

Seminario 4: Tests de Hipótesis 1

Vamos a comenzar la práctica de hoy analizando algunos conjuntos de datos, y luego estudiaremos mediante simulaciones algunas propiedades del test t y del test para la varianza. Específicamente, estudiaremos cómo se comportan cuando falla la hipótesis de normalidad. Finalmente, introduciremos un test (el test de Wilcoxon) para detectar diferencias en la media(na) de dos poblaciones no necesariamente normales.

1. Tests con datos reales

Los siguiente problemas están en los apuntes del curso.

1. Los datos del fichero `http://mat.uab.cat/~acabana/data/skull.txt` se pueden leer en R con la instrucción

```
skull=read.table("http://mat.uab.cat/~acabana/data/skull.txt")
```

son mediciones (en cm) de la parte más ancha de una serie de cráneos, publicados por Barnicott y Brothwell (1959), (*The evaluation of metrical data in the comparison of ancient and modern bones in medical biology and Etruscan origins*) .

La pregunta que se plantean los arqueólogos es si hay evidencias de diferencias de razas en estas dos mediciones, ¿Qué les podemos decir?

NOTA: conviene crear un vector para los cráneos italianos que NO tenga los “NA”, esto se puede hacer con la instrucción (una vez leídos los datos y guardados) `Italian=na.omit(Italian)` Se puede verificar que este nuevo vector no contiene los datos NA (*not available*), por ejemplo haciendo `length(Italian)`, o simplemente mirándolo, o pidiendo `summary(Italian)`.

2. Los datos que se pueden leer con la instrucción

```
read.table("http://mat.uab.cat/~acabana/data/darwin.txt")
```

corresponden a un experimento hecho por Darwin en el que se intentaba mostrar que la fecundación cruzada supera a la autofecundación. Se plantan 15 pares de semillas de *zea mays* de la misma edad, una obtenida mediante autofecundación y otra mediante fecundación cruzada, y se planta cada par de semillas en un mismo recipiente, de manera que las plantas de cada par crecen en condiciones casi iguales (experimento pareado). Los datos corresponden a la altura en pulgadas de cada planta después de un período de tiempo. Darwin consultó a Galton sobre el análisis de estos datos, que posteriormente fueron discutidos por Fisher en su *Design of Experiments*. Analiza los datos, haciendo los gráficos y tests que resulten pertinentes. ¿Tenía razón Darwin?

2. Potencia del test para media de las normales

Supongamos que tenemos una m.a.s. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, con σ^2 conocido, fijemos $\sigma^2 = 1$. Graficar la función de potencia de los siguientes tests:

U- $H_0 : \mu = 0$ contra $H_1 : \mu > 0$

B- $H_0 : \mu = 0$ contra $H_1 : \mu \neq 0$

3. ¿Cómo se comporta el test t cuando los datos no son normales?

En el fichero <http://mat.uab.cat/~acabana/data/lipids.txt> hay datos sobre el colesterol en sangre obtenidos en un estudio (National Health and Nutrition Examination Survey (US)) para 3026 personas mayores de 12 años.

Vamos a estudiar cuán bien se comporta el test t mediante el siguiente experimento: supongamos que la población entera son estos 3026 individuos, y tomemos $N = 1000$ sub-muestras de tamaño $n = 25$ de esta población. Veamos cuántos de los Intervalos de confianza $1 - \alpha$ construidos a partir de las muestras de tamaño n contienen la media de la población global, o, lo que es equivalente, cuántos de los tests t para $H_0 : \mu = \mu_0$ contra la alternativa de que la media es diferente no rechazan H_0 , donde μ_0 es la media de la población de los 3026 individuos.

El siguiente código nos servirá para este propósito

```
lipids=read.table("//mat.uab.cat/~acabana/data/lipids.txt",header=T)
pop=lipids$TRG
mu=mean(lipids$TRG);mu
N = 10000 ## Numero de simulaciones
n =25 ## tamaño de la muestra
covered=numeric(N) ## aquí pondremos los resultados
for (i in 1:N){
  sam <- sample(pop, n)
  covered[i] <- t.test(sam, mu=mu)$p.value > 0.05
}
mean(covered)
```

4. Escribir un programa que repita los cálculos anteriores para distintos valores de n

```
n = c(2, 3, 4, 5, 10, 20, 40, 80, 160, 320, 640)
```

Guardar los resultados en una matriz (de $N \times 11$) y graficar la cobertura en función de n .

