Prueba de permutaciones

Román Castillo

CIMAT Unidad Monterrey

Prueba de permutaciones

La prueba de permutación es un método no paramétrico (de distribución libre) probar que una muestra pertenece cierta distribución F no especificada.

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 $N!$ (distribución de permutación), posibles formas de asignar estos valores N a los datos originales son igualmente probables cuando a hipótesis nula es cierta.

Prueba de permutaciones (algoritmo)

1. Dado el estadístico $T(X_1, X_2, \ldots, X_N)$, calcula T_{obs} con los datos obtenidos en la muestra
2. Permuta los datos y recalcula T_i para i-ésima permutación
3. Repetir el paso anterior hasta computar las $N!$ permutaciones de los datos y así obtener $T_1, T_2, \ldots, T_N!$
4. El *p-valor* se calcula por:

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

Se obtiene una probabilidad exacta, no una aproximación

Ejemplo Consideremos un conjunto de datos $A : (7, 5, 8)$ y definamos a $T() = (X_1 + X_2 - X_3)$ ¿Hay evidencia que pertenecen a la misma F?

- Calculamos T con los datos observados $T_{obs} = (7 + 5 - 8) = 4$
- Permutamos y calculamos T para cada permutación

Table 1: Permutaciones y estadístico

X1	X2	X3	T
5	7	8	4
5	8	7	6
7	5	8	4
7	8	5	10
8	5	7	6
8	7	5	10

$$\text{p-valor} = P(Y > 4) = 4/6$$

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?
Solución Probemos la hipotesis alternativa:

