base a un nivel de significancia del 1 %

```
> mean(X)
[1]  11.23052
> length(X)
[1]  50
```

Solución:

Primero se debe encontrar el estimador máximo verosímil de ν, suponiendo que se tiene una muestra de n variables aleatorias iid Gamma(k, ν) con k conocido, entonces la función de verosimilitud es:

$$\begin{aligned} L(\nu) &= \prod_{i=1}^{n} \frac{\nu^k}{\Gamma(k)} X_i^{k-1} e^{-\nu X_i} \\ &= \left(\frac{\nu^k}{\Gamma(k)}\right)^n \left(\prod_{i=1}^{n} X_i\right)^{k-1} \exp\left(-\nu \sum_{i=1}^{n} X_i\right) \end{aligned}$$

La función de log-verosimilitud es:

$$\begin{aligned} \ln(L) &= n \ln\left(\frac{\nu^k}{\Gamma(k)}\right) + (k-1) \sum_{i=1}^{n} \ln(X_i) - \nu \sum_{i=1}^{n} X_i \\ &= nk \ln(\nu) - n \ln[\Gamma(k)] + (k-1) \sum_{i=1}^{n} \ln(X_i) - \nu \sum_{i=1}^{n} X_i \end{aligned}$$

Derivando respecto a ν e igualando a 0 se obtiene el EMV buscado:

$$\begin{aligned} \frac{d \ln(L)}{d\nu} &= 0 \\ \frac{nk}{\nu} - \sum_{i=1}^{n} X_i &= 0 \\ \frac{nk}{\nu} &= \sum_{i=1}^{n} X_i \\ \frac{k}{\nu} &= \frac{1}{n} \sum_{i=1}^{n} X_i \\ \frac{1}{\nu} &= \frac{\overline{X}_n}{k} \\ \hat{\nu} &= \frac{k}{\overline{X}_n} \end{aligned}$$

Para el test de hipótesis pedido se tienen dos opciones, test de hipótesis sobre la media o sobre el parámetro ν.

Alternativa 1: Test de hipótesis sobre la media, se tienen las siguientes dos hipótesis:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_a : \mu > \mu_0$$

donde $\mu = g(\nu) = \frac{k}{\nu}$, su EMV por invarianza es $\hat{\mu} = g(\hat{\mu}) = \frac{k}{\hat{\nu}} = \overline{X}_n$ y $g(\nu_0) = \mu_0 = 10$, el estadístico

de prueba en su versión general es:

$$Z_0 = \frac{g(\hat{\nu}) - g(\nu_0)}{\sqrt{\frac{[g'(\nu_0)]^2}{I_n(\nu_0)}}} \dot{\sim} \text{Normal}(0,1)$$

donde:

$$g'(\nu) = \frac{dg}{d\nu} = -\frac{k}{\nu^2} \longrightarrow g'(\nu_0) = -\frac{k}{\nu_0^2}$$

e $I_n(\nu) = -\text{E}\left(\frac{\partial^2}{\partial \nu^2}\ln(L)\right)$ es la información de Fisher:

$$\frac{\partial^2}{\partial \nu^2}\ln(L) = -\frac{nk}{\nu^2}$$

$$l_n(\nu) = -\text{E}\left(\frac{\partial^2}{\partial \nu^2}\ln(L)\right) = -\text{E}\left(-\frac{nk}{\nu^2}\right) = \frac{nk}{\nu^2}$$

$$I_n(\nu_0) = \frac{nk}{\nu_0^2}$$

Por lo que el pivote a utilizar es:

$$Z_0 = \frac{\overline{X}_n - \mu_0}{\sqrt{\frac{k}{n\nu_0^2}}} = \frac{\overline{X}_n - \mu_0}{\sqrt{\frac{\mu_0^2}{kn}}} \dot{\sim} \text{Normal}(0,1)$$

Evaluando el pivote se tiene:

$$\begin{aligned} Z_0 &= \frac{11.23052 - 10}{\sqrt{\frac{10^2}{4 \cdot 50}}} \\ &= 1.740218 \end{aligned}$$