Una cosa que no es agradable sobre este gráfico es que la parte *interesante* (donde la cobertura cambia) está sólo a la izquierda. Una solución es usar un eje no lineal. Hay varias formas de hacer esto en R, una de las más simples es

```
plot(log2(n), coverage, type="l", ylim=c(0.85,1), xaxt="n", xlab="n")
x=seq(1, 9, 2)
axis(1, at=x, labels=2^x)
abline(h=0.95, col="red")
```

5. ¿Qué pasará con el test para la varianza? Repetir el análisis anterior para el test $H_0 : \sigma^2 = \sigma_0^2$ contra $H_1 : \sigma^2 \neq \sigma_0^2$.

El test t es bastante robusto frente a alejamientos de la normalidad, en cambio los intervalos para la varianza son mucho más sensibles a estos alejamientos. Este es uno de los motivos por los cuales las simulaciones son tan útiles en estadística: para examinar el comportamiento de diferentes métodos cuando las suposiciones en que se basan no se cumplen (este tipo de cuestiones suele ser difícil de responder analíticamente).

4. El test de Wilcoxon.

Dado un conjunto de números reales $\{x_1, \dots, x_m\}$ llamamos rango de uno cualquiera de sus elementos x_i , y lo denotamos $R(x_i)$, al lugar que ocupa x_i en el conjunto, luego de ordenarlo de menor a mayor, esto es: $R(x_i) = \sum_{h=1}^m \mathbf{1}_{\{x_h \leq x_i\}}$.

Cuando los elementos del conjunto son todos diferentes, el conjunto de sus rangos es $\{1, 2, \dots, m\}$. Cuando esto no ocurre el conjunto de los rangos contiene números repetidos, y hay números que no son rangos de ningún elemento. En lo que sigue, vamos a aplicar la definición de rangos a variables aleatorias, y supondremos que sus distribuciones son tales que los empates quedan excluidos con probabilidad 1.

En particular, este es el caso cuando $X_1, \dots, X_m \sim F$ y $Y_1, \dots, Y_n \sim G$ y F y G son funciones de distribución continuas.

La primera prueba basada explícitamente en rangos fue propuesta por Frank Wilcoxon¹.

Es una prueba sensible a desplazamientos relativos de las dos distribuciones.

El estadístico utilizado es

$$W = \sum_{i=1}^m R(X_i),$$

que tiende a dar resultados significativamente grandes cuando las X están desplazadas hacia la derecha de las Y , e, inversamente, significativamente pequeños cuando están desplazadas hacia la izquierda.

El valor mínimo de W es $\sum_{i=1}^m i = \frac{m(m+1)}{2}$ y el máximo $\sum_{i=n+1}^{m+n} i = mn + \frac{1}{2}m(m+1)$.

El comando `wilcox.test (X,Y)` de R calcula la versión de Mann-Whitney de este estadístico, restando el mínimo que puede alcanzar W :

$$W^* = \sum_{i=1}^m R(X_i) - \frac{m(m+1)}{2}$$

La esperanza y la varianza de W se pueden calcular fácilmente. $\mathbf{E}W = \frac{m}{m+n} \frac{(m+n)(m+n+1)}{2} = \frac{m(m+n+1)}{2}$, $\mathbf{Var}W = \frac{mn}{(m+n)(m+n-1)} \left(\frac{(m+n)(m+n+1)(2m+2n+1)}{6} - \frac{(m+n)(m+n+1)^2}{4} \right) = \frac{mn(m+n+1)}{12}$.

La probabilidad $\mathbf{P}(W \leq w)$ se obtiene con la instrucción `pwilcox(w,m,n)`, y con `wilcox.test(x,y, conf.int=T)` obtenemos un intervalo de confianza para la diferencia en la localización de las dos muestras.

6. Repetir el problema 2 usando el test de Wilcoxon.

El estadístico de Wilcoxon puede adaptarse fácilmente para estudiar el problemas de localización para una sola muestra, o la diferencia de dos muestras emparejadas.

Si tenemos muestras $X_1, \dots, X_n \sim F$ y $Y_1, \dots, Y_n \sim G$ tomadas sobre los mismos individuos en dos circunstancias diferentes (por ejemplo antes y después de un tratamiento), estaremos interesados

¹Probability tables for individual comparisons by ranking methods. Biometrics 3, (1947), 119–122

en hacer un test para $F(t) = G(t - \Delta)$. Si miramos las diferencias $Z_i = Y_i - X_i$ estamos reduciendo el problema anterior al de saber si la localización de una muestra es mayor o menor que cero.

