Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática
MA3402-1 Estadística Autor: Arturo Lazcano

Tarea 3

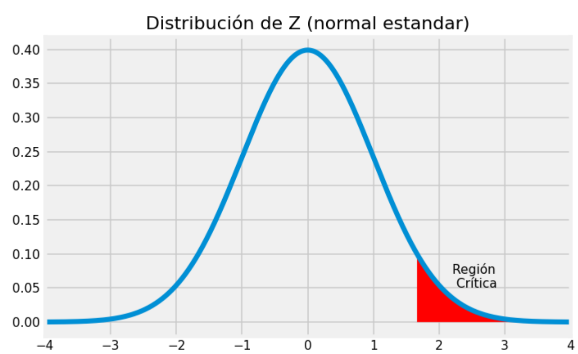
Pregunta 1: Test de Hipótesis

(a) Denotando las observaciones como ξ, tendremos que su promedio (variable desconocida) se denotará por $\bar{\xi}$. Con esto tenemos que $X = X_1, ..., X_{16} \sim \mathcal{N}(100, 16^2)$ y por lo tanto $\bar{X} = \frac{1}{16}\sum_{i=1}^{16} X_i \sim \mathcal{N}(100, \frac{16^2}{16}) = \mathcal{N}(100, 16)$.

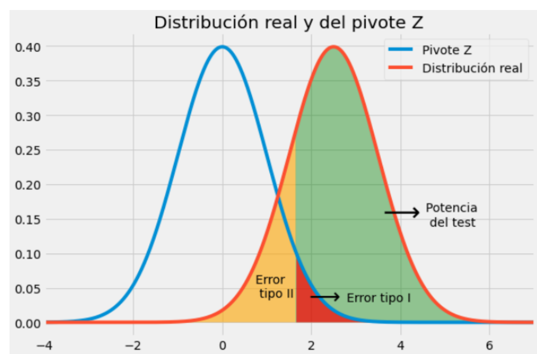
Así, el pivote $Z = \frac{\bar{X}-100}{\frac{16}{\sqrt{16}}} \sim \mathcal{N}(0,1) \Longrightarrow \mathcal{P}(\bar{X} \geq \bar{\xi}) = \mathcal{P}\left(\frac{\bar{X}-100}{4} \geq \frac{\bar{\xi}-100}{4}\right)$

Por otro lado, la región crítica se puede ver en python usando norm.ppf(1-α)=1,645 para $\alpha = 0,05$.

(b),(c)



(a) Distribución pivote Z



(b) Distribución pivote Z y real

Figura 1: Distribuciones con regiones críticas y potencia

El error tipo I viene dado por $\mathbb{P}(Z \geq 1,645) = 1 - 0{,}95 = 0{,}05$ donde Z distribuye como $\mathcal{N}(0,1)$. Mientras que el error tipo II viene dado por $\mathbb{P}(W \leq 1,645)$ donde W distribuye como $\mathcal{N}(\frac{5}{2}, 1)$ por lo que es equivalente a trasladar la media y calcular $\mathbb{P}(Z \leq 1,645 - \frac{5}{2}) = 0,196$. Por último, la potencia se calcula como 1-error tipo II, por lo tanto la potencia del test es 1-0,196=0,804 y toda esta información se puede apreciar en la Figura 1 (b).

(d)

