Clase 4: Tests de permutaciones II

Matías Carrasco

6 de octubre de 2019

Índice

1. Diferencia entre tratamientos I

Supongamos que un nuevo tratamiento para la recuperación posquirúrgica se compara con un tratamiento estándar al observar los tiempos de recuperación (en días) de los pacientes en cada tratamiento. De los N sujetos disponibles para el estudio, n son asignados aleatoriamente para recibir el nuevo tratamiento, mientras que los $m = N - n$ restantes reciben el tratamiento estándar.

Supongamos que $n = 4$ pacientes son asignados al nuevo tratamiento y que $m = 3$ al estándar. La siguiente tabla muestra los resultados obtenidos:

Tabla 1: Tiempo de recuperación (en días) para ambos grupos de pacientes.

Tratamiento **nuevo**				Tratamiento **esándar**		
19	22	25	26	23	33	40

Queremos responder a la pregunta: ¿hay alguna diferencia entre los tratamientos? Una manera natural de comparar los tratamientos es usando el tiempo de recuperación promedio de cada uno. En el grupo asignado al nuevo tratamiento el tiempo de recuperación promedio es 23, y en el grupo asignado al tratamiento estándar el tiempo promedio es 32. Dicho de otra manera, el promedio en el nuevo tratamiento es 9 veces más chico que el promedio en el otro grupo. Ver la Figura 1.

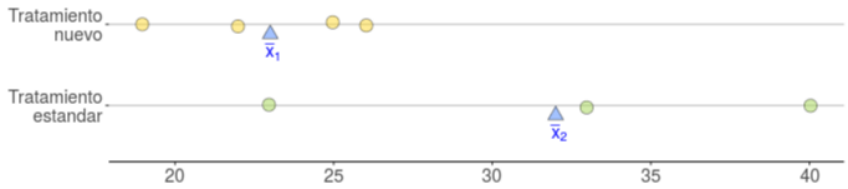


Figura 1: Distribución de los datos originales. Arriba se muestran los tiempos de recuperación de los pacientes del grupo de tratamiento nuevo (puntos en amarillo) y su promedio $\bar{x}_1 = 23$. Abajo se muestran los tiempos de recuperación de los pacientes del grupo de tratamiento estándar (puntos en verde) y su promedio $\bar{x}_2 = 32$.

¿Debemos considerar esta diferencia como evidencia significativa de que el nuevo tratamiento es mejor que el estándar? Recordando que los pacientes fueron asignados a los grupos aleatoriamente, debemos tener presente que la diferencia observada puede deberse a la suerte, tal vez los pacientes más "fuertes" quedaron, de casualidad, asignados en el grupo con el nuevo tratamiento.

Supongamos entonces que el tratamiento no tiene ningún efecto sobre el tiempo de recuperación de los pacientes. Si no hay diferencia entre los tratamientos, entonces el tiempo de recuperación para cada paciente será el mismo independientemente del tratamiento que reciba. Por ejemplo, el paciente de la Tabla 1 que se recuperó en 19 días con el nuevo tratamiento, se habría recuperado en la misma cantidad de tiempo con el tratamiento estándar, si no hay efecto del tratamiento.

Usemos como estadístico la diferencia de promedios

$$X = \left(\begin{smallmatrix}\text{promedio del}\\ \text{nuevo tratamiento}\end{smallmatrix}\right) - \left(\begin{smallmatrix}\text{promedio del}\\ \text{tratamiento estándar}\end{smallmatrix}\right).$$

El valor observado del estadístico es entonces

$$X_{\text{obs}} = 23 - 32 = -9.$$

Notar que los tiempos de recuperación no son aleatorios, porque los pacientes no fueron elegidos al azar, solo su asignación a los tratamientos es aleatoria. Por lo tanto, la base para construir una distribución de probabilidad para X proviene de la asignación aleatoria de los pacientes en los tratamientos.

Esta asignación al azar da como resultado que n pacientes reciban el nuevo tratamiento, y m reciban el tratamiento estándar, pero esta es solo una de las $\binom{N}{n}$ asignaciones igualmente probables que podrían haber ocurrido.

La distribución de X, llamada distribución de aleatorización, puede determinarse calculando X para cada una de las posibles asignaciones. La Tabla 2 enumera las $\binom{7}{4} = 35$ asignaciones posibles de los siete tiempos de recuperación observados en nuestro caso concreto, y la diferencia en los promedios para cada asignación. La probabilidad de cualquiera de estas asignaciones es $1/35$. La Figura 2 muestra la distribución de aleatorización de la diferencia de promedios X, con la diferencia observada marcada con un triángulo amarillo.

El p-valor es la probabilidad de obtener una diferencia de promedios tanto o más extrema que la observada. Esto es

$$\text{p-valor} = \mathbf{P}\left(X \leq X_{\text{obs}}\right).$$

Tabla 2: Todas las asignaciones posibles de los tiempos de recuperación (en días) en dos grupos de tratamiento de tamaños $n = 4$ y $m = 3$.

