Examen computacional sobre Test de Hipótesis
Métodos estadísticos en física experimental
No son todos iguales

A. Rabinovich (LU:316/08)
Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, 1428 Buenos Aires, Argentina.
(Dated: July 20, 2017)

I. INTRODUCCIÓN

Sean dos hipótesis especificadas completamente por dos valores distintos de un parámetro θ en una función de distribución de probabilidad $f(x|\theta)$. Para ejemplificar, la hipótesis nula, H_0, supone que $\theta = \theta_0$ mientras que la alternativa, H_1, asume $\theta = \theta_1$.
Suponiendo que la hipótesis nula es cierta, es posible encontrar una región R en el espacio muestral W para la observación x tal que la probabilidad de que x pertenezca a R es igual a algún valor numérico asignado previamente. La región R es llamada la región de rechazo o región crítica para H_0, mientras que $(W - R)$ es la región de aceptación de H_0.
Las dos regiones están separadas por un valor x_c y se llama significancia a la probabilidad preasignada de que una observación x pertenezca a R. Ésta determina el nivel de significancia $100\alpha\%$.

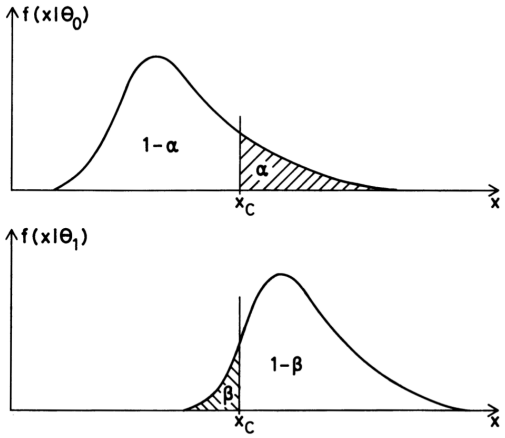

FIG. 1: Ejemplo de Error de Tipo I, Error de Tipo II y potencia de test. Tomado de Frodesen[1].

De ésta definición se desprende que existe una probabilidad α de rechazar H_0 cuando era en realidad verdadera, error conocido como Error de Tipo I.
Por otro lado, el Error de Tipo II se define como la probabilidad de aceptar H_0 cuando era en realidad falsa, y la probabilidad β de que ocurra depende de H_1.
Finalmente, la potencia de un test se define como la probabilidad $(1 - \beta)$ de rechazar H_0 cuando era realmente falsa. La potencia de un test depende de la cantidad de muestras, el nivel de significancia y el tamaño del efecto. Todas éstas definiciones se ejemplifican en la figura 1.[1]

II. EL PROBLEMA

Sean dos muestras independientes $X_1, ..., X_m$ e $Y_1, ..., Y_n$ con distribuciones gaussianas $N(\mu_x, \sigma^2)$ y $N(\mu_y, \sigma^2)$ con μ_x, μ_y, σ desconocidas.
Se quiere estudiar computacionalmente cual de los siguientes test es el de mayor potencia para testear a un nivel de significancia α la hipótesis H_0: $\mu_x \leq \mu_x$ contra H_1 : $\mu_x > \mu_y$.
El primero consiste en rechazar H_0 cuando $U \geq T_{m+n-2,(1-\alpha)}$, donde:

$$U_1 = (\overline{X} - \overline{Y})\sqrt{\frac{m+n-2}{(\frac{1}{m}+\frac{1}{n})(S_x^2+S_y^2)}} \tag{1}$$

con $S_x^2 = \sum_i^m (x_i - \overline{X})^2$ y $S_y^2 = \sum_i^n (y_i - \overline{Y})^2$ y $T_{m+n-2,(1-\alpha)}$ es el cuantil $(1-\alpha)$ de la destribución T con $n + m - 2$ grados de libertad.