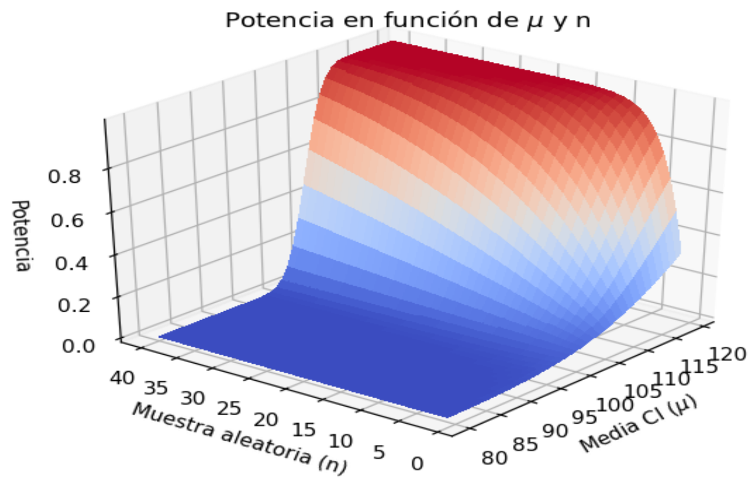


Figura 2: Potencia para distintos valores de μ y n

(e) El error tipo I viene dado por la región crítica del pivote Z que distribuye como una normal estándar, por lo cual solo depende de α y viene dado por $\mathbb{P}(Z \geq \text{norm.ppf}(1\text{-}\alpha))$, por lo que no dependede otros parámetros. Por otro lado, el error tipo II viene dado por 1-Potencia, por lo tanto, a mayor potencia, menor valor de error tipo II, y como se observa en el gráfico, la variable con mayor peso para la potencia es la media (μ) por lo que este error depende del valor que tenga μ.
Con esto podemos concluir que el error tipo I viene dado solamente por el valor de α mientras que las variables n y α no afectan significativamente a la potencia mientras que μ si lo hace por lo que el error tipo II depende de esta última variable (ver Figura 2 y 3).

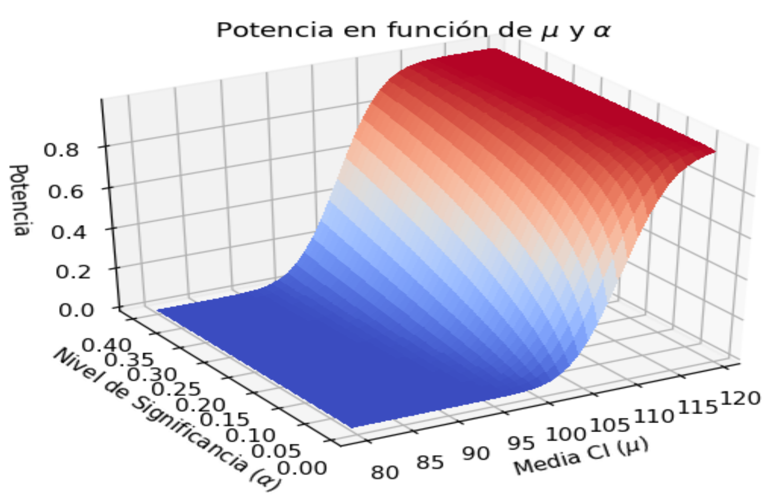


Figura 3: Potencia para distintos valores de μ y α

Pregunta 2: Intervalo de Confianza

(a)

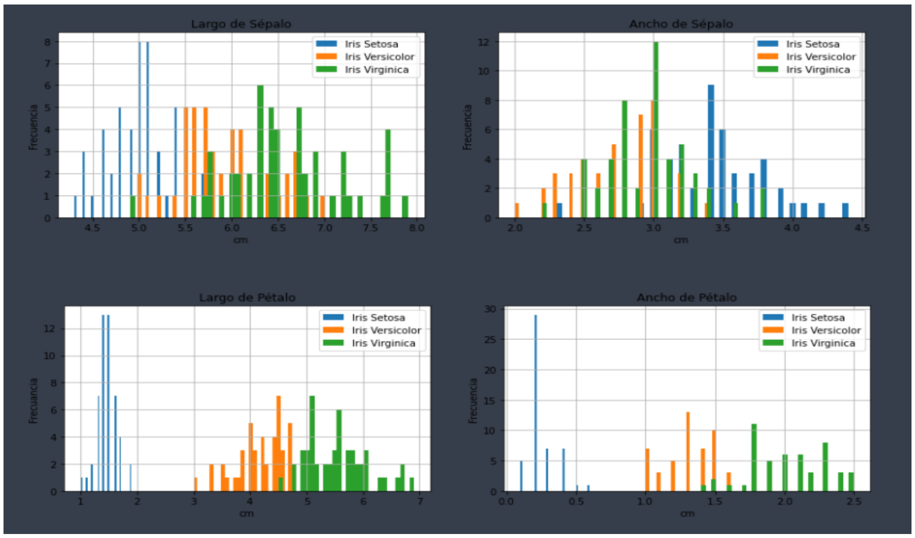


Figura 4: Largo y ancho de sépalo y pétalo para Iris Setosa, Versicolor y Virginica