Nº	Tratamiento nuevo				Tratamiento esándar			Diferencia de promedios	Suma de nuevo	Suma de estándar	Diferencia de medianas
1	19	22	25	26	23	33	40	-9.00	92	96	-9.5
2	22	23	25	26	19	33	40	-6.67	96	92	-9.0
3	22	33	25	26	19	23	40	-0.83	106	82	2.5
4	22	25	26	40	19	23	33	3.25	113	75	2.5
5	19	23	25	26	22	33	40	-8.42	93	95	-9.0
6	19	25	26	33	22	23	40	-2.58	103	85	2.5
7	19	25	26	40	22	23	33	1.50	110	78	2.5
8	19	22	23	26	25	33	40	-10.17	90	98	-10.5
9	19	22	26	33	23	25	40	-4.33	100	88	-1.0
10	19	22	26	40	23	25	33	-0.25	107	81	-1.0
11	19	22	23	25	26	33	40	-10.75	89	99	-10.5
12	19	22	25	33	23	26	40	-4.92	99	89	-2.5
13	19	22	25	40	23	26	33	-0.83	106	82	-2.5
14	23	25	26	33	19	22	40	-0.25	107	81	3.5
15	22	23	26	33	19	25	40	-2.00	104	84	-0.5
16	22	23	25	33	19	26	40	-2.58	103	85	-2.0
17	19	23	26	33	22	25	40	-3.75	101	87	-0.5
18	19	23	25	33	22	26	40	-4.33	100	88	-2.0
19	19	22	23	33	25	26	40	-6.08	97	91	-3.5
20	23	25	26	40	19	22	33	3.83	114	74	3.5
21	22	23	26	40	19	25	33	2.08	111	77	-0.5
22	22	23	25	40	19	26	33	1.50	110	78	-2.0
23	19	23	26	40	22	25	33	0.33	108	80	-0.5
24	19	23	25	40	22	26	33	-0.25	107	81	-2.0
25	19	22	23	40	25	26	33	-2.00	104	84	-3.5
26	25	26	33	40	19	22	23	9.67	124	64	7.5
27	22	26	33	40	19	23	25	7.92	121	67	6.5
28	22	25	33	40	19	23	26	7.33	120	68	6.0
29	19	26	33	40	22	23	25	6.17	118	70	6.5
30	19	25	33	40	22	23	26	5.58	117	71	6.0
31	19	22	33	40	23	25	26	3.83	114	74	2.5
32	23	26	33	40	19	22	25	8.50	122	66	7.5
33	23	25	33	40	19	22	26	7.92	121	67	7.0
34	22	23	33	40	19	25	26	6.17	118	70	3.0
35	19	23	33	40	22	25	26	4.42	115	73	3.0

En nuestro caso, solamente 3 asignaciones dieron como resultado una diferencia de promedios X menor o igual a X_{obs}. Luego el p-valor es en este caso

$$\text{p-valor} = \frac{3}{35} = 0.0857.$$

Es decir, bajo el supuesto de que el tratamiento no tiene efecto sobre los tiempos de recuperación de los pacientes, vemos que solamente el $8.57\,\%$ de las asignaciones tienen un valor tanto o más extremo del estadístico que el observado. Podemos decir entonces que hay evidencia moderada en contra de nuestro supuesto.

Podríamos, por supuesto, elegir un estadístico diferente de X para medir la efectividad del nuevo tratamiento, como la diferencia en las medianas de cada grupo. La distribución de aleatorización de la diferencia en las medianas de cada grupo da como resultado el mismo p-valor de $3/35$.

También podríamos usar la suma de los tiempos de recuperación en uno de los grupos de tratamiento como estadístico. Los valores pequeños de la suma en el tratamiento nuevo o los valores grandes de la suma en el tratamiento estándar indicarían una mejoría en los tiempos de recuperación con el nuevo tratamiento. Estos también se muestran en la Tabla 2.

Si llamamos $x_i^{(1)}$ con $i = 1, \ldots, n$, a los tiempos de recuperación de los pacientes en el grupo

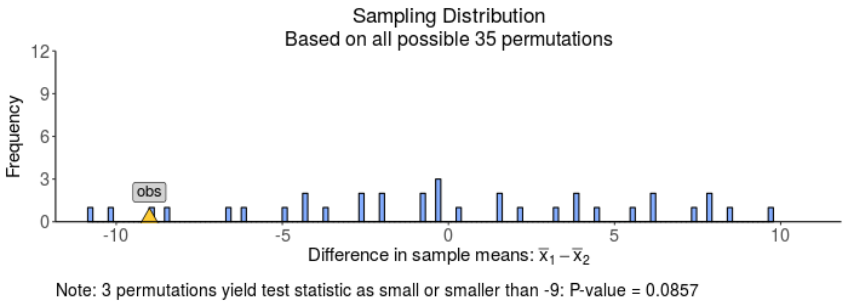


Figura 2: Distribución de aleatorización de X, basada en las 35 asignaciones posibles. Con un triángulo amarillo se muestra el valor de X observado $X_{\text{obs}} = -9$. Notar que solamente 3 asignaciones dieron un valor de X menor o igual a X_{obs}. Por lo que el p-valor es $3/35 = 0.857$.

de tratamiento nuevo, y $x_j^{(2)}$ con $j = 1, \ldots, m$ a los tiempos de recuperación de los pacientes en el grupo de tratamiento estándar, podemos reescribir la diferencia de promedios X como

$$X = \bar{x}_1 - \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{(1)} - \frac{1}{m}\sum_{j=1}^{m} x_j^{(2)} = \frac{m+n}{mn}\sum_{1=1}^{n} x_i^{(1)} - \frac{1}{m}\left(\sum_{i=1}^{n} x_i^{(1)} + \sum_{j=1}^{m} x_j^{(2)}\right)$$
$$= \frac{m+n}{mn}\sum_{1=1}^{n} x_i^{(1)} - \frac{s}{m}$$

en donde $s = \sum_{i=1}^{n} x_i^{(1)} + \sum_{j=1}^{m} x_j^{(2)}$. Como la suma de los dos grupos es siempre la misma, $s = 188$, vemos que X es una función monótona de la suma $\sum_{1=1}^{n} x_i^{(1)}$. Esto implica que el orden de los estadísticos en las 35 asignaciones es igual para ambos. Por lo tanto, la suma de los tiempos de recuperación en uno de los grupos es un estadístico equivalente a la diferencia de promedios. Ambos dan el mismo p-valor.

Debido a su equivalencia, la suma de los tiempos de un grupo de tratamiento a menudo se utiliza en lugar de la diferencia de promedios, ya que es computacionalmente más eficiente. Esto importante ya que la cantidad de asignaciones $\binom{N}{n}$ crece rápidamente a medida que m y n aumentan.

2. Diferencia entre tratamientos II

El supuesto que hemos hecho, el tratamiento no tiene efecto, y que hemos descartado como poco probable, se conoce en estadística como *hipótesis nula*. Se la suele escribir como H_0. Es decir, en nuestro ejemplo hemos puesto a prueba la hipótesis nula

$$H_0 : \text{el tratamiento no tiene efecto.}$$

Nuestro razonamiento se basó en el razonamiento por improbable. Asumiendo H_0, hemos calculado la probabilidad de observar algo tanto o más extremo que lo observado. Obtuvimos que esta probabilidad es 0.0857. Como juzgamos dicha probabilidad bastante chica, dedujimos que hay evidencia suficiente para *rechazar* H_0.

