Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Barcelona

Seminario 6: Tests de Hipótesis 3

Vamos a empezar la práctica de hoy introduciendo (visto anteriormente) el test de Wilcoxon para detectar diferencias en la media(na) de dos poblaciones no necesariamente normales. Finalmente, presentaremos los tests de permutaciones

1. El test de Wilcoxon.

Dado un conjunto de números reales $\{x_1, \ldots, x_m\}$ 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 $\{1, 2, \ldots, m\}$. 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 excluídos con probabilidad 1.

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

La primera prueba basada explícitamente en rangos fue propuesta por Frank Wilcoxon[1].

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.

$$\begin{aligned}
\mathbf{E}\, W &= \frac{m}{m+n}\frac{(m+n)(m+n+1)}{2} = \frac{m(m+n+1)}{2}, \\
\mathbf{Var} W &= \frac{mn}{(m+n)(m+n-1)}\left(\frac{(m+n)(m+n+1)(2m+2n+1)}{6} - \frac{(m+n)(m+n+1)^2}{4}\right) \\
&= \frac{mn(m+n+1)}{12}.
\end{aligned}$$

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

La probablidad $\mathbf{P}(W \leq w)$ 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.

1. Volver a analizar los datos del fichero `skull.txt` usando este estadístico y compararlo con lo que se obtiene realizando un test t.

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

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 $X_1, \ldots, X_n \sim F$ y $Y_1, \ldots, Y_n \sim G$ tomadas sobre los mismos individuos en dos circunstancias diferentes (por ejemplo antes y después de un tratamiento), estaremos interesados 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 $X_1, \ldots, X_n$ es m, considerando en este caso las variables $Z_i = X_i - m$.

2. 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). Empezar realizando un análisis gráfico. De nuevo, comparar los resultados con lo que se obtendría mediante un test t.

3. Leer la documentación de `binom.test` y usarlo para analizar los mismos datos.

```
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.

2. El test de Ansari-Bradley

En la prueba de Wilcoxon, los números $1, 2, \ldots, m+n$ 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 $X_1, \ldots, X_m \sim F$ y $Y_1, \ldots, Y_n \sim G$, 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:

$$1\ 2\ 3\ 4\ \ldots\ 4\ 3\ 2\ 1$$

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}	X	R_{AB}
0.8	3	1.15	7
0.83	4	0.88	5
1.89	2	0.9	6
1.04	7	0.74	2
1.45	5	1.21	8
1.38	6		
1.91	1		
1.64	3		
0.73	1		
1.46	4		

Cuadro 1:

```
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.

3. Pruebas Exactas o de Permutaciones

El procedimiento que vamos a describir a continuación proporciona una forma sencilla de obtener la distribución muestral de cualquier estadístico para poder encontrar regiones críticas, o calcular p-valores, bajo la hipótesis nula muy general de que las observaciones son intercambiables. Son tests no paramétricos, y tienen la ventaja de que pueden usarse con cualquier estadístico, sea o no conocida su distribución bajo la hipótesis nula.

Los primeros en proponer tests de este tipo fueron Fisher y Pitman, cerca de 1930, pero hasta la década de los 80 eran casi imposibles de implementar excepto para conjuntos de datos pequeños.

Su limitación, por otra parte, es la suposición (fuerte) de que las observaciones son intercambiables bajo la hipótesis nula, de manera que, por ejemplo, si queremos un test para comparar la media de dos poblaciones, tenemos que asegurarnos que las varianzas son iguales (para que los datos resulten intercambiables), así que no nos escapamos de los problemas (Behrens-Fisher) que tiene el test t.

Tenemos que pensar, entonces, que el procedimiento de las permutaciones sirve para hipótesis relativas a las distribuciones, más que a parámetros.

Comencemos con un ejemplo clásico de comparación de poblaciones, con datos de la cantidad de litros obtenidos el mismo día y en la misma región geográfica, a partir de 26 nubes bombardeadas con yoduro de platay 26 nubes naturales.

Queremos saber si la cantidad de lluvia que producen las nubes bombardeadas es significativamente mayor que la de las nubes sin bombardear.