El segundo consiste en rechazar H_0 cuando $U \geq T_{dof,(1-\alpha)}$, donde:

$$U_2 = \frac{(\overline{X} - \overline{Y})}{S} \tag{2}$$

con $S^2 = \frac{S_x^2}{m(m-1)} + \frac{S_y^2}{n(n-1)}$ y

$$dof \sim \frac{(\frac{S_x^2}{m(m-1)}+\frac{S_y^2}{n(n-1)})^2}{(\frac{S_x^2}{m(m-1)})^2/(m-1)+(\frac{S_y^2}{n(n-1)})^2/(n-1))}$$

III. SIMULACIONES

Se tomaron m muestras para $X \sim N(0.1, 1)$ y n muestras para $Y \sim N(0, 1)$, variando m y n desde 10 hasta 1100 cada una (es decir, en una iteración anidada dentro de la otra) y se calcularon los estadísticos U_1 y U_2. Luego, para cada par m y n se obtuvo el $p - value$ definido como $1-$ la probabilidad acumulada de la T correspondiente hasta U. Ésto se repitió 100 veces y se calculó el promedio del $p - value$ para cada par, como se muestra en la figura 2.
Se observa que los $p - values$ obtenidos con ámbos tests son muy similares, y en ambos casos fue necesario tomar aproximadamente 900 muestras de cada distribución normal para obtener un $p - value < 0.05$.

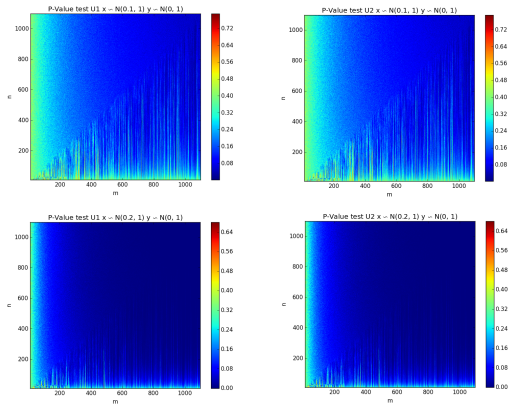

FIG. 2: De izquierda y derecha y de arriba a abajo, P-Value para U1 y U2 para diferencia de medias de 0.1 y de 0.2 para distintos tamaños de muestra

Se realizó el mismo proceso tomando muestras de $X \sim N(0.2, 1)$ graficando el $p - value$, y se encontró que con tomar alrededor de 300 muestras de cada distribución normal era suficiente. Ésto concuerda con la dependencia del test con el tamaño del efecto. A un efecto mayor (mayor diferencia entre las medias), se requiere una menor cantidad de muestras para obtener la significancia necesaria.
Por otro lado, se midió la potencia de cada test tomando 2000 muestras de $U1$ y $U2$ para cada n y m, con $\mu_x > \mu_y$ y con $\mu_x < \mu_y$. En la figura 3 se observan histogramas para $n = m = 100$, $n = m = 500$ y $n = m = 900$ para diferencia de medias de 0.1 y $n = m = 10$, $n = m = 100$ y $n = m = 300$ para diferencia de medias de 0.2 respectivamente. En rojo, x_c.

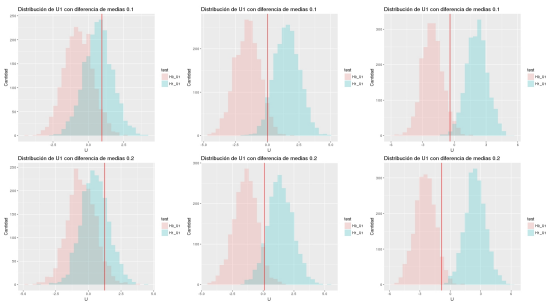

FIG. 3: Histogramas de H_0 y H_1 para test U1 con diferencia de media de 0.1 para n=m con n=100, 500 y 900 respectivamente (arriba) y para diferencia de medias de 0.2 con n=m con n=10, 100, 300. Se observa que las distribuciones se separan a medida que aumenta el tamaño de la muestra o el tamaño del efecto. En rojo, x_c.

Se verifica que a medida que aumenta el tamaño de las muestras, ambas distribuciones se separan. También se observa que a mayor efecto, menor cantidad de muestras alcanza para separar las distribuciones. La figura 4 muestra la potencia de cada test y su dependencia con el tamaño de la muestra. Nuevamente, se observa que para tamaños de efecto mayores se requiere de una menor cantidad de muestras para conseguir aumentar la potencia del test.