Que una probabilidad sea pequeña o no es bastante subjetivo. Un procedimiento es el de tomar un valor umbral p_u, aceptado por todos, para el cual probabilidades más chicas que p_u

sean consideradas chicas, y probabilidades más grandes que p_u sean consideradas grandes. Es común usar como valor umbral $p_u = 0.05$ o $p_u = 0.1$. En nuestro ejemplo, si usamos el valor $p_u = 0.05$ obtenemos una probabilidad grande, y debemos concluir que no hay evidencia para rechazar H_0. Si usamos el valor $p_u = 0.1$, deducimos que si la hay.

Convengamos en usar para este ejemplo el umbral $p_u = 0.1$. Entonces tenemos evidencia para rechazar la hipótesis de que el tratamiento no tiene efecto.

Si estamos dispuestos a suponer que el nuevo tratamiento tiene un efecto aditivo constante comparado con el tratamiento estándar, podemos estimar este efecto del tratamiento invirtiendo el método que usamos hasta ahora.

Supongamos que el nuevo tratamiento reduce los tiempos de recuperación en Δ días. Entonces, si cambiamos los tiempos de recuperación del grupo con tratamiento nuevo por

$$x_1^{(1)} + \Delta, \ldots, x_n^{(1)} + \Delta,$$

estos tiempos cambiados deberían ser similares en magnitud a los tiempos del tratamiento estándar, $x_1^{(2)}, \ldots, x_m^{(2)}$, y el razonamiento por improbable basado en estos dos conjuntos de tiempos no debería rechazar H_0.

Por ejemplo, supongamos que el nuevo tratamiento reduce el tiempo de recuperación en $\Delta = 1$ día. Si sumamos 1 a todos los tiempos del grupo del tratamiento nuevo, tenemos los tiempos:

Tratamiento **nuevo** $+1$				Tratamiento **esándar**		
20	23	26	27	23	33	40

Si calculamos el p-valor para estos nuevos tiempos, obtenemos

$$\text{p-valor} = \frac{4}{35} = 0.1143.$$

Como es mayor que 0.1, deducimos que no hay evidencia para rechazar H_0. Es decir, es plausible que Δ sea igual a 1.

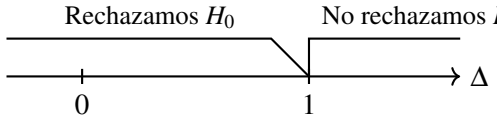
Probemos con $\Delta = 0.5$. En este caso los tiempos son:

Tratamiento **nuevo** $+0.5$				Tratamiento **esándar**		
19.5	22.5	25.5	26.5	23	33	40

y el p-valor es nuevamente 0.0857, por lo que rechazamos H_0. De hecho, se puede ver que para todo $\Delta < 1$ el p-valor es menor que el umbral 0.1.

El procedimiento consiste entonces en calcular la distribución del estadístico X_Δ para cada Δ, el nuevo valor observado $(X_\Delta)_{\text{obs}}$, y a partir de esto calcular el nuevo p-valor $\text{pval}(\Delta)$. Si conocemos la función $\Delta \mapsto \text{pval}(\Delta)$ (típicamente se calcula usando la computadora), podemos deducir cuáles son los valores de Δ para los cuales no rechazamos H_0. Ver la Figura 3.

Esto significa que tenemos una situación como la que se muestra en el siguiente diagrama:



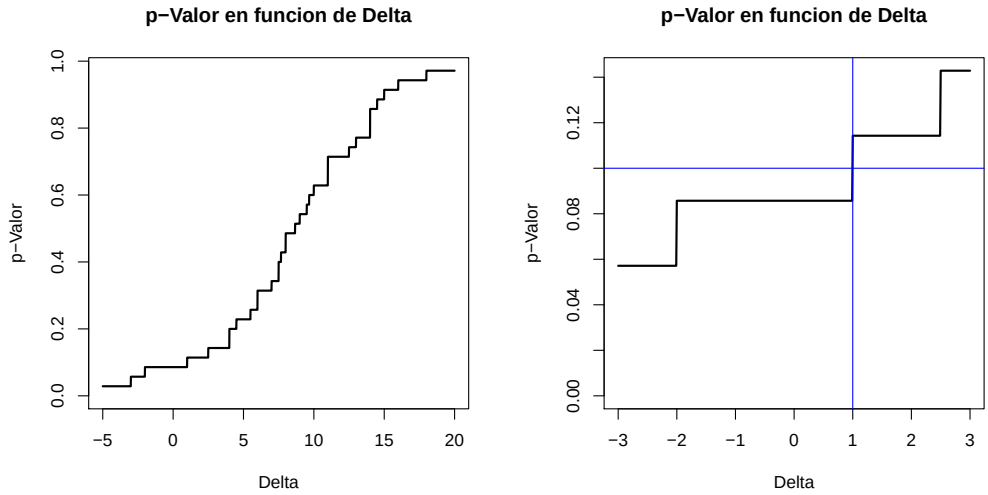


Figura 3: A la izquierda se muestra el gráfico de la función $\Delta \mapsto \text{pval}(\Delta)$. A la derecha se muestra el valor de $\Delta = 1$ en donde el p-valor pasa de ser menor que 0.1 (se rechaza H_0) a ser mayor (no se rechaza H_0).

Esto lo podemos interpretar del siguiente modo. Cualquier valor de $\Delta \geq 1$ es compatible con el supuesto de que los tiempos de recuperación del grupo con el tratamiento nuevo se reducen en Δ días en relación a los tiempos del grupo con el tratamiento estándar. Podemos concluir que el nuevo tratamiento reduce el tiempo de recuperación en al menos 1 día.

3. Test de significancia

Todos los ejemplos que hemos visto entran dentro de un mismo contexto teórico. Se trata de un método de inferencia que se llama test de significancia. A su vez, los test de significancia son casos particulares de lo que se conoce en estadística como test de hipótesis, o prueba de hipótesis.

Existen esencialmente dos modelos que estudiaremos en nuestro curso. El modelo de tratamientos y el modelo de población:

- En el modelo de tratamientos, lo que se busca es comparar dos tipos distintos de tratamientos, primero saber si existe una diferencia entre ellos y segundo estimar cuantitativamente esta diferencia.
- En los modelos de población, se busca inferir alguna propiedad de una población a partir de una muestra de la misma. Ejemplos típicos son las encuestas, en las cuales queremos estimar el número de votantes por un determinado candidato en una población dada, a partir del número de votantes por ese candidato en una muestra de la misma.

