Clase 8: El test binomial y el test z

Matías Carrasco

6 de octubre de 2019

Índice

1. Generalidades sobre tests de hipótesis

Con los test de permutaciones hemos intentado analizar las situaciones en las que se desea comparar un nuevo tratamiento con un control o estándar. Hemos respondido a preguntas como: ¿Qué tipo de evidencia te convencería de que el nuevo tratamiento es mejor que el estándar? Esto nos llevo a la definición de p-valor, y de errores de tipo I y II.

Ahora continuaremos con el análisis de este tipo de problemas, pero desde una perspectiva un poco diferente, basada en el modelo de población. Estudiaremos por ejemplo, cómo responder a las siguientes preguntas: supongamos que queremos decidir si una moneda es justa. Si la lanzamos 100 veces y obtenemos 87 caras, ¿crees que es probable que la moneda sea sesgada? ¿Qué hay de 61 caras? ¿O 53 caras? La mayoría de la gente diría que 87 caras es evidencia fuerte de que la moneda es sesgada, mientras que 53 caras no constituye ninguna prueba. 61 caras es menos claro. Un test de hipótesis (TdH para abreviar, a veces llamado prueba de significación de hipótesis nula) es un enfoque frecuentista para responder cuantitativamente estas preguntas.

Comenzaremos enumerando los ingredientes de un TdH. Formalmente son bastante simples, pero hay un cierto arte en elegirlos bien. Es como el cocinero que elige los ingredientes para una receta; la receta puede ser simple y clara, pero si no tenemos experiencia es probable que no nos quede muy bien. Un TdH es como una receta, y exploraremos el arte de los TdH en ejemplos.

Ingredientes de un TdH

Son los mismo que para los Test de permutaciones (TdP), pero los modelos probabilísticos son distintos, ya que están basados en el modelo de población. Los repasamos:

- H_0: la hipótesis nula. Este es el supuesto por defecto para el modelo que genera los datos.
- H_A: la hipótesis alternativa. Si rechazamos la hipótesis nula, aceptamos esta alternativa como la mejor explicación para los datos.
- X: el estadístico de prueba. Calculamos esto a partir de los datos.
- *Distribución nula*: la distribución de probabilidad de X asumiendo H_0. En los TdP la llamábamos distribución de aleatorización, pues el azar provenía simplemente de la asignación aleatoria en grupos. Ahora los datos son producidos aleatoriamente, como en un muestreo.
- *Región derechazo*: si X está en la región de rechazo se rechaza H_0 a favor de H_A.
- *Región de no rechazo*: el complemento a la región de rechazo. Si X está en esta región no rechazamos H_0. Notar que decimos "no rechazar" en lugar de "aceptar" porque generalmente lo mejor que podemos decir es que los datos no prueban que H_0 es falsa.

La hipótesis nula H_0 y la hipótesis alternativa H_A desempeñan diferentes roles. Por lo general, elegimos que H_0 sea una hipótesis simple o por defecto (e.g. las diferencias observadas se deben simplemente al azar), que solo rechazaremos si tenemos pruebas suficientes contra ella.

Terminología de los TdH

En esta sección usaremos el ejemplo de la moneda para introducir y explorar la terminología utilizada en los TdH.

Para probar si una moneda es justa, la lanzamos 10 veces. Si obtenemos un número inesperado, grande o pequeño, de caras sospecharemos que la moneda es sesgada. Para esto elegimos los ingredientes del TdH de la siguiente manera. Sea θ la probabilidad de que la moneda salga cara.

1. Hipótesis nula H_0: "la moneda es justa", es decir, $\theta = 0.5$.
2. Hipótesis alternativa H_A: "la moneda es sesgada", es decir, $\theta \neq 0.5$.
3. Estadístico: $X =$ número de caras en 10 lanzamientos.
4. Distribución nula: es la función de probabilidad puntual basada en la hipótesis nula

$$p(x|\theta = 0.5); X \sim \text{Bin}(10, 0.5).$$

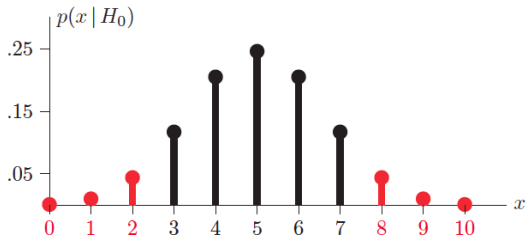
   La tabla que muestra la f.p.p. de X para la distribución nula es la siguiente