El valor-p a calcular es:

$$\begin{aligned} \text{valor-p} &= P(Z > 1.740218) \\ &= 1 - P(Z \leq 1.740218) \\ &= 1 - \Phi(1.74) \\ &= 0.04092 \end{aligned}$$

al 1 % de significancia no existe evidencia para afirmar que la velocidad media es mayor a 10 Mbps

$$\text{valor-p} > \alpha$$

Alternativa 2: Test de hipótesis sobre ν, en este caso las hipótesis a utilizar son:

$$H_0 : \nu = \nu_0 \qquad \text{vs} \qquad H_a : \nu < \nu_0$$

El estadístico de prueba a utilizar para ν es:

$$Z_0 = \frac{\hat{\nu} - \nu_0}{\sqrt{\frac{1}{I_n(\nu_0)}}} \dot{\sim} \text{Normal}(0,1)$$

donde $\nu_0 = \frac{k}{\mu_0} = \frac{4}{10} = 0.4$ y la información de Fisher no cambia de la calculada anteriormente, por lo que:

$$Z_0 = \frac{\frac{k}{\overline{X}_n} - \nu_0}{\sqrt{\frac{\nu_0^2}{nk}}} \dot{\sim} \text{Normal}(0,1)$$

evaluando se obtiene:

$$\begin{aligned} Z_0 &= \frac{\frac{4}{11.23052} - \frac{4}{10}}{\sqrt{\frac{0.4^2}{50 \cdot 4}}} \\ &= -1.549544 \end{aligned}$$

El valor-p a calcular es:

$$\begin{aligned} \text{valor-p} &= P(Z < -1.549544) \\ &= 1 - \Phi(1.55) \\ &= 0.06063 \end{aligned}$$

al 1 % de significancia no existe evidencia para afirmar que la velocidad media es mayor a 10 Mbps

$$\text{valor-p} > \alpha$$

Pregunta 5 *Tamaño de Muestra*

En la reciente elección presidencial del Perú (en realidad se refiere al 2016), finalmente la diferencia entre los candidatos fue de un 0.32 %. ¿Qué tamaño de preferencia debería haber tenido una encuesta previa a la elección, para que la estimación del porcentaje de preferencia de un candidato tuviese ese margen de error? Utilice una confianza del 95 % y el criterio de varianza maximiza.

Solución:

Considerando $X_1, ..., X_n$ una muestra aleatoria Bernoulli(p), donde p representa el porcentaje de preferencia por un candidato cualquiera.

Para el caso de varianza máxima se tiene que el tamaño de la población bajo una confianza del 95 % ($\alpha = 5\,\%$) está dada por:

$$n = \left(\frac{k_{1-\alpha/2}}{2\omega}\right)^2$$

donde $k_{1-\alpha/2} = k_{0.975}$ es el percentil 97.5 %, de la tabla Normal(0,1) se tiene que este valor es de aproximadamente 1.96, el error de estimación por enunciado es:

$$\omega = 0.32\,\% = 0.0032$$

por lo que:

$$n = \left(\frac{1.96}{2 \cdot 0.0032}\right)^2 \approx 93789.06$$

Por lo que se necesita realizar la encuesta a aproximadamente 93790 personas para que se produzca este margen de error igual a la diferencia real observada.

Pregunta 6 *Intervalos de Confianza*

Recientemente ocurrió un sismo de gran magnitud (imaginario obviamente), producto del desaste en la infraestructura, la ocurrencia de daños en una vivienda puede ser modelada por una distribución Bernoulli(p), con p la probabilidad de daño. Un indicado para realizar estudios referidos al tema es la "chance de daño", la cual se define como:

$$\text{CD} = g(p) = \frac{p}{1-p}$$

a) Determine el estimador de máxima verosimilitud para la chance de daño

b) Determine la distribución asintótica del estimador en (a) y, a partir de este resultado, obtenga un intervalo de confianza al $(1-\alpha)\times 100\,\%$ para la chance

c) Determine un nivel de confianza tal que, en una muestra de 60 casas, cuando 28 de ellas poseen daño, no hay posibilidad de considerar que la chance de error es 1

Solución:

Solución a)