Ambos modelos se aplican a una gran variedad de situaciones, para las cuales deben hacerse ciertas suposiciones. No es posible hacer inferencia estadística sin hacer suposiciones. La famosa frase “dejar que los datos hablen por sí solos” no es aplicable cuando queremos inferir (aprender) propiedades sobre un tratamiento o una población a partir de un análisis estadístico.

En una primera aproximación, podemos dividir los estudios estadísticos en dos clases. Los experimentos controlados y los estudios observacionales. Por ejemplo, en un estudio que compara dos tratamientos distintos, si el investigador asigna de forma aleatoria y controlada los individuos a los diferentes tratamientos, se trata de un experimento controlado. Si en cambio, el investigador no puede realizar esta asignación, como por ejemplo determinar qué individuos fuman y cuáles no, se trata de un estudio observacional. Del mismo modo ocurre con los modelos de población, si en una encuesta la muestra se elige al azar de forma controlada, se trata de un experimento, pero si la muestra consiste de personas que pasan por una determinada esquina a una determinada hora, se trata de un estudio observacional. Nosotros estudiaremos solo experimentos controlados.

Los ejemlos anteriores que hemos visto son todos ejemplos de experimentos controlados en el modelo de tratamiento. Los modelos de población los estudiaremos más adelante.

Modelo de tratamientos

El marco teórico general de los modelos de tratamientos es el siguiente. Se desea comparar dos tratamientos, que por comodidad los llamaremos *tratamiento* y *control*. Cada individuo de un grupo de N individuos es asignado al azar para tratamiento o control. La asignación puede ser hecha de modo que las cantidades de individuos en cada grupo sean diferentes.

Para evaluar el efecto de los tratamientos, el investigador mide una *respuesta* a cada individuo. Por ejemplo, en la sección anterior la respuesta se midió en los días de recuperación de cada paciente.

En el modelo de tratamientos se asume que cada individuo tiene una respuesta potencial para cada tratamiento. Denotamos r_i^T la respuesta potencial del individuo i al tratamiento, y r_i^C la respuesta potencial del individuo i al control. Los valores de r_i^T y r_i^C se suponen fijos, y no aleatorios. El problema principal de la inferencia en estos modelos es que el investigador no puede medir las dos respuestas, si no solo una de ellas. Es decir, solo una de las dos respuestas es observable.

Si ambas respuestas fueran observables, el problema estaría resuelto. Basta compararlas entre ellas para deducir el efecto del tratamiento. Sin embargo, como solo una de ellas es observable, el problema inferencial de los modelos de tratamiento se basa en "adivinar" la respuestas no observadas, en base a las respuestas sí observadas.

En general se desea poner a prueba la siguiente hipótesis nula

$$H_0: \text{ El tratamiento no tiene ningún efecto.}$$

Dicho así es un poco vago. Lo que esta hipótesis quiere decir, es que la respuesta de un individuo no depende del tratamiento que reciba. Es decir, podemos escribir de forma más precisa la hipótesis nula de la siguiente manera

$$H_0: \ r_i^T = r_i^C \text{ para todo } i = 1, \ldots, N.$$

El objetivo es poner a prueba H_0. Es decir, suponer que es verdadera, y a partir de esto, calcular la probabilidad de observar algo tanto o más extremo que lo observado.

Notar que si asumimos H_0, entonces de hecho conocemos todas las respuestas, pues de cada individuo hemos observado una de ellas.

En estos modelos el azar proviene de la asignación aleatoria de los individuos a los diferentes grupos, y no de los valores de las respuestas potenciales. Las respuestas potenciales están fijas, lo que es aleatorio es la asignación de cada individuo, y por ende qué respuesta potencial observamos.

Para poner a prueba H_0 la idea es crear un conjunto de *realidades hipotéticas* de todos los resultados que pudieron haber ocurrido, aunque en la realidad hayamos observado uno solo de ellos.

Por ejemplo, supongamos que en nuestro estudio hay $N = 5$ individuos 1, 2, 3, 4, 5, divididos en dos grupos de $n = 3$ (tratamiento) y $m = 2$ (control). Supongamos que los individuos del grupo de control son el 2 y el 5. Entonces, lo que observamos son los valores de respuesta $r_1^T, r_2^C, r_3^T, r_4^T, r_5^C$. Sin embargo, hay $\binom{5}{2} = 10$ asignaciones posibles, por lo que el conjunto de realidades hipotéticas tiene las 10 posibles realidades que hubiéramos podido observar:

Realidad **observada**					Realidad **hipotética**				
T	*C*	*T*	*T*	*C*	*T*	*T*	*T*	*C*	*C*
1	2	3	4	5	1	2	3	4	5
					T	*T*	*C*	*T*	*C*
					1	2	3	4	5
					T	*T*	*C*	*C*	*T*
					1	2	3	4	5
					T	*C*	*T*	*C*	*T*
					1	2	3	4	5
					T	*C*	*T*	*T*	*C*
					1	2	3	4	5
					C	*T*	*T*	*T*	*C*
					1	2	3	4	5
					C	*T*	*T*	*C*	*T*
					1	2	3	4	5
					T	*C*	*C*	*T*	*T*
					1	2	3	4	5
					C	*T*	*C*	*T*	*T*
					1	2	3	4	5
					C	*C*	*T*	*T*	*T*
					1	2	3	4	5

Notar que los tamaños de los dos grupos se mantienen constantes, el de tratamiento con $n = 3$ individuos y el de control con $m = 2$ individuos. Simplemente permutamos los tratamientos.

Cada una de estas asignaciones tiene la misma probabilidad de ocurrir que las demás, por lo que la probabilidad de cada una de ellas es $1/\binom{N}{n}$. Si la hipótesis nula es verdadera, en cualquiera de estas asignaciones hubieramos observado las mismas respuestas que observamos realmente.

El siguiente paso es elegir un *estadístico* X. Un estadístico es una función que a cada asignación asocia un número real. Este debe ser calculable a partir de las respuestas de los individuos. Estadísticos comunes son por ejemplo la diferencia de promedios, la suma de las respuestas de un grupo, la diferencia de medianas, o cualquier otra cantidad calculable a partir de las respuestas potenciales.