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Región de rechazo: bajo la hipótesis nula esperamos obtener alrededor de 5 caras en 10 lanzamientos. Rechazaremos H_0 si el número de caras es mucho menor o mayor que 5. Definimos la región de rechazo como $\{0, 1, 2, 8, 9, 10\}$. Es decir, si el número de caras en 10 lanzamientos está en esta región, rechazaremos la hipótesis de que la moneda es justa a favor de la hipótesis de que no lo es.

Podemos resumir todo esto en el gráfico y la tabla de probabilidad a continuación. La región de rechazo consiste de los valores de x en rojo, y las probabilidades correspondientes también están en rojo. La distribución nula se muestra en el gráfico, con los valores de rechazo de x en rojo.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001



La hipótesis nula es la opción cautelosa por defecto, no reclamaremos que la moneda sea sesgada a menos que tengamos pruebas convincentes. La región de rechazo consiste en datos que son extremos bajo la hipótesis nula. Es decir, consiste en los resultados que están en la cola de la distribución nula lejos del centro de mayor probabilidad. Como veremos pronto, cuán lejos depende del nivel de significación α del test.

Si obtenemos 3 caras en 10 lanzamientos, entonces el estadístico del test se encuentra en la región de no rechazo. En el lenguaje científico habitual equivale a decir que los datos"no apoyan el rechazo de la hipótesis nula". Incluso si obtuviéramos 5 caras, no diríamos que los datos prueban que la hipótesis nula es cierta.

Si la moneda justa, ¿cuál es la probabilidad de que decidamos incorrectamente que es sesgada? La hipótesis nula es que la moneda es justa. La pregunta equivale a calcular la probabilidad de que los datos de una moneda justa estén en la región de rechazo. Es decir, la probabilidad de que obtengamos 0, 1, 2, 8, 9 o 10 caras en 10 lanzamientos. Esta es la suma de las probabilidades en rojo. Es decir,

$$\mathbf{P}(\text{rechazar } H_0 | H_0 \text{ es cierta}) = 0.11.$$

A continuación, vamos a seguir analizando este ejemplo y definiremos más términos utilizados en los TdH.

Hipótesis simples y compuestas

Una *hipótesis simple* es una para la cual podemos especificar su distribución por completo. Una hipótesis simple típica es que un parámetro de interés toma un valor específico.

Una *hipótesis compuesta* es una para la cual no podemos especificar completamente su distribución. Una hipótesis compuesta típica es que un parámetro de interés se encuentre en un rango de valores.

En el ejemplo de la moneda, la hipótesis nula es que $\theta = 0.5$, por lo que la distribución nula es $\text{Bin}(10, 0.5)$. Dado que la distribución nula está completamente especificada, H_0 es simple. La hipótesis alternativa es que $\theta \neq 0.5$. Esto es realmente muchas hipótesis en una: θ podría ser 0.51, 0.7, 0.99, etc. Dado que la distribución binomial alternativa $\text{Bin}(10, \theta)$ no está completamente especificada, H_A es compuesta.

Si la hipótesis alternativa fuera

$$H_A : \theta = 0.7$$

entonces sí sería simple. En cambio, si fuera

$$H_A : \theta > 0.5$$

sería nuevamente compuesta.

Dos tipos de error

Al igual que en los TdP, hay dos tipos de errores que podemos cometer. Podemos rechazar (incorrectamente) la hipótesis nula cuando es verdadera o (incorrectamente) no podemos rechazarla cuando es falsa. El primero se llama Error de tipo I y el segundo Error de tipo II. Resumimos esto en la siguiente tabla.

Tabla 1: Cuadro de decisiones y errores.

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

Es decir:

- Error de tipo I: falso rechazo de H_0.
- Error de tipo II: falso no rechazo ("aceptación") de H_0.

En nuestro ejemplo de la moneda, el error de tipo I ocurre cuando juzgamos la moneda sesgada siendo justa, y el error de tipo II es cuando concluimos que no hay suficiente evidencia para probar que la moneda es sesgada pero sí lo es.