Como se puede apreciar en el gráfico, existe una diferencia significativa para los valores de atributos para distintas especies, por ejemplo, para el largo y ancho de pétalo, la Iris Setosa esta muy por debajo con respecto a las otras dos especies. Por otro lado, la mayor incertidumbre podría darse en el largo y ancho de sépalo para las especies Versicolor y Virginica para rangos de 5.5 a 6.5 cm en el largo y 2.5 a 3 cm en el ancho.

(b) Sabemos por enunciado que $X_i \sim \mathcal{N}(\mu, \sigma^2)$ con σ conocido. Luego, normalizando tenemos que $\dfrac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \sim \mathcal{N}(0,1)$ donde n=50 al ser dividido el grupo total por especies. Con esto, podemos tomar Z_1, Z_2 tales que

$$\mathbb{P}\left(\bar{X} - \frac{Z_2\sigma}{\sqrt{50}} < \mu < \bar{X} - \frac{Z_1\sigma}{\sqrt{50}}\right) = 1 - \alpha \ (\bigstar).$$

Tomando $\alpha = 0{,}35 \in [0,1]$ resulta $(\bigstar)$=0,65 y podemos calcular la probabilidad acumulada en python para Z_1, Z_2 simétricos y la gráfica centrada con lo que resulta $|Z_1| = |Z_2| = 0,9346$

```
Intervalos de confianza (A(x),B(x)) con α=0.35
Largo sépalo Iris Setosa:  (4.959878839366244, 5.052121160633757)
Largo sépalo Iris Versicolor:  (5.8684621024335755, 6.003537897566424)
Largo sépalo Iris Virginica:  (6.504798970497599, 6.671201029502398)
Ancho sépalo Iris Setosa:  (3.378401701689965, 3.477598298310035)
Ancho sépalo Iris Versicolor:  (2.7289413721758757, 2.811058627824125)
Ancho sépalo Iris Virginica:  (2.93180324987494, 3.0161967501250606)
Largo pétalo Iris Setosa:  (1.439277108044547, 1.4847228919554534)
Largo pétalo Iris Versicolor:  (4.198514968094015, 4.3214850319059845)
Largo pétalo Iris Virginica:  (5.479787890440543, 5.624212109559458)
Ancho pétalo Iris Setosa:  (0.23221092794315729, 0.25978907205684265)
Ancho pétalo Iris Versicolor:  (1.3001252483374852, 1.3518747516625145)
Ancho pétalo Iris Virginica:  (1.9900636881202762, 2.0619363118797245)
```

Figura 5: Intervalos de confianza por atributo y especie, $\alpha = 0{,}35$

Con esto podemos observar que se puede identificar una especie de Iris teniendo sus 4 atributos de forma relativamente segura, esto pues se puede ver en la Figura 4 como sus atributos están relativamente separados, y por lo tanto, la especie es identificables al juntar sus 4 características.

(c) Si pudiéramos recolectar una gran cantidad de datos, la probabilidad sobre los intervalos sería de 100(1-α)=65 %. Sin embargo, no es así con solo una nueva muestra.
Por otro lado, en general, en punto medio de los intervalos de confianza es un estadístico pues depende de los datos, ya que se calcula su promedio y desviación estándar para la creación de los 12 intervalos de confianza.