La hipótesis nula garantiza que podamos calcular el valor de X para cualquiera de las realidades hipotéticas (asignaciones). Entre estas se encuentra las realidad observada, y el valor observado de X lo denotamos por X_{obs}.

Para rechazar o no rechazar H_0 debemos calcular el p-valor asociado al estadístico X. Para esto, se debe tener en cuenta el motivo con el cual fue diseñado el experimento. Debemos distinguir dos casos:

1. Si no se tiene un juicio a priori sobre que el tratamiento sea más o menos efectivo que el control, se toma como convención la siguiente definición de p-valor:

$$\text{pval}\,(X_{\text{obs}}) = 2\,\text{mín}\{\mathbf{P}\,(X \leq X_{\text{obs}})\,, \mathbf{P}\,(X \geq X_{\text{obs}})\}$$

   En este caso decimos que se trata de un p-valor a dos colas.

2. Si por conocimientos a priori se sabe que el tratamiento no es peor que el control, y el objetivo del experimento es "demostrar" que es es más efectivo, convenimos en calcular el p-valor como:

$$\text{pval}\,(X_{\text{obs}}) = \mathbf{P}\,(X \leq X_{\text{obs}}) \quad \text{o} \quad \text{pval}\,(X_{\text{obs}}) = \mathbf{P}\,(X \geq X_{\text{obs}})$$

   dependiendo de si los valores grandes o pequeños de X indican una mejor eficacia del tratamiento. En este caso decimos que se trata de un p-valor a una cola.

Una forma de pensar sobre si debemos tomar un p-valor a dos colas o a una cola es considerar la *hipótesis alternativa* H_1. Esta consiste simplemente en la negación de H_0. En general, si

$$H_0 :\ r_i^T = r_i^C \text{ para todo } i = 1, \ldots, N;$$

entonces

$$H_1 :\ r_i^T \neq r_i^C \text{ para algún } i = 1, \ldots, N.$$

Esto quiere decir que $r_i^T < r_i^C$ o $r_i^T > r_i^C$ para algún i. Si no tenemos ninguna información adicional sobre la relación entre las respuestas potenciales, debemos considerar las dos desigualdades como posibles. En este caso el p-valor es a dos colas. Pero si por alguna razón sabemos que $r_i^T \geq r_i^C$ siempre, entonces la alternativa consiste solamente de $r_i^T > r_i^C$. En este caso debemos considerar un p-valor a una cola. Disponer o no de esa información adicional no es parte del test, es un supuesto adicional en caso de que exista. Más adelante en el curso vamos a profundizar sobre el rol de la hipótesis alternativa.

En cualquiera de los dos casos, la probabilidad $\mathbf{P} = \mathbf{P}_{H_0}$ es la probabilidad suponiendo H_0 verdadera, para la cual todas las asignaciones posibles (realidades hipotéticas) son igualmente probables.

Si juzgamos que el p-valor es pequeño, entonces rechazamos H_0. Sin embargo, si juzgamos el p-valor moderadamente grande, entonces debemos concluir que no hay evidencia suficiente para rechazar H_0.

Debido a la subjetividad al juzgar probabilidades grandes o pequeñas, es común definir un valor umbral p_u para el cual:

1. Si $\text{pval}\,(X_{\text{obs}}) \leq p_u$, entonces rechazamos H_0;
2. Si $\text{pval}\,(X_{\text{obs}}) \geq p_u$, entonces no rechazamos H_0.

En general se usa como umbral $p_u = 0.05$ o $p_u = 0.1$.

El rol de la asignación aleatoria

Dos científicos deben realizar un experimento para comparar un tratamiento, T, que se cree que mejora el rendimiento, con un control C. Se deben usar cuatro unidades, dos para T y C. Las seis asignaciones posibles de T y C se enumeran en la primera columna de la Tabla 3.

El primer científico, A, decide seleccionar uno de los seis diseños al azar. El segundo científico, B, considera que el primer y el último diseño serían insatisfactorios, porque todos los tratamientos y todos los controles están juntos, y por lo tanto selecciona un diseño al azar de los cuatro restantes.

Tanto A como B llevan a cabo sus respectivas asignaciones y ambos obtienen el diseño TCTC, segunda fila de la tabla. En la implementación del experimento, ambos científicos obtienen las respuestas 5, 4, 3, 2 que se muestran en la última fila de la tabla. La suma de las respuestas en el tratamiento es 8, y en el control 6, y el efecto se mide por la diferencia, 2.

Hasta acá los científicos están de acuerdo, pero ahora veamos qué pasa si usan el argumento de aleatorización para el análisis.

Si las respuestas observadas son las mismas que las de cualquier otro diseño que pudiera haberse utilizado, las diferencias habrían sido las enumeradas en la segunda columna de la Tabla 3.

Consideremos al científico B primero. El científico B excluyó el primero y el último diseño, por lo que las posibles diferencias son $(2, 0, 0, -2)$, de las cuales la primera, la que realmente se obtiene, es la más grande. Por lo tanto, el resultado es significativo al 25 por ciento, porque todos los diseños tienen el mismo 25 por ciento de posibilidades de ser utilizados.

Tabla 3: Asignaciones posibles para los dos científicos

		Asignación	Diferencia
A		TTCC	4
	B	TCTC	2
		TCCT	0
		CTTC	0
		CTCT	-2
		CCTT	-4
Respuestas		5432	

El científico A, sin embargo, incluyó el primero y el último diseño en la aleatorización, por lo que debe incluir las diferencias 4 y -4 que podrían haber surgido al usarlos. De las seis diferencias, 4 es la más grande y 2, la que realmente se observa, la siguiente más grande. Por lo tanto, la probabilidad de diferencias más extremas o iguales a la diferencia observada, es 2 en 6 y el resultado es significativo en 33.3 por ciento.

Entonces, hay dos científicos que han realizado exactamente el mismo experimento, TCTC, obtuvieron exactamente el mismo resultado, y aún así uno obtiene un nivel de significancia superior al otro. Y la razón de esta diferencia de nivel es que A contemplaba realizar experimentos que B no (es decir, aquellos en la primera y la última fila de la tabla), aunque, de hecho, A no realizó uno de estos experimentos. Expresado de manera ligeramente diferente,

el análisis de los resultados del experimento dependió de lo que podría haberse hecho, pero en realidad no se hizo (las realidades hipotéticas).