Nivel de significación y potencia

El nivel de significación y la potencia se utilizan para cuantificar la calidad del TdH. Lo ideal sería que un TdH no cometiera errores. Es decir, no rechazaría H_0 cuando H_0 fuera cierta, y rechazaría H_0 a favor de H_A cuando H_A fuera cierta. En total, hay cuatro probabilidades importantes que se corresponden con la tabla anterior de errores: Las dos probabilidades en

Tabla 2: Cuadro de decisiones y probabilidades de los errores.

		Decisión	
		Rechazamos H_0	No rechazamos H_0
Realidad	H_0 cierta	$\mathbf{P}(\text{rechazar } H_0 \mid H_0)$	$\mathbf{P}(\text{no rechazar } H_0 \mid H_0)$
	H_A cierta	$\mathbf{P}(\text{rechazar } H_0 \mid H_A)$	$\mathbf{P}(\text{no rechazar } H_0 \mid H_A)$

las que nos centramos son:

$$\begin{aligned}\text{Nivel de significación} &= \alpha = \mathbf{P}(\text{rechazar } H_0 | H_0) \\ &= \text{probabilidad de rechazar incorrectamente } H_0 = \mathbf{P}(\text{error de tipo I}). \\ \text{Potencia} &= \pi = \mathbf{P}(\text{rechazar } H_0 | H_A) \\ &= \text{probabilidad de rechazar correctamente } H_0 \\ &= 1 - \mathbf{P}(\text{error de tipo II}) = 1 - \beta.\end{aligned}$$

Idealmente, un TdH debería tener un nivel de significación pequeño (cerca de 0) y una potencia grande (cerca de 1). Pensemos en dos analogías para ayudarte a recordar los significados de significación y potencia:

1. Pensar en H_0 como la hipótesis de que "nada está pasando", es decir, "la moneda es justa", "el tratamiento no es mejor que el control", etc. Pensar en H_A como lo contrario: "algo interesante está sucediendo". Entonces la potencia es la probabilidad de detectar algo interesante cuando está presente, y el nivel de significación es la probabilidad de afirmar erróneamente que algo interesante ha ocurrido.

2. En los juicios, los acusados de delitos se consideran inocentes hasta que se demuestre su culpabilidad más allá de toda duda razonable. Podemos expresar esto en términos de un TdH como

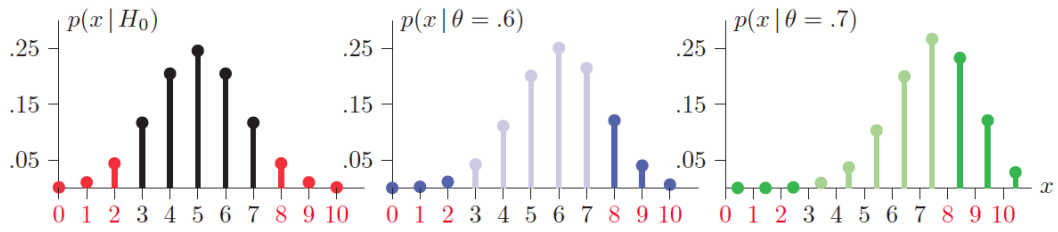
$$\begin{cases} H_0 : \text{el acusado es inocente (la opción por defecto)} \\ H_A : \text{el acusado es culpable.} \end{cases}$$

   El nivel de significancia es la probabilidad de juzgar a una persona inocente como culpable. La potencia es la probabilidad de juzgar correctamente a un culpable como culpable. "Más allá de toda duda razonable" significa que debemos exigir que el nivel de significación sea muy pequeño.

Cuando la hipótesis alternativa es compuesta, debemos considerar la potencia para cada posibilidad de la alternativa. En el ejemplo de la moneda H_A es compuesta, por lo que la potencia es diferente para diferentes valores de θ. Ampliemos la tabla de probabilidad anterior para incluir algunos valores alternativos de θ.

x	0	1	2	3	4	5	6	7	8	9	10
$H_0 : p(x\|\theta = .5)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$H_A : p(x\|\theta = .6)$	.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$H_A : p(x\|\theta = .7)$	.000	.000	.001	.009	.037	.103	.200	.267	.233	.121	.028

Notar que pensamos a la distribución alternativa como función de θ, y lo mismo para la potencia $\pi(\theta)$. En ese sentido estas probabilidades juegan un rol similar a la función de verosimilitud.



