Un enfoque Bayesiano al problema de Test de Hipótesis

La entrega de este TP debe incluir:

a) Un informe detallado en formato pdf, hecho en latex o word. En dicho informe toda afirmación que forme parte de la resolución debe ser debidamente justificada. No es aceptable presentar en el informe solamente el resultado numérico o un gráfico. En caso de realizar un desarrollo que conduce al resultado numérico, debe incluirse.

b) El script en R que permite obtener los resultados numricos.

En un hospital, se desea diseñar una experiencia para probar una nueva droga contra la diabetes, es decir, para disminuir la glucosa. Dicha experiencia involucrará n individuos. En cada individuo se mide el nivel de glucemia en ayunas antes y después de administrarles la droga que se deja actuar durante 20 minutos.

Sea X_i la glucemia del individuo i-ésimo antes de la administración de la droga e Y_i la glucemia después de la droga y sea $D_i = X_i - Y_i$. Es decir, que si en promedio observamos muchos valores positivos de D_i esto indicará que la droga tiende a cumplir su cometido.

Se sabe que

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \eta \end{pmatrix}, \mathbf{\Sigma} \right) \quad \text{donde} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Los pares (X_i, Y_i) conforman una **muestra apareada**, es decir, son independientes entre sí pero X_i no es independiente de Y_i. Notemos que el valor medio de glucemia en sangre depende del tiempo de medición (si fue antes o después del consumo de la droga). Todos los parámetros en cuestión son desconocidos.

El objetivo de los investigadores es poder sacar alguna conclusión sobre $\theta = \mu - \eta$. Para diseñar la experiencia, necesitan determinar el número de pacientes que van a estudiar. Como el costo de la experiencia es alto, los investigadores no quieren tomar muchos más pacientes de los que sean necesarios para sacar conclusiones sobre la efectividad de la droga, pero tampoco quieren realizar la experiencia y no poder obtener una conclusión confiable por haber tomado pocos individuos.

1. Sea $\mathbf{D} = (D_1, \ldots, D_n)$. Hallar la distribución de D_1 y un estimador IMVU para θ.

2. Se sabe que la droga no aumenta la glucemia y por lo tanto, para medir su efectividad, nos interesa saber si la droga produce una disminución estadísticamente significativa sobre la media de la glucemia en ayunas. Es decir, queremos efectuar el siguiente test de hipótesis sobre θ

$$H_0 : \theta = 0 \quad vs \quad H_1 : \theta > 0.$$

Supongamos que σ_X, σ_Y y ρ son conocidos, entonces se propone el siguiente estadístico,

$$T = \sqrt{n}\frac{\overline{D}}{\sigma_D} = \sqrt{n}\frac{\overline{X} - \overline{Y}}{\sigma_D}$$

con $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$.

   a) Probar que bajo H_0 $T \sim N(0, 1)$.

   b) Deducir que un test para H_0 está dado por

$$\phi(\mathbf{D}) = \begin{cases} 1 & T > K \\ 0 & T < K\ . \end{cases} \tag{1}$$

   c) Hallar el valor de K para que el test (1) tenga nivel α exacto.

   Indicaremos por z_α el valor tal que $\mathbb{P}(Z > z_\alpha) = \alpha$ con $Z \sim N(0, 1)$, o sea, $\Phi(z_\alpha) = 1 - \alpha$.

3. Supongamos conocido el valor de n, queremos expresar la probabilidad de error de Tipo II en función de n y de la alternativa elegida.

   Sea $\beta_\phi(\theta_1)$ la potencia del test (1) de nivel α en θ_1. Llamaremos

$$\pi\left(n, \frac{\theta_1}{\sigma_D}, \alpha\right) = \beta_\phi(\theta_1) \quad \text{y} \quad p_{II}\left(n, \frac{\theta_1}{\sigma_D}, \alpha\right) = 1 - \pi\left(n, \frac{\theta_1}{\sigma_D}, \alpha\right)$$

   a la potencia y a la probablidad del Error de Tipo II en θ_1, donde ha reforzado la dependencia de n, α y θ_1/σ_D.