Aunque esto parezca raro, se puede entender razonando que está bien que A y B argumenten de forma diferente porque B pensó que el primero y el último experimento podrían ser insatisfactorios, mientras que A no lo hizo. En otras palabras, ambos científicos tenían ideas diferentes antes del experimento; ¿No es razonable que los dos científicos tengan ideas diferentes después? Este argumento viola la afirmación que a menudo se hace de las pruebas de significancia, que dicen permitir que los datos hablen por sí mismos y no se vean afectadas por consideraciones ajenas a los datos.

Nosotros siempre supondremos que la asignación aleatoria se efectúa entre todas las asignaciones posibles.

4. Errores de tipo I y II

Otra forma de definir el p-valor es a través de el error de tipo I. Un error de tipo I se produce cuando rechazamos H_0 pero H_0 es cierta. Existe otro error, el error de tipo II que se produce cuando no rechazamos H_0 pero H_0 es falsa. La Tabla 4 muestra ambos errores.

Tabla 4: Cuadro de decisiones y errores.

		Decisión	
		Rechazamos H_0	No rechazamos H_0
Realidad	H_0 cierta	Error de tipo I	Correcto
	H_0 falsa	Correcto	Error de tipo II

Decidir entre si rechazar H_0 o no, se puede hacer mediante el uso de *regiones de rechazo*. Esto es, definimos una región crítica I_r, que por lo general es una unión de intervalos de la recta real.

De este modo, la regla de decisión es que si el valor observado del estadístico X cae en la regíon de rechazo, rechazamos H_0, y si cae afuera no rechazamos H_0:

1. Si $X_{\text{obs}} \in I_r$, entonces rechazamos H_0;
2. Si $X_{\text{obs}} \notin I_r$, entonces no rechazamos H_0.

El error de tipo I se puede escribir entonces como

$$\boxed{\alpha = \mathbf{P}(X \in I_r | H_0)}$$

Se suele usar por convención la letra α para el error de tipo I, y β para el error de tipo II. Nosotros no usaremos por ahora los errores de tipo II, los estudiaremos más adelante al profundizar el rol de la hipótesis alternativa.

La metodología consiste entonces en buscar una región de rechazo que tenga un error de tipo I pequeño, por ejemplo menor o igual al valor umbral p_u. Pero es muy importante tener presente que el valor de α se debe fijar de ante mano.

Supongamos que los valores chicos del estadístico X indican una mejor eficacia del grupo de tratamiento. Pensar en el ejemplo de los tiempos de recuperación, en donde menores tiempos de recuperación significan mejor rendimiento del tratamiento. Otros casos se pueden analizar de forma análoga.

En este caso, es razonable tomar como región de rechazo un intervalo de la forma $I_r(c) = (-\infty, c]$. Es decir, rechazamos H_0 si el estadístico observado es suficientemente chico: $X_{\text{obs}} \leq c$. Una vez fijado el valor de α, podemos calcular c resolviendo la ecuación

$$\alpha = \mathbf{P}(X \in I_r(c)|H_0) = \mathbf{P}(X \leq c|H_0).$$

Esto define el valor de c sin ambigüedad si el valor de α es un valor posible para la función

$$c \mapsto \mathbf{P}(X \leq c|H_0).$$

Si esto no es así, convenimos en tomar c como el real más grande que verifica la desigualdad

$$\mathbf{P}(X \leq c|H_0) \leq \alpha,$$

de forma tal de asegurarnos que el error de tipo I sea menor que el valor de α que hemos elegido antes.

Para simplificar la exposición, supongamos que la función $c \mapsto \mathbf{P}(X \leq c|H_0)$ es continua, aunque esto no es cierto en los casos que nos interesan. Cuando $c \to -\infty$ la función tiende a 0, y cuando $c \to +\infty$ la función tiende a 1. Ver la Figura 4.

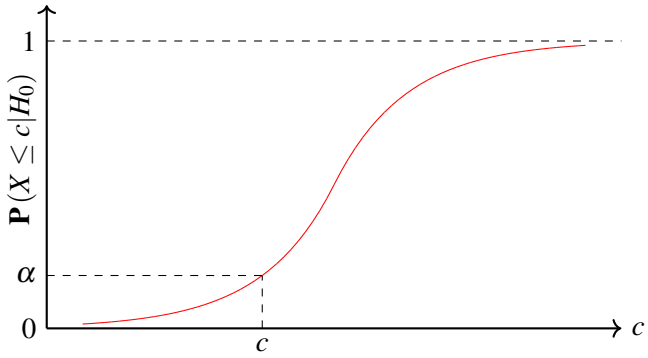


Figura 4: Determinación de c a partir del valor de α elegido.

Notar que si evaluamos la función $c \mapsto \mathbf{P}(X \leq c|H_0)$ en el valor observado del estadístico X_{obs}, obtenemos el p-valor $\text{pval}(X_{\text{obs}})$:

$$\text{pval}(X_{\text{obs}}) = \mathbf{P}(X \leq X_{\text{obs}}|H_0).$$

En la Figura 5 se muestra la relación entre el p-valor, el error de tipo I y la región de rechazo.

Vemos así que α juega un rol similar al valor umbral p_u, pues podemos escribir la regla de decisión como:

1. Si $\text{pval}(X_{\text{obs}}) \leq \alpha$, entonces rechazamos H_0;
2. Si $\text{pval}(X_{\text{obs}}) > \alpha$, entonces no rechazamos H_0.