Ordenaremos los valores de $|Z_i|$, asignaremos rangos y sumaremos solamente los rangos de las Z_i positivas.

Este test se puede adaptar al caso en que nos interesa saber si la mediana de una población X_1, \dots, X_n es m , considerando en este caso las variables $Z_i = X_i - m$.

7. Volveremos a analizar los datos del cociente entre el ancho y el largo de los rectángulos de los indios Shoshone, para probar $H_0 : med = 0.618$ (el golden ratio)

```
ratio=c (0.693,0.670,0.654,0.749,0.606,0.553,0.601,0.609,0.672,0.663,
0.606,0.615,0.844,0.570,0.933,0.576,0.668,0.628,0.690,0.611)
wilcox.test(ratio,mu=0.618)
```

Un comentario acerca de la eficiencia de la prueba de Wilcoxon: contrariamente a lo que se pueda pensar, no siempre resultan más ineficientes los estadísticos basados en rangos que los que utilizan los valores de la muestra. Por ejemplo, si comparamos la eficiencia relativa asintótica (ARE) del test basado en W con la del test t (diseñado para muestras normales), observamos que, si n_i es el número de datos que necesitamos para tener una prueba con nivel α y potencia $1 - \beta$ usando la prueba de Wilcoxon y m_i es el número de datos que necesitamos con la prueba t para obtener el mismo nivel y potencia, bajo la alternativa i , entonces, la eficiencia relativa asintótica, $ARE = \lim m_i/n_i$ para diferentes alternativas es:

i	Alternativa	ARE
1	Normal	0.955
2	Uniforme	1
3	Logística	1.097
4	Laplace	1.5
5	Cauchy	∞
6	Exponencial	3

Por lo tanto, para muestras grandes, la prueba t es un poco más eficiente que la de Wilcoxon para muestras normales, y bastante más ineficiente en otros casos.

5. El test de Ansari-Bradley

En la prueba de Wilcoxon, los números $1, 2, \dots, m+n$ se ubican en orden creciente en los $m+n$ lugares del vector ψ .

Si quisiéramos una prueba de *dispersión* para muestras con igual parámetro de localización, tendríamos que distribuir los números de otra forma: los valores más bajos en los extremos, y los más altos en el centro.

A partir de $X_1, \dots, X_m \sim F$ y $Y_1, \dots, Y_n \sim G$, muestras independientes, vamos a probar $H_0 : F = G$ contra la alternativa de que las muestras difieren en dispersión,

$$\frac{X - \theta}{\eta_1} \stackrel{d}{=} \frac{Y - \theta}{\eta_2}.$$

Si las varianzas son finitas, su cociente es $\gamma^2 = (\eta_1/\eta_2)^2$ y H_0 equivale a $\eta_1 = \eta_2$ y a $\gamma^2 = 1$. En la prueba de Ansari-Bradley la asignación de rangos se repite de manera simétrica:

1 2 3 4 ... 4 3 2 1

En R se utiliza el comando `ansari.test`.

Ejemplo Con los datos de la permeabilidad de la placenta, podemos probar H_0 : “Los datos tienen igual mediana pero la dispersión de la permeabilidad es diferente”.

Si ahora asignamos los rangos correspondientes a la prueba de Ansari Bradley obtenemos $AB = 7 + 5 + 6 + 2 + 8 = 28$.

Y	R_{AB}	R_{ST}	X	R_{AB}	R_{ST}
0.8	3	5	1.15	7	15
0.83	4	8	0.88	5	9
1.89	2	3	0.9	6	12
1.04	7	13	0.74	2	4
1.45	5	10	1.21	8	14
1.38	6	11			
1.91	1	2			
1.64	3	6			
0.73	1	1			
1.46	4	7			

```
ansari.test(x,y)
```

Ansari-Bradley test

```
data: x and y
```

```
AB = 28, p-value = 0.1372
```

```
alternative hypothesis: true ratio of scales is not equal to 1
```

En realidad, en este ejemplo, podríamos preguntarnos si los datos son razonablemente normales y hacer en caso afirmativo una prueba t para comparar las medias, y una prueba F para comparar las varianzas. Pero hay que tener en cuenta que una de las muestras tiene tamaño 5, por lo tanto las pruebas de normalidad no tiene por qué ser muy fiables.