Usamos la tabla para calcular el nivel de significación y la potencia de cada θ.

$$\begin{aligned}\text{Nivel de significación} &= \text{probabilidad de rechazar } H_0 \text{ cuando es cierta}\\ &= \text{prob. de que } X \text{ esté en la región de rechazo cuando } H_0 \text{ es cierta}\\ &= \text{suma de cuadros rojos en la fila } \theta = .5\\ &= 0.11\end{aligned}$$

$$\begin{aligned}\text{Potencia cuando } \theta = .6 &= \text{probabilidad de rechazar } H_0 \text{ cuando } \theta = .6\\ &= \text{prob. de que } X \text{ esté en la región de rechazo cuando } \theta = .6\\ &= \text{suma de cuadros azul oscuro en la fila } \theta = .6\\ &= 0.18\end{aligned}$$

$$\begin{aligned}\text{Potencia cuando } \theta = .7 &= \text{probabilidad de rechazar } H_0 \text{ cuando } \theta = .7\\ &= \text{prob. de que } X \text{ esté en la región de rechazo cuando } \theta = .7\\ &= \text{suma de cuadros verde oscuro en la fila } \theta = .7\\ &= 0.384\end{aligned}$$

Vemos que la potencia es mayor para $\theta = .7$ que para $\theta = .6$. Esto no es sorprendente ya que esperamos que sea más fácil reconocer que una moneda de .7 es sesgada que una moneda de .6. Normalmente, la potencia es mayor cuando la hipótesis alternativa está más alejada de la hipótesis nula. En este ejemplo, sería bastante difícil distinguir una moneda justa de una con $\theta = .51$.

¿Cómo se relacionan estos conceptos?

Para ilustrar mejor la relación entre los conceptos de nivel de significación, potencia, hipótesis nula, e hipótesis alternativa, veamos un ejemplo con alternativa simple.

Un emprendedor produce dos tipos distintos de alfajores, que se diferencian solamente en la cantidad de dulce de leche. Unos tienen μ_0 gramos y los otros tienen μ_1 gramos, con $\mu_0 < \mu_1$. Sin embargo, el proceso de producción presenta cierta variabilidad, de modo que la cantidad de dulce de leche vertida es una variable aleatoria de la forma $\mu_i + X$, en donde Z es normal de esperanza nula y varianza σ^2. La varianza es conocida por el emprendedor.

El emprendedor está muy entusiasmado con las nuevas tecnologías y desea automatizar el etiquetado de las cajas de alfajores. Cada caja tiene 16 alfajores, todos con el mismo μ_i, y se debe decidir qué etiqueta poner. Para esto se pesan los 16 alfajores, y luego se divide por 16. Es decir, el estadístico es el promedio de los pesos de los 16 alfajores

$$X = \frac{1}{16}\sum_{j=1}^{16} P_j = \frac{1}{10}\sum_{j=1}^{16}(m + \mu_i + X_j) = m + \mu_i + \overline{Z}_{16}$$

en donde m es el peso del alfajor que no es dulce de leche, y $\overline{Z}_{16}$ es el promedio de 16 variables $N(0,\sigma^2)$ independientes. Restando m (que suponemos constante e igual para todos los alfajores) podemos asumir que $m = 0$.

Llamemos μ a la cantidad de dulce de leche que tienen los alfajores de la caja. La hipótesis nula es que los alfajores tienen μ_0 gramos