   Encontrar una expresión para $\pi(n, \theta_1/\sigma_D, \alpha)$ y $p_{II}(n, \theta_1/\sigma_D, \alpha)$.

4. Se quiere elegir el tamaño de muestra de modo que el test (1) de nivel α tenga potencia igual a β (o sea, error de Tipo II $1 - \beta$) cuando $\theta = \theta_1 > 0$. Es decir, se quiere controlar el error de Tipo I y el error de Tipo II en θ_1.

   Supongamos σ_D^2 conocido, $\theta_1 > 0$, α y β prefijados.

   a) Encontrar el menor valor de n tal que el test (1) de nivel α tenga potencia por lo menos β en θ_1, o sea, $\beta_\phi(0) = \alpha$ y $\beta_\phi(\theta_1) \geq \beta$. Como n depende de θ_1 y de σ_D sólo a través del cociente θ_1/σ_D, llamaremos $n(\alpha, \beta, \theta_1/\sigma_D)$ a ese valor.

b) Supongamos que $\theta_1/\sigma_D = 0.4$, $\alpha = 0.025$. Qué tamaño de muestra elegiría para garantizar que el test (1) de nivel 0.025 tenga potencia por lo menos $\beta = 0.9$ en θ_1 (o sea, probabilidad de tipo II a lo sumo 0.1)? Es decir, calcule $n_0 = n(0.025, 0.9, 0.4)$.

Observación: *Recuerde que n es un natural*

5. Los investigadores están dispuestos a tener cierta incertidumbre ya que no conocen exactamente el valor de σ_D y de θ_1 y quieren incluir esa incertidumbre en la determinación del tamaño de muestra y en el cálculo de la probabilidad de tipo II.

Supongamos que los investigadores quieren tener una probablididad de error de tipo I de a lo sumo 0.025. Para ello eligen el test (1) que tiene nivel $\alpha = 0.025$.

Por experiencias previas, se sabe que es razonable suponer que θ y σ_D son independientes, Más aún, supondremos que $\theta \sim N(0.4, 0.01)$ y que $\sigma_D \sim |V|$ donde $V \sim N(1, 0.09)$.

El objetivo de este ejercicio es tratar de simular el comportamiento de n y π ya que la distribución exacta de n y π en función de la de θ y σ_D no es fácilmente deducible.

a) Genere 10000 valores $(\theta_j^\star, \sigma_{D,j}^\star)$, $1 \leq j \leq 10000$ utilizando la distribución de (θ, σ_D).

b) Haga un histograma de los valores $\theta^\star$ y $\sigma_D^\star$ obtenidos. Comparelos con la densidad de θ y σ_D sobreimponiendo dichas densidades.

c) Para cada valor $(\theta_j^\star, \sigma_{D,j}^\star)$ obtenido, calcule

   i) $\pi_j^\star = \pi(n_0, \theta_j^\star/\sigma_{D,j}^\star, 0.025)$

   ii) $n_j^\star = n(0.025, 0.9, \theta_j^\star/\sigma_{D,j}^\star)$

d) i) Haga un histograma y un boxplot de la distribución de $\pi^\star$.

   ii) Le parece adecuado usar un boxplot para esta distribución?

   iii) Qué pasa con la potencia?

   iv) Qué valor de $\pi^\star$ parece ser el más representativo?

   v) Cuál es la potencia mediana cuando el tamaño de muestra es n_0?

   vi) Cuál sería la probabilidad de que la potencia fuera menor que 0.75, es decir, la probabilidad de error II mayor que 0.25? En base a eso le recomendaría a los investigadores usar como tamaño de muestra n_0?

e) i) Haga un histograma y un boxplot de la distribución de $n^\star$.

   ii) Le parece adecuado usar un boxplot para esta distribución?

   iii) Qué valor de n parece ser el más representativo?

   iv) Qué tamaño de muestra le recomendaría usar a los investigadores si ellos quieren tener una certidumbre del 80% de tener una potencia del 0.9 cuando θ y σ_D tienen las distribuciones descriptas? Y si quisieran una certidumbre del 95%?