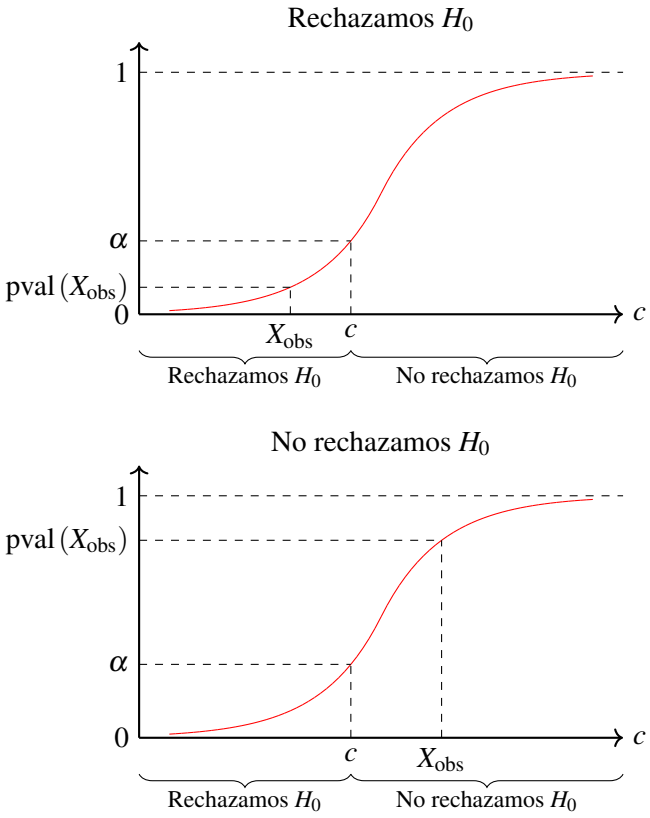


Figura 5: Relación entre pval (X_{obs}), α y c.

De este modo podemos redefinir el p-valor como

$$\text{pval}\,(X_{\text{obs}}) = \text{mín}\,\{\alpha \in (0,1) : \text{ no rechazamos } H_0\}\,.$$

Esta forma de pensar la regla de decisión se adapta mejor a pruebas de significancia que se replican muchas veces a lo largo de un cierto tiempo, como por ejemplo en el control de calidad. Esto es así pues el valor de α es una tasa de error de tipo I. Dicho de otro modo, adoptando la regla de decisión anterior, cometeremos un error de tipo I solamente en el $100\alpha\,\%$ de las pruebas.

En nuestro contexto de modelos de tratamiento, es mejor interpretar el p-valor como evidencia en contra de la hipótesis nula, pues este tipo de experimentos no se repiten. Un p-valor pequeño indica evidencia fuerte en contra de H_0.

No confundir el p-valor pval (X_{obs}) con α el error de tipo I. El primero depende de los datos observados y cambia de valor si los datos son diferentes. Se debe interpretar como evidencia en contra de H_0. El segundo es una tasa de error, que se fija de antemano. Indica la frecuencia de veces que cometeremos un error al rechazar H_0 en varias repeticiones del experimento.

5. Estimación de la diferencia entre dos tratamientos

Los métodos que hemos desarrollado también sirven para estimar la diferencia entre los dos tratamientos. La idea es muy simple: si el tratamiento mejora la respuesta en una cierta can-

tidad Δ, es decir, si $r_i^T = r_i^C \pm \Delta$ para todo $i = 1, \ldots, N$, entonces cuando restamos/sumamos Δ a las respuestas observadas de quienes recibieron tratamiento, obtendremos dos grupos cuyas respuestas son similares. En ese caso, la prueba de significancia no debería rechazar H_0. Recíprocamente, si al restar/sumar Δ a las respuestas de quienes recibieron tratamiento, la prueba no rechaza H_0, es porque las respuestas de ambos grupos son similares y por lo tanto las respuestas originales difieren de las de control en la cantidad Δ.

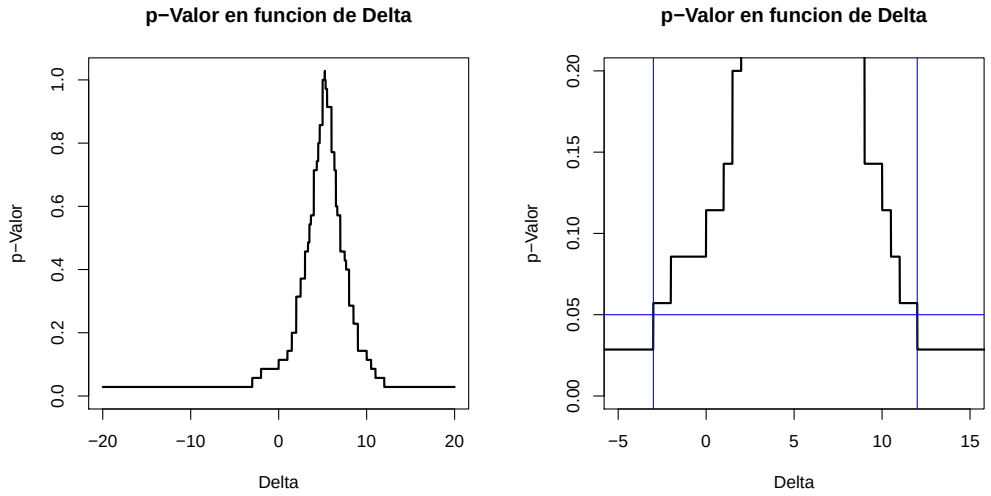


Figura 6: A la izquierda se muestra el gráfico de la función $\Delta \mapsto \text{pval}(\Delta)$. A la derecha se muestra el intervalo $\Delta \in [-3, 12]$ para el cual el p-valor es mayor a 0.05, y por lo tanto no se rechaza H_0.

Veamos como funciona la idea con un ejemplo en concreto. Supongamos que un agricultor está interesado en saber cuál fertilizante es más efectivo para el crecimiento de un cierto cultivo. Tiene dos opciones, el fertilizante A y el fertilizante B. Para decidir cuál comprar realiza un experimento controlado con cuatro plantas que reciben el fertilizante A y cuatro que reciben el fertilizante B. Sus resultados se muestran en la Tabla 5.

Tabla 5: Altura de plantas en cm.

Fertilizante A	Fertilizante B
21 26 29 31	24 23 19 20

Para determinar si hay una diferencia de efectos en las alturas de los fertilizantes, consideremos el estadístico

$$X = \bar{x}_A - \bar{x}_B$$

que es la diferencia de los promedios de las respuestas en cada grupo. De este modo, el valor observado por el agricultor es

$$X_{\text{obs}} = 26.75 - 21.5 = 5.25.$$

Notar que queremos captar diferencias grandes tanto positivas como negativas, pues no sabemos de antemano qué fertilizante es mejor. Por último, fijemos como umbral $\alpha = p_u = 0.05$.

Para calcular el p-valor, debemos calcular la distribución de X bajo H_0. Es decir, debemos considerar todas las asignaciones hipotéticas, para cada una de ellas calcular X, y registrar las frecuencias con que toma los diferentes valores. Cada una de las asignaciones es igualmente probable, y hay $\binom{8}{4} = 70$ asignaciones posibles.