$$H_0 : \mu = \mu_0$$

y la hipótesis alternativa es que tienen μ_1 gramos

$$H_A : \mu = \mu_1$$

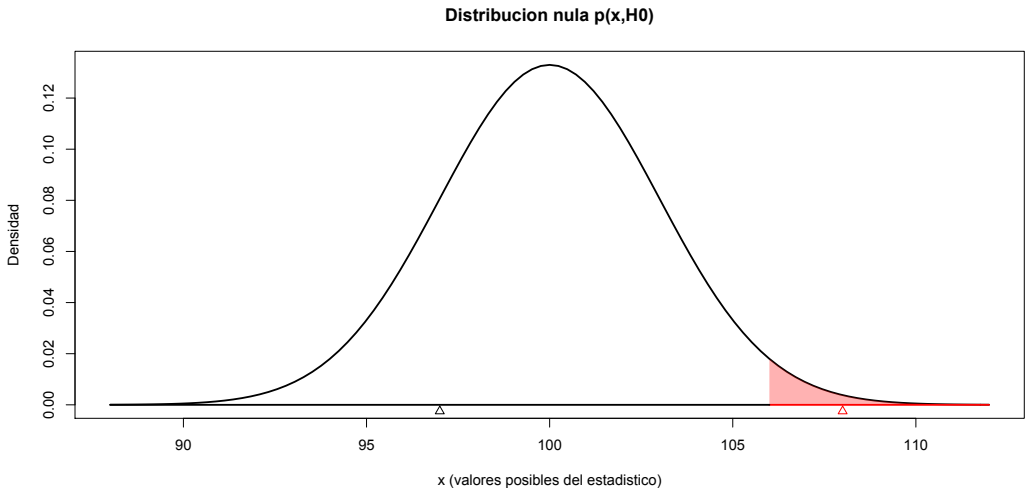
Como $\mu_0 < \mu_1$, es razonable rechazar H_0 cuando X es mayor que μ_0 por una cantidad importante. Así que una región de rechazo puede ser

$$I = [\mu_0 + c, +\infty),$$

en donde $c > 0$ es una constante. Es decir, nuestra *regla de decisión* es la siguiente:

1. Si X cae en I, rechazamos H_0 y ponemos la etiqueta que dice μ_1 en la caja.
2. Si X cae fuera de I, no rechazamos H_0 y ponemos la etiqueta que dice μ_0 en la caja.

Pongamos números concretos al ejemplo. Supongamos que $\mu_0 = 100$ gr y que $\sigma = 12$ gr. La primer figura a continuación ilustra la distribución nula con las regiones de rechazo I (en rojo) y de no rechazo I^c (en negro). También se muestran dos valores posibles del estadístico: x_1 (triángulo negro) y x_2 (triángulo rojo).



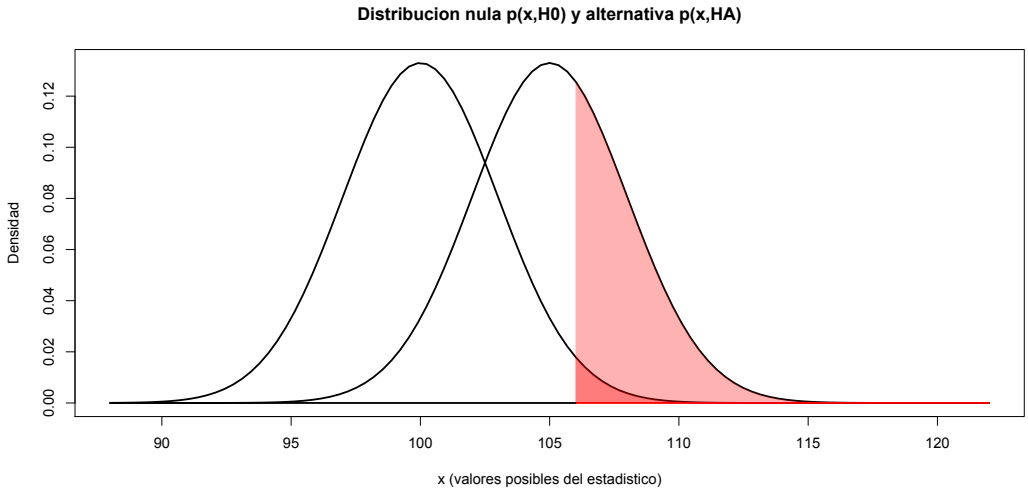
El valor x_1 está en la región de no rechazo. Entonces, si nuestros datos produjeran el valor de estadístico x_1, no rechazaríamos la hipótesis nula H_0. Por otro lado, el valor x_2 está en la región de rechazo, por lo que si nuestros datos produjeran x_2, rechazaríamos la hipótesis nula en favor de la hipótesis alternativa.