Se pide el EMV de CD, por invarianza se tiene que si $\hat{p}$ es el EMV de p, entonces $\widehat{\text{CD}} = g(\hat{p})$ es el EMV de CD:

$$\widehat{\text{CD}} = g(\hat{p}) = \frac{\hat{p}}{1-\hat{p}} = \frac{\overline{X}}{1-\overline{X}}$$

Solución b)

La distribución asintótica de $\hat{p}$ es:

$$\hat{p} = \overline{X} \dot{\sim} \text{Normal}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Por la propiedad de invarianza, la distribución asintótica de $\widehat{\text{CD}}$ es:

$$\widehat{\text{CD}} = \frac{\overline{X}}{1-\overline{X}} \dot{\sim} \text{Normal}\left(\frac{p}{1-p}, \sqrt{\frac{p}{n(1-p)^3}}\right)$$

El intervalo de confianza para una parámetro general $g(\theta)$ con un nivel de confianza $(1-\alpha)\times 100\,\%$ es:

$$\langle g(\theta)\rangle_{1-\alpha} \in g(\hat{\theta}) \pm k_{1-\alpha/2}\cdot\sqrt{\widehat{\text{Var}(g(\hat{\theta})}}$$

Para este problema, el intervalo pedido es:

$$\langle \text{CD}\rangle_{1-\alpha} \in \frac{\hat{p}}{1-\hat{p}} \pm k_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}}$$

$$\langle \text{CD}\rangle_{1-\alpha} \in \frac{\overline{X}}{1-\overline{X}} \pm k_{1-\alpha/2}\cdot\sqrt{\frac{\overline{X}}{n(1-\overline{X})^3}}$$

Solución c)

Se observa que la chance de daño no puede ser mayor a 1, lo que se pide es determinar para que nivel de confianza, el extremo superior del intervalo es menor a 1, las casas totales de la muestra son 60 y las casas que presentan daño son 28, entonces:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{60} X_i = \frac{28}{60} = 0.466667$$

por lo que valor estimado de la chance de daño es:

$$\widehat{\text{CD}} = \frac{\overline{X}}{1-\overline{X}} = 0.8750001$$

La desviación estándar estimada de la chance de daño es:

$$\sqrt{\frac{\overline{X}}{n(1-\overline{X})^3}} = 0.226428$$

Entonces:

$$\begin{aligned}
\frac{\overline{X}}{1-\overline{X}} + k_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n(1-\overline{X})^3}} &< 1 \\
0.8750001 + k_{1-\alpha/2} \cdot 0.226428 &< 1 \\
k_{1-\alpha/2} &< 0.5520524 \\
\Phi^{-1}(1-\alpha/2) &< 0.5520524 \\
1-\alpha/2 &< \Phi(0.55) \\
1-\alpha/2 &< 0.7088 \\
2-\alpha &< 1.4177 \\
1-\alpha &< 0.41768
\end{aligned}$$

El nivel de confianza debe ser menor a 41.7 %.

Pregunta 7 *Comparación de Poblaciones*

Frente al tema de género que se ha instalado en la discusión nacional, un investigador busca determinar si existen diferencias basales en el desempeño según género. Específicamente, su hipótesis es que frente a situaciones de stress (por ejemplo, desarrollo de una evaluación) las mujeres tienen un mayor control. Con el fin de verificar o refutar su hipótesis lleva a cabo un cuasi-experimento que consisten en someter a una situación estresantes a dos grupos - Hombres y Mujeres - seleccionados al azar dentro de los alumnos y alumnas del curso, registrando la presión arterial durante la evaluación. Los resultados son:

Resultados	Hombres	Mujeres	Ambos
Num Casos	15	17	32,00
Promedio	85	92	88,00
Mediana	80	85	84,00
Desv. Estándar	8	13	10,00
Promedio LN	-	-	4,48
Desv. Estándar LN	-	-	0,12

Asumiendo que la presión arterial diastólica se comporta de acuerdo a una distribución Normal y teniendo claro que un alza de presión implica una pérdida de control frente a situaciones estresantes, ¿es válida la afirmación del experto para un nivel de significancia de 10 %? Debe plantear hipótesis, indicar el test a utilizar, especificar y validar supuestos (cuando sea necesario) y sus conclusiones deben estar basadas en el valor-p.