`H0:"El desempeño es el mismo"`
`HA:"Hay mejoría (incremento) en el desempeño"`

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

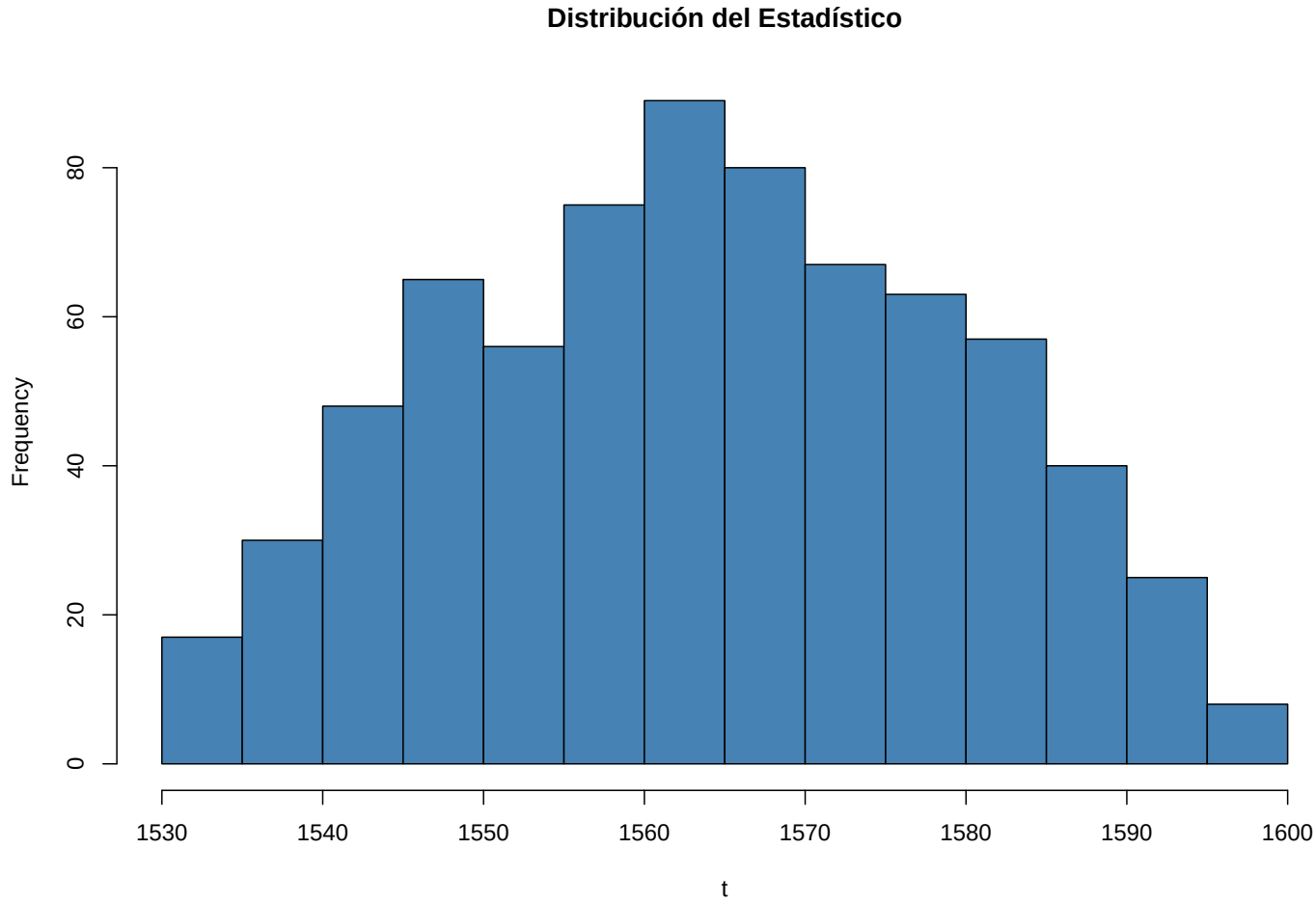
Table 2: Permutaciones y estadístico

	X1	X2	X3	X4	X5	X6	T
143	72	69	79	78	74	75	1579
164	72	74	79	69	78	75	1573
180	72	75	74	79	78	69	1564
376	75	69	78	74	79	72	1570
566	78	75	74	69	79	72	1553
581	78	79	69	75	72	74	1547
618	79	69	75	78	74	72	1556
658	79	74	72	75	78	69	1547

Table 3: T de permutación

	T	$P(x)$
60	1589	0.95
61	1590	0.95
62	1591	0.96
63	1592	0.97
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


Distribución del Estadístico
Frequency
0
20
40
60
80
1530
1540
1550
1560
1570
1580
1590
1600
t

Muestras de dos poblaciones

Supongamos que $X_1, \ldots, X_m \sim F_X$ y $Y_1, \ldots, Y_n \sim F_Y$ son dos muestras independientes y que bajo H_0 son identicamente distribuidas, sea $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ un estadístico, tal como una diferencia de medias $\overline{X_m} - \overline{Y_n}$

Ahora consideremos que $N = m + n$, y que con los datos $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ existen $N!$ permutaciones de ellos, y que en cada caso podemos calcular valores para $T_1, \ldots, T_{N!}$

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 $T = |\overline{X_m} - \overline{Y_n}|$ en nuestra muestra original:

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

Table 4: Permutaciones y Estadístico (10 obs)

	X1	X2	X3	Y1	Y2	T
37	2	9	1	5	14	5.50
41	2	9	14	1	5	5.33
63	5	9	2	1	14	2.17
105	14	2	5	1	9	2.00
116	14	9	1	5	2	4.50

Table 5: 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

p-valor $= P(T > 2.17) = 1 - 0.40 = 0.60.$

Ejemplo 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 $T = |\nu_1 - \nu_2|$, Entonces $T_{obs} : |\nu_1 - \nu_2| = 710$. Ahora calculemos las permutaciones y el estadístico.

Dado que $N = 10$, hay $10! = 3,628,800$ permutaciones posibles. 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

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

Efectuamos la prueba con una muestra de 1000 de las permutaciones y obtenemos:

```
## 0.087 sec elapsed
```

Table 6: Tabla de frecuencia

T obs	P acumulada
10	0.270
150	0.413
210	0.559
310	0.607
350	0.736
510	0.812
570	0.865
710	0.942
870	1.000

p-valor $= P(T > 710) = 1 - 0.954 = 0.046.$

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

¡Gracias!

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).

Table of Contents