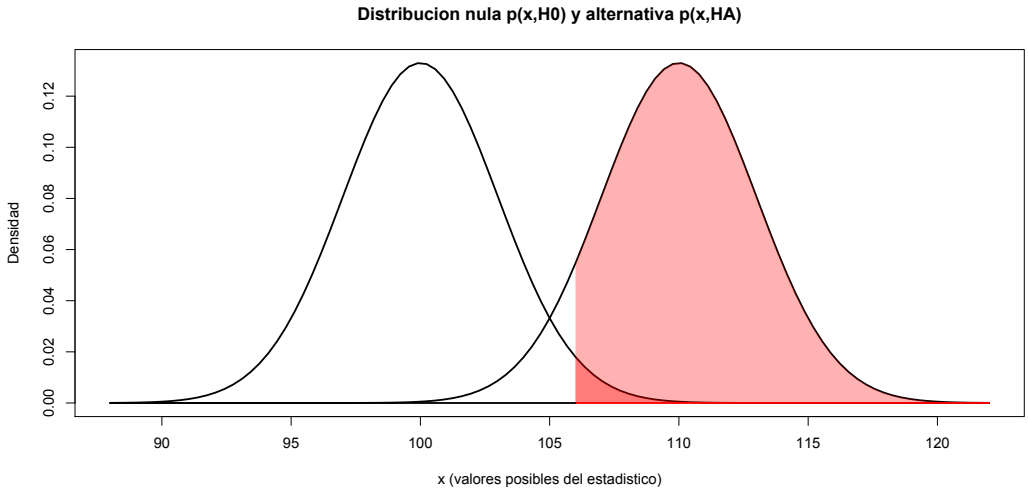
Hay varias cosas a tener en cuenta en esta figura:

1. La región de rechazo consiste en valores alejados hacia la derecha del centro de la distribución nula.
2. La región de rechazo es de una cola. También veremos ejemplos de regiones de rechazo de dos colas.
3. No se menciona la hipótesis alternativa. Rechazamos o no rechazamos H_0 basándonos solo en la región de rechazo. Como veremos, es mejor considerar la hipótesis alternativa H_A al elegir una región de rechazo, pero no desempeña un papel formal en rechazar o no rechazar H_0.
4. A veces, llamamos a la región de no rechazo región de aceptación. Esto es técnicamente incorrecto porque nunca se acepta realmente una hipótesis nula. Rechazamos o

decimos que los datos no son compatibles con rechazar H_0. Esto se resume a menudo en la frase: *nunca se puede probar la hipótesis nula*.

Las siguientes dos figuras muestran tests de alta y baja potencia. La primera corresponde al caso en que $\mu_1 = 105$ gr, y la segunda a $\mu_1 = 110$ gr. El área sombreada debajo de $p(x|H_0)$ representa el nivel de significación. Del mismo modo, el área sombreada debajo de $p(x|H_A)$ representa la potencia, es decir, la probabilidad de que el estadístico esté en la región de rechazo cuando H_A es verdadera. Ambos TdH tienen el mismo nivel de significación, pero si $p(x|H_A)$ tiene una superposición considerable con $p(x|H_0)$, la potencia es mucho menor. Merece la pena comprender a fondo estas representaciones gráficas.





En ambos casos, la distribución nula, la región de rechazo y el nivel de significación son los mismos. En la figura inferior vemos que las medias de las dos distribuciones están separadas por más de 3 desviaciones estándar. Dado que las áreas bajo las densidades tienen muy poca superposición, el test tiene potencia alta. Es decir, si los alfajores corresponden a una caja de H_A, es casi seguro que su peso promedio estará en la región de rechazo.

En la figura superior, las medias de las dos distribuciones están a menos de 2 desviaciones estándar, y por lo tanto el test tiene menor potencia. Esto es, si los alfajores corresponden a una caja de H_A, es menos probable que su peso promedio caiga en la región de no rechazo.

Por lo general, podemos aumentar la potencia de un test aumentando la cantidad de datos (la cantidad de alfajores en una caja) y, por lo tanto, disminuyendo la varianza de ambas distribuciones. Por eso, el diseño experimental previo es muy importante para determinar de antemano el número de mediciones necesarias para lograr la potencia deseada.