Preguntа 3: Prior de Jeffreys

(a) Usaremos la fórmula $p(\lambda|Y) = \dfrac{p(Y|\lambda)p(\lambda)}{p(Y)} \propto p(Y|\lambda)p(\lambda)$
Por un lado, $p(Y|\lambda) = \displaystyle\prod_{i=1}^{n} \frac{\lambda^y e^-\lambda}{y!} \propto e^{-\lambda n}\lambda^{\sum_{i=1}^n Y_i}$ Por independencia de datos y ya que el término $\displaystyle\prod_{i=1}^{n} y_i!$ no depende de λ.
Por otro lado, $p(\lambda) = \dfrac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$ donde $\frac{\beta^\alpha}{\Gamma(\alpha)}$ es un término constante y conocido, por lo cual, lo anterior es proporcional a $\lambda^{\alpha-1}e^{-\beta\lambda}$.
Con esto tenemos que
$$p(\lambda|Y) \propto e^{-\lambda n}e^{-\beta\lambda}\lambda^{\alpha-1}\lambda^{\sum_{i=1}^n Y_i} = e^{-\lambda(n+\beta)}\lambda^{\sum_{i=1}^n Y_i+\alpha-1} \sim \text{Gamma}\left(\sum_{i=1}^{n} Y_i + \alpha, n+\beta\right)$$
Por lo tanto, la istribución gamma es la familia de prior conjugados para lambda de una distribución Poisson ya que la posterior sigue la misma ley del prior.

(b) Al estar usando un modelo de Poisson para λ, por la parte (a) sabemos que el modelo gamma es de la familia conjugada de λ por lo que este será nuestro prior.
Con el modelo ya escogido, solo resta determinar lo hiperparámetros, es decir, α y β. Se sabe que lambda

(tasa de ocurrencia de un suceso) es aproximadamente 1, con un error también de 1 orden de magnitud. Como la relación entre los modelos de poisson y gamma es que $\lambda = \alpha$, entonces $\alpha = 1$ y β es el parámetro calculado como el ratio de la ocurrencia de sucesos, por lo tanto el problema descrito puede ser modelado con un $\lambda \sim \text{Gamma}(1, \frac{1}{365})$.

(c) Al ser el prior una distribución uniforme, sabemos que esta distribucion es constante en un cierto intervalo [a,b] y que integrando desde 0 a ∞, esto resultara ∞ y no 1 como las distribuciones propias pues $\lambda > 0$. Por lo tanto, esta distribución se denomina impropia.
Veamos ahora su posterior, sabemos que
$p(\lambda) \propto 1$ y que
$P(\lambda|Y) \propto e^{-\lambda n}\lambda^{\sum Y_i + 1 - 1}$ (ni quita ni pone para ver mejor los parámetros α y β).
Por lo tanto, la posterior es una $\text{Gamma}\left(\sum_{i=1}^{n} Y_i + 1 \ , \ n\right)$. Luego, al ser la posterior una distribución gamma, por definición, su pdf siempre integra 1 con $\alpha, \beta > 0$. Por lo tanto resulta en una posterior propia.

(d) Sabemos que $p_j(\lambda \propto \sqrt{I(\lambda)}$ por lo tanto
$p_j(\lambda) \propto \sqrt{-\mathbb{E}\left(\frac{\partial^2}{\partial\lambda^2} log \ p(Y|\lambda)\right)} = \sqrt{-\mathbb{E}\left(\frac{\partial^2}{\partial\lambda^2} log \ \lambda^y \frac{e^{-\lambda}}{y!}\right)} = \sqrt{\mathbb{E}\left(\frac{y}{\lambda^2}\right)} = \sqrt{\frac{1}{\lambda}}$ pues el valor esperado de una distribución de Poisson es λ.

(e) Sea la transformación $\eta = h(\lambda) = \sqrt{\lambda}$.
Entonces, $\frac{1}{\lambda} = ((\sqrt{\lambda})')^2 I_q(\eta) \iff \frac{4\lambda}{\lambda} = I_q(\eta) \iff I_q(\eta) = 4$. Por lo tanto
$p_j(\phi) \propto \sqrt{4} = 2$ lo que es quivalente a un prior uniforme.

(f) El prior de Jeffreys resulta no ser propio al no integrar 1 (pues $\lambda > 0$). Veamos ahora su posterior:
$p(\lambda) \propto \frac{1}{\lambda}$
$p(Y|\lambda) \propto e^{-\lambda n}\lambda^{\sum y_i}$
$\Rightarrow p(\lambda|Y) \propto e^{-\lambda n}\lambda^{\sum y_i - \frac{1}{2} + 1 - 1} \sim \text{Gamma}\left(\sum_{i=1}^{n} y_i + \frac{1}{2} \ , \ n\right)$ y como los parámetro α, β de la distribución gamma son mayores que 0, por definición, esta integra 1, por lo que resulta en una posterior propia.