Los datos se pueden leer con la instrucción:

```
lluvia=read.table("http://mat.uab.cat/~acabana/data/lluvia.txt")
x=lluvia$nobomb ; y=lluvia$bomb
```

1. Hacer un análisis gráfico de los datos. ¿Qué se puede decir sobre su distribución?
2. Hacer un test t para comparar las medias de ambas poblaciones. ¿Qué se concluye? ¿Es el test t adecuado?
3. Una alternativa posible es transformar los datos para que su distribución sea más simétrica. Hacerlop y volver a hacer un test t. ¿Cambia la conclusión?
4. Hacer un test de Wilcoxon. ¿Es más adecuado?¿Por qué?¿Cambia la conclusión?

El problema general en este caso es que

- La distribución de la muestra no es conocida
- Tamaños muestrales moderados
- La comparación es muy general. ¿Queremos comparar sólo las medias? ¿Varianzas? ¿$F_X = F_Y$?

Una alternativa razonable en estos casos es hacer un *test de permutaciones*.

El procedimiento general para hacer un test de permutaciones es el siguiente:

1. Especificar las hipótesis nula y alternativa.
2. Decidir qué estadístico se usará para hacer el test.
3. Crear un *nuevo* conjunto de datos, obtenido a partir de **permutaciones aleatorias** de los datos originales. Cómo se reordenarán, dependerá de la hipótesis nula.
4. Calcular el estadístico, y compararlo con el valor sobre la muestra original.

5. Repetir este procedimiento muchas veces (si n es pequeño, podemos hacer todas las permutaciones posibles y en ese caso diremos que el test es *exacto*).
6. Si el valor observado con la muestra original excede el cuantil $1-\alpha$ de la distribución del estadístico obtenida con las permutaciones, se rechaza la hipótesis nula con nivel α.

En el ejemplo de las nubes, para $\mathrm{H}_0 : \mu_x = \mu_y$ contra $H_1 : \mu_x < \mu_y$ podemos rechazar H_0 si $\mu_x - \mu_y < const.$

```
muestra=c(x,y)
rep=1000
original=mean(x)-mean(y)
distrib=numeric(rep)
l=length(x)+length(y)
for(i in 1:rep){
  sam=sample(muestra,l)
newx=sam[1:length(x)]
newy=sam[(length(x)+1):l]
distrib[i]=mean(newx)-mean(newy)
}
hist(distrib)
abline(v=quantile(distrib,0.05))
points(original,0)
```

3.1. La distribución de las permutaciones

Vamos a ver por qué este procedimiento funciona:

Supongamos que observamos dos muestras independientes $X_1, \ldots, X_n \sim F_X$ y $Y_1, \ldots, Y_m \sim F_Y$.

Llamemos Z al conjunto ordenado $Z = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ con índices

$$\nu = \{1, 2, \ldots, n, n+1, \ldots, n+m\} = \{1, \ldots, N\}$$

y $Z^* = (X^*, Y^*)$ a una permutación de la Z original, es decir, si π es una permutación de los enteros ν, $Z_i^* = Z\{\pi(i)\}$, y consideraremos que los primeros n elementos de Z^* corresponden a X^* y los restantes m a Y^*.

El número de posibles particiones es $\binom{N}{n}$.

Bajo $H_0 : F_X = F_Y$, cualquier Z^* elegido al azar tiene probabilidad $\frac{1}{\binom{N}{n}} = \frac{n!m!}{N!}$, es decir, si $F_X = F_Y$ todas las permutaciones son igualmente probables.

Si $\hat{\theta}(X, Y) = \hat{\theta}(Z, \nu)$ es un estadístico, entonces, la distribución de $\hat{\theta}^*$ (es decir, la distribución de las réplicas del estadístico) es

$$F_{\hat{\theta}^*}(t) = \mathbf{P}(\hat{\theta}^* \leq t) = \binom{N}{n}^{-1} \sum_{j=1}^{N} 1_{\{\hat{\theta}^* \leq t\}}$$

Si los valores grandes de $\hat{\theta}$ van a favor de la alternativa, rechazaremos H_0 cuando $\hat{\theta} > c_{1-\alpha}^*$, donde $c_{1-\alpha}^*$ es el cuantil $1-\alpha$ de la distribución de $\hat{\theta}^*$. Similarmente si el test es para la cola de abajo o bilateral.

Si el número de permutaciones es muy grande (es decir, si el tamaño de la muestra es grande), se puede obtener test aproximado eligiendo al azar un número relativamente bajo de permutaciones. Estos tests se conocen como tests de permutaciones *aproximados, o de Monte Carlo, o aleatorios*. Davison y Hinkley (*Bootstrap methods and their application*, p.159) dicen que "entre 99 y 999 permutaciones aleatorias deberían ser suficientes".