Prueba de permutación

Román Castillo

DEFINICIÓN

La prueba de permutación es un método no paramétrico (de distribución libre).Supongamos que queremos probar que $H_0 : X_1, \ldots, X_N$ es una muestra de cierta distribución F no especificada. La prueba de permutaciones se usa para este tipo de hipótesis, el p-valor se calcula al conocer el conjunto **S** de datos observados pero sin saber (o ignorando) qué valor de datos corresponde a X_1, X_2 así sucesivamente. El cálculo hace uso del hecho de que, condicional al conjunto de valores de datos S, cada una de las $N!$ posibles formas de asignar estos valores N a los datos originales son igualmente probables cuando a hipótesis nula es cierta.

Consideremos a $T(X1, X_2; \ldots . X_n)$ como un estadístico de la muestra, y como la **distribución de permutación** a la que asigna masa de probabilidad de $\frac{1}{N!}$ a cada permutación T_j, ahora sea t_{obs} el valor observado del estadístico, asumimos como criterio de rechazo $T > t_{obs}$:

$$\text{p-value} = P(T > t_{obs}) = \frac{1}{N!} \sum_{j=1}^{N!} I(T > t_{obs})$$

La probabilidad que se obtiene es exacta y no una aproximación asíntótica.

Ejemplo 1 Un grupo de estudiantes recibieron ayuda extraclase, se muestran las calificaciones del grupo durante las seis semanas que duró la ayuda: *72,69,74,78,75,79.* ¿Existe evidencia de una mejoría en el desempeño durante el tiempo que se brindo la asesoría?

SoluciónProbemos la hipotesis alternativa H_A :*"Hay mejoría (incremento) en el desempeño"* y como H_0 :*"El desempeño es el mismo"*. En caso de la que la hipótesis alternativa sea cierta, deberiamos encontrar una correlación positiva entre las calificaciones (x_j) y el momento del tiempo j en la que fueron tomadas. Definamos a $T() = \sum_{i=1}^{6}(x_i) \cdot (i)$, para la muestra se obtiene:

$$T_{obs} = (72)(1) + (69)(2) + (74)(3) + (78)(4) + (75)(5) + (79)(6) = 1593$$

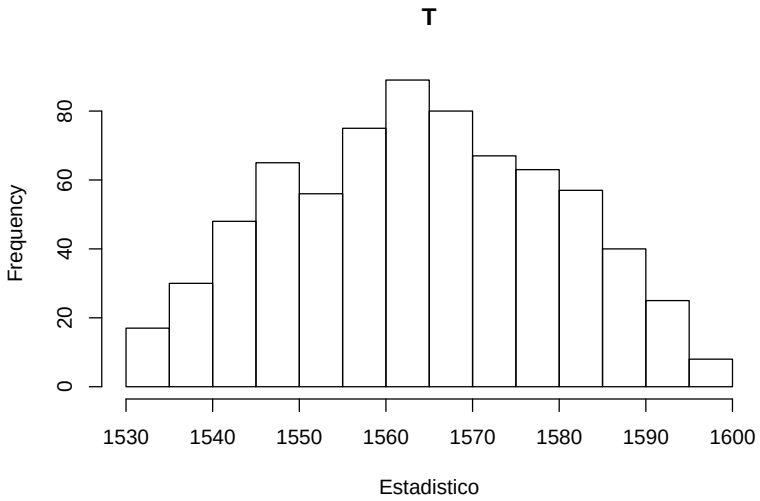
Si encontramos una "proporción significativa" de valores T que sean iguales o mayores que el observado, deberiamos rechazar la hipotesís nula, ya que se tiene evidencia de un aumento en la nota con respecto al tiempo.

Ahora efectuamos las permutaciones y se calcula T para cada una de ellas

Table 1: Permutaciones y estadístico

	X1	X2	X3	X4	X5	X6	T
2	69	72	74	75	79	78	1598
30	69	74	72	79	78	75	1589
224	72	79	74	69	78	75	1568
323	74	78	72	79	69	75	1557
326	74	78	75	69	79	72	1558
456	75	78	79	74	72	69	1538
551	78	74	79	75	69	72	1540
567	78	75	74	72	69	79	1557
574	78	75	79	72	74	69	1537
649	79	74	69	72	75	78	1565

	T	$P(x$
64	1593	0.98
65	1594	0.98
66	1595	0.99
67	1596	0.99
68	1597	1.00
69	1598	1.00
70	1599	1.00



Entonces p-valor $= P(T > 1593) = 1 - 0.98 = 0.02$. Hay evidencia para rechazar la hipótesis nula. Podemos asumir una mejora en el desempeño durante el periódo de ayuda

DOS MUESTRAS

Suponga dos muestras independientes:

$$X_1, \dots X_m \sim F_X \qquad Y_1, \dots Y_n \sim F_Y$$

Entonces deseamos poner a prueba:

$$H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y$$

Sea $T(X_1, \dots X_m, Y_1, \dots Y_n)$ un estadístico con las muestras, tal como $|\bar{X}_m - \bar{Y}_n|$, tenemos un total de $N = m + n$ observaciones, con ellas podemos formar $N!$ permutaciones y calcular T para cada una de ellas $(T_1, \dots, T_{N!})$ y obtener el *p-valor* del T_{obs}

Ejemplo 2 Consideremos una muestra de la forma $(X_1, X_2, X_3, Y_1, Y_2) = (2, 5, 9, 14, 1)$.