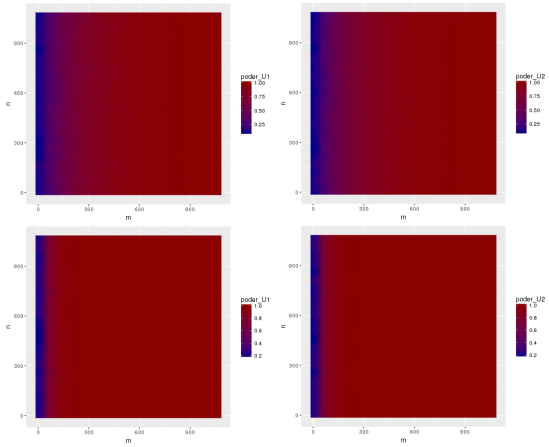

FIG. 4: De izquierda y derecha y de arriba a abajo, potencia para U1 y U2 para diferencia de medias de 0.1 y de 0.2 para distintos tamaños de muestra.

Se observa además que basta con aumentar el tamaño de las muestras de x para lograr aumentar la potencia del test.
Finalmente, se generaron las distribuciones de $U1$ para H_0 y H_1 tomando 20000 muestras con una diferencia de media de 0.1 y n=m=100. Se calculó la probabilidad acumulada p_i para cada valor de $U1$ de la distribución H_1 cómo:

$$p_i = \int_{-\inf}^{U_i} f_X(t)dt \tag{3}$$

tomando $f_X(t)$ como U_1 para H_0 y H_1.
La figura 5 muestra los histogramas de p_i para U_1 bajo cada una de las hipótesis.

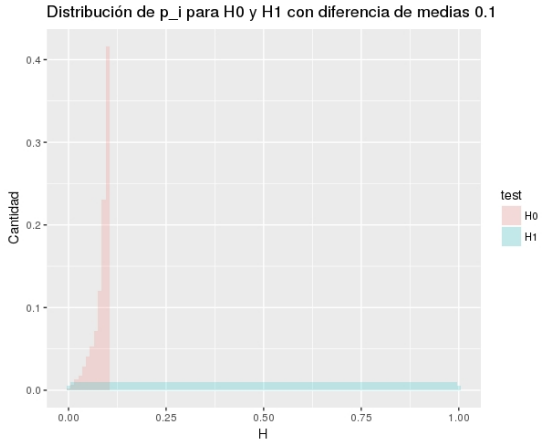

FIG. 5: Histograma de p_i para la distribución de U_1 para H_0 y H_1.

Se observa que si la función que se está integrando es la distribución de la que proviene la variable aleatoria que se usa como límite de integración se obtiene un histograma uniforme, mientras que si se utiliza la proveniente de H_0 se obtiene un histograma no uniforme. Es decir, cuanto

menos uniforme sea el histograma de p_i habrá más razones para rechazar H_0.

IV. CONCLUSIONES

Ambos tests resultaron tener una potencia similar para testear a un nivel de significancia $\alpha = 0.05$ la hipótesis H_0: $\mu_x \leq \mu_x$ contra H_1: $\mu_x > \mu_y$.
La potencia en ambos casos resultó dependiente del tamaño de la muestra y del tamaño del efecto, verificándose que a mayor tamaño de efecto, menor cantidad de muestras requeridas para testear al nivel de significancia deseado. Se encontró que para una diferencia en las medias de 0.1, se requieren más de 900 muestras de cada distribución normal para alcanzar una significancia de $\alpha = 0.05$ mientras que para una diferencia de medias de 0.2, alcanza con aproximadamente 300 muestras.
Finalmente, fue posible además, distinguir las distribuciones de U obtenidas bajo la hipótesis nula y la hipótesis alternativa a partir del estadístico p_i.

[1] Frodesen, *Probability and statistics in particle physics* (UNIVERSITETSFORLAGET).

Todos los códigos utilizados para generar las muestras se pueden encontrar en:
https : *//github.com/andresrabinovich/examen_estadistica*