Solución:

Un mayor control ante situaciones de estrés indica un bajo valor de presión arterial, por lo que las hipótesis a analizar son en base al promedio de la presión arterial entre hombres y mujeres:

$$H_0 : \mu_H = \mu_M \qquad \text{vs} \qquad H_a : \mu_H > \mu_M$$

ya que las varianzas poblacionales no se conocen, se pueden utilizar dos test, comparación de medias con σ_H y σ_M desconocidos pero iguales o diferentes, para saber cual utilizar es necesario realizar primero un test de varianzas bajo las siguientes hipótesis:

$$H_0 : \sigma_H = \sigma_M \qquad \text{vs} \qquad H_a : \sigma_H \neq \sigma_M$$

Bajo H_0, el estadístico de prueba a utilizar es:

$$F_0 = \frac{S_H^2}{S_M^2} \sim \text{Fisher}(n_H - 1, n_M - 1)$$

con $S_H^2 = 8^2$, $S_M^2 = 13^2$, $n_H = 15$ y $n_M = 17$, por lo que el estadístico tiene un valor de:

$$F_0 = 0.3786982 \sim \text{Fisher}(14, 16)$$

El criterio de rechazo de H_0 es el siguiente:

- $F_0 > F_{1-\alpha/2}(14, 16)$
- $F_0 < F_{\alpha/2}(14, 16)$

Ambos criterios implican que valor-p $< 10\,\%$, para concluir se debe obtener por percentiles $F_{0.05}(14, 16)$ y $F_{0.95}(14, 16)$, de la tabla Fisher se puede obtener el segundo valor fijando $\text{df}_1 = 14$ y $\text{df}_2 = 16$:

$$F_{0.95}(14, 16) = 2.37$$

Mediante la siguiente relación se obtiene el segundo percentil:

$$F_{0.05}(14, 16) = \frac{1}{F_{0.95}(16, 14)}$$

de tabla se tiene que $F_{0.95}(16,14) = 2.44$, por lo que

$$F_{0.05}(14,16) = \frac{1}{2.44} = 0.409836$$

Comparando se tiene que:

$$F_0 < F_{0.05}(14,16) \longrightarrow \text{valor-p} < 10\,\%$$

en base a esto se concluye que existe evidencia para rechazar H_0, es decir, las varianzas poblaciones se pueden considerar diferentes con un 10 % de significancia.

Continuando con el test de comparación de medias, como σ_H y σ_M son desconocidas pero diferentes, entonces se utiliza el siguiente estadístico de prueba:

$$T_0 = \frac{\overline{H} - \overline{M}}{\sqrt{\dfrac{S_H^2}{n_H} + \dfrac{S_M^2}{n_M}}} \sim \text{t-Student}(\nu)$$

con $\nu = \dfrac{\left(\dfrac{S_H^2}{n_H} + \dfrac{S_M^2}{n_M}\right)^2}{\dfrac{(S_H^2/n_H)^2}{n_H - 1} + \dfrac{(S_M^2/n_M)^2}{n_M - 1}}$, reemplazando con los datos se obtiene:

$$T_0 = -1.85709 \sim \text{t-Student}(105)$$

El valor-p a calcular es:

$$\text{Valor-p} = P(T > T_0) = 1 - P(T \leq T_0)$$

En la tabla se fija $\nu = \infty$ y se debe buscar dos valores donde se encuentre T_0, ya que este es negativo no se encuentra en la tabla, pero se puede utilizar la siguiente relación:

$$t_p(\nu) = -t_{1-p}(\nu)$$

entonces, $-T_0$ se encuentra entre $t_{0.95}(105) = 1.645$ y $t_{0.975}(105) = 1.960$, por lo que:

$$t_{0.95}(105) = 1.645 < 1.85709 < t_{0.975}(105) = 1.960$$

entonces:

$$t_{0.025}(105) = -1.960 < -1.85709 < t_{0.05}(105) = -1.645$$

$$0.025 < P(T \leq T_0) < 0.05$$

$$0.95 < P(T > T_0) < 0.975$$

Como valor-p $> 10\,\%$, por lo tanto no se rechaza H_0 y no se apoya la afirmación del experto.