Solución Hallamos T en nuestra muestra original:

$$T = \left| \frac{2+5+9}{3} - \frac{14+1}{2} \right| = 2.17$$

las permutaciones de la muestra son:

Table 3: Permutaciones y Estadístico (10 obs)

	X1	X2	X3	Y1	Y2	T
3	1	2	9	5	14	5.50
15	1	9	5	2	14	3.00
20	1	14	2	9	5	1.33
21	1	14	5	2	9	1.17
27	2	1	9	5	14	5.50
68	5	14	1	9	2	1.17
89	9	5	14	1	2	7.83
93	9	14	2	1	5	5.33
112	14	5	2	9	1	2.00
117	14	9	2	1	5	5.33

Table 4: Tabla de frecuencia

T obs	P acumulada
1.17	0.1
1.33	0.2
2	0.3
2.17	0.4
3	0.5
4.5	0.6
5.33	0.7
5.5	0.8
7.83	0.9
8.83	1.0

Entonces p-valor $= P(T > 2.17) = 1 - 0.40 = 0.60$. No hay evidencia para rechazar la hipótesis nula. Podemos asumir que las muestras pertenecen a la misma población

Ejemplo 3 Los *DNA microarrays* permiten a los investigadores medir los niveles de expresión de miles de genes. Los datos son los niveles de ARN mensajero (ARNm) de cada gen, que se piensa que proporciona una medida de cuánta proteína produce ese gen. A grandes rasgos, cuanto mayor sea el número, más activo será el gen. La siguiente tabla, *reproducida de Efron et al. (2001)*, muestra los niveles de expresión para los genes de diez pacientes con dos tipos de células de cáncer de hígado. Hay 2,638 genes en este experimento, pero aquí mostramos solo los dos primeros. Los datos son *log-ratios* de de los niveles de intensidad de dos tintes de color diferentes utilizados en los *arrays*.

	Type I					Type II				
Patient	1	2	3	4	5	6	7	8	9	10
Gene 1	230	-1,350	-1,580	-400	-760	970	110	-50	-190	-200
Gene 2	470	-850	-.8	-280	120	390	-1730	-1360	-1	-330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Solución Probemos si el nivel medio del gen 1 es diferente entre los dos grupos. ν_1 representa la mediana del gen 1 del Tipo I y ν_2 la del gen 1 del Tipo II. Definimos a $T = |\nu_1 - \nu_2|$, Probemos que $H_0 : |\nu_1 - \nu_2| = 710$. Ahora estimamos la distribución de permutaciones por simulación, se presenta el resultado a continuación.

Encontramos que el valor p estimado es .045. Por lo tanto, si usamos un nivel de significancia de $\alpha = 0.05$,

diríamos que existe evidencia para rechazar la hipótesis nula de no diferencia.

Cuando el tamaño de la muestra crece, calcular las $N!$ permutaciones se vuelve más complicado, podemos aproximar el *p-valor* usando una aleatoria, de tamaño B, de todas las permutaciones y realizar el cálculo del *p-valor* con esa muestra

$$\text{p-value} \approx P(T > t_{obs}) = \frac{1}{B}\sum_{j=1}^{B} I(T > t_{obs})$$

Repetimos el ejercicio anterior usando 1,000 de las 3,628,800 permutaciones posibles

```
## 0.077 sec elapsed
```

ElVector	Freq
10	0.299
150	0.427
210	0.559
310	0.608
350	0.760
510	0.842
570	0.887
710	0.957
870	1.000

Obtenemos un *p-valor* de 0.46, prácticamente idéntico al que se obtiene con todas las permutaciones. En general, cuando se trata de muestras de gran tamaño *la prueba de permutaciones* devuelve un resultado bastante similar al paramétrico. Es notorio que resulta más útil en muestras pequeñas debido al poder computacional requerido

EJERCICIOS

1. Un jugador de béisbol tiene la reputación de comenzar "flojo" al inicio de una temporada, y luego mejorar continuamente a medida que la temporada avanza. Los siguientes datos son la cantidad de hits que tiene en las series consecutivas de juegos de la temporada, ¿ Hay evidencia que valide la reputación del jugador?

   8, 3, 7, 7, 13, 6, 12, 4, 4, 6

2. Un grupo de 10 ratones se expuso fuente de radiación. El grupo se dividió al azar en dos subgrupos. Los ratones del primer subgrupo vivían en un entorno normal de laboratorio, mientras que los del otro se criaron en un entorno especial sin gérmenes. Los siguientes datos muestran los días que vivieron los ratones después de la radiación:

   Grupo 1: 133, 145, 156, 159, 164

   Grupo 2: 145, 148, 157, 171, 178

¿Hay evidencia de que los tiempos de vida en los grupos sea diferente?

REFERENCIAS

1. Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.

2. Ross, S. M. (2014). Introduction to probability and statistics for engineers and scientists. Academic Press.
3. Good, P. I. (2004). Permutation, parametric, and bootstrap tests of hypotheses (Springer series in statistics).