En la Figura 7 se muestra la distribución de aleatorización de X. Notar que 4 de las asignaciones verifican que X es mayor o igual a 5.25, por lo que el p-valor *a dos colas* es $\text{pval}(5.25) = 8/70 = 0.1143$. Como es mayor que 0.05, no tenemos evidencia suficiente para rechazar H_0.

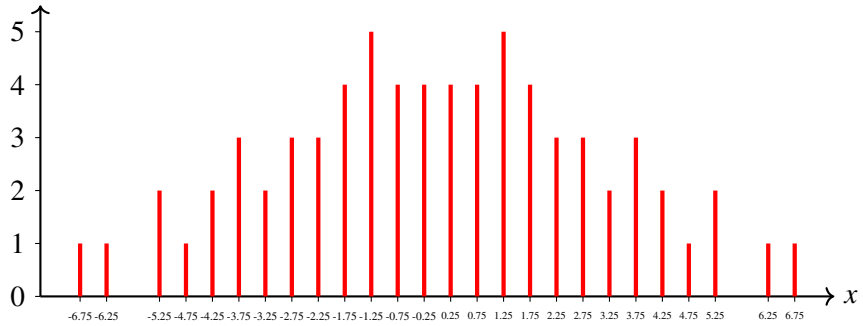


Figura 7: Distribución de aleatorización de X.

Pero ahora nos preguntamos, ¿cuán grande puede ser la diferencia entre ambos fertilizantes? Si restamos Δ a todas las respuesta del fertilizante A, tendremos un nuevo estadístico X_Δ y un nuevo valor observado $(X_\Delta)_{\text{obs}} = X_{\text{obs}} - \Delta$. Ahora realizamos el mismo test para ver si rechazamos o no H_0. Es decir, debemos calcular el nuevo p-valor $\text{pval}(\Delta)$ y ver cuándo es más chico que 0.05.

La Figura 6 muestra la función $\Delta \mapsto \text{pval}(\Delta)$. Para calcular esta función debemos usar una computadora, pues la cantidad de asignaciones (70) es demasiado grande para hacer las cuentas una por una. De la gráfica vemos que los valores de Δ para los cuales no rechazamos H_0 son los que están en el intervalo $\Delta \in [-3, 12]$. Este intervalo lo interpretamos de la siguiente manera: en relación al fertilizante B, el fertilizante A aumenta en a lo sumo 12 cm la altura de las plantas, y recíprocamente, en relación al fertilizante A, el fertilizante B aumenta en a lo sumo 3cm la altura de las plantas. Notar que este intervalo contiene al 0, lo cual es consistente con nuestra decisión anterior.

Se puede obtener este mismo resultado usando regiones de rechazo. Nuevamente, como los cálculos son tediosos usamos una gráfica hecha en la computadora como la que se muestra en la Figura 8. En la misma, se grafica la región de rechazo $I_r^\Delta = (-\infty, a(\Delta)] \cup [b(\Delta), +\infty)$ en función de Δ. La curva $\Delta \mapsto b(\Delta)$ es la curva superior, y $\Delta \mapsto a(\Delta)$ la inferior. Cuando trazamos una linea vertical sobre un valor de Δ, esta corta a las dos curvas del diagrama en los extremos de la región de rechazo . Luego, la zona entre las dos curvas consiste de los valores de Δ y valores observados para los cuales no rechazamos H_0. En el exterior de las curvas se rechaza H_0. Esto implica que si trazamos una linea horizontal sobre el valor X_{obs}, obtenemos el intervalo para estimar el efecto de los tratamientos. En nuestro ejemplo, obtenemos nuevamente el intervalo $\Delta \in [-3, 12]$.

Este mismo tipo de argumentos funciona en general, cuando queremos estimar la diferencia entre dos tratamientos.

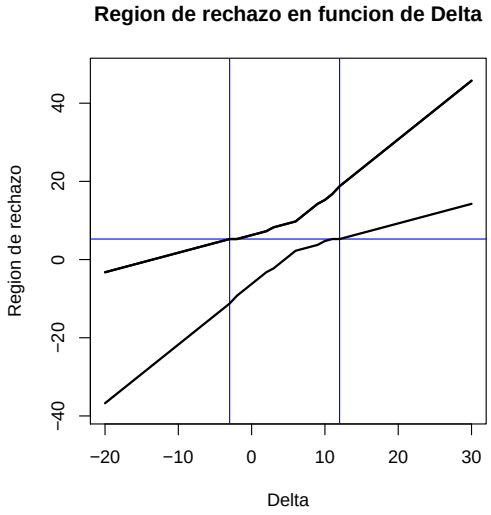


Figura 8: Diagrama que muestra la region de rechazo en función de Δ. Cortando con la linea horizontal a la altura del valor observado X_{obs} se obtiene el intervalo de valores de Δ para los cuales no se rechaza H_0. Notar que obtenemos el mismo intervalo que antes $\Delta \in [-3, 12]$.

6. Elegir bien el estadístico

En las pruebas de hipótesis se debe elegir de forma inteligente el estadístico a usar. Veamos un ejemplo.

Supongamos que un investigador arroja una moneda diez veces y asume una hipótesis nula de que la moneda es justa. El estadístico natural a usar es el número total de caras, y es razonable hacer un test a dos colas.

Supongamos que el investigador observa cruces y caras alternadamente

$$CXCXCXCXCX.$$

Esto da un estadístico observado igual a 5 y un p-valor igual a 1, ya que es el número esperado de caras.

Supongamos, en cambio, que el estadístico para este experimento es el "número de alternancias", es decir, el número de veces que X sigue a C o C sigue a X. En este caso, el valor observado es 9, y tiene un p-valor de

$$2/2^9 = 1/256 \approx 0.0039.$$

Esto se considera extremadamente significativo, mucho más chico que el umbral 0.05. El conjunto de datos es extremadamente improbable que haya ocurrido por casualidad, pero no sugiere que la moneda esté sesgada hacia caras o cruces.

Según el primer estadístico, los datos dan un p-valor grande, lo que sugiere que la cantidad de caras observadas no es improbable. Con el segundo estadístico, los datos dan un p-valor muy bajo, lo que sugiere que el patrón de alternancias observado es muy poco probable.

Este ejemplo muestra que el p-valor depende completamente del estadístico utilizado.