2. Diseñando un test de hipótesis

Formalmente, todo lo que requiere un test de hipótesis es H_0, H_A, un estadístico y una región de rechazo. En la práctica, el diseño se realiza a menudo siguiendo los siguientes pasos.

1. *Elegir la hipótesis nula* H_0. La elección de H_0 y H_A no es matemática. Es arte y costumbre. A menudo elegimos H_0 de modo que sea simple. En general H_0 representa la explicación más simple o cautelosa de los datos, por ejemplo, un fármaco no tiene efecto, la moneda no es sesgada, etc.

2. *Decidir si* H_A *es a una o a dos colas*. En el ejemplo de la moneda, queríamos saber si la moneda era sesgada. Una moneda sesgada podría estar sesgada a favor o en contra de las caras, por lo que $H_A : \theta \neq 0.5$ es una hipótesis a dos colas. Si solo nos importa si la moneda está sesgada a favor de caras, podríamos utilizar la hipótesis a una cola $H_A : \theta > 0.5$. En muchas situaciones se desea comparar con una H_A a una sola cola pues se sabe, por conocimientos previos, que la otra alternativa no es posible o relevante.

3. *Elegir un estadístico*. Por ejemplo, la media muestral, la mediana, o la varianza muestral. A menudo la elección es obvia. Algunos estadísticos habituales que encontraremos son z, t y χ^2. Aprenderemos a usar estos estadísticos a medida que trabajemos en los ejemplos de las próximas clases. Una aspecto importante que veremos repetidamente es que las distribuciones que acompañan a estos estadísticos son siempre condicionadas bajo la hipótesis nula.

4. *Elegir un nivel de significación y determinar la región de rechazo*. Usualmente usaremos α para denotar el nivel de significación. Es imprescindible elegir α por adelantado. Los valores típicos son 0.1, 0.05, 0.01. El valor que elijamos dependerá de las consecuencias de un error de tipo I.

   Si $\alpha = 0.1$ entonces esperamos una tasa de error tipo I del 10 %. Es decir, esperamos rechazar la hipótesis nula en el 10 % de los experimentos en los que la hipótesis nula es cierta. Si 0.1 es un nivel de significación razonable depende de las decisiones que se tomarán al usarlo.

   Por ejemplo, si el experimento es para determinar si una barra de chocolate tiene más de 72 % de cacao, entonces probablemente esté bien una tasa de error del 10 %. Creer falsamente que el cacao de la barra es mayor a 72 %, es probablemente aceptable. Por otro lado, si un laboratorio forense está identificando huellas dactilares para un juicio por asesinato, entonces una tasa de error de tipo I del 10 %, es decir, afirmar erróneamente que las huellas dactilares encontradas en la escena del crimen pertenecían a una persona verdaderamente inocente, definitivamente no es aceptable.

   Una vez elegido el nivel de significación, podemos determinar la región de rechazo, casi siempre en la(s) cola(s) de la distribución nula.

En el ejemplo de la moneda, H_A es a dos colas, por lo que la región de rechazo es la unión de dos colas de la distribución nula. Recordar que esta distribución es:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Si elegimos $\alpha = 0.05$, entonces la región de rechazo debe tener probabilidad como máximo .05. Para una región de rechazo a dos colas obtenemos

$$I = \{0, 1, 9, 10\}.$$

Si elegimos $\alpha = 0.01$ la región de rechazo es

$$I = \{0, 10\}.$$

Supongamos que cambiamos H_A por "la moneda está sesgada en favor de las caras". Ahora tenemos una hipótesis a una cola $\theta > 0.5$. Nuestra región de rechazo ahora estará en la cola de la derecha, ya que no queremos rechazar H_0 a favor de H_A si obtenemos una pequeña cantidad de caras. En este caso, si $\alpha = 0.05$, la región de rechazo es

$$I = \{9, 10\}.$$

Si elegimos $\alpha = 0.01$ entonces la región de rechazo es

$$I = \{10\}.$$

En general, los extremos de la región de rechazo se llaman *valores críticos* de la distribución nula, asociados al nivel de significación α.

Si H_0 es compuesta, entonces $\mathbf{P}(\text{error de tipo I})$ depende de qué miembro de H_0 es verdadero. En este caso, el nivel de significación se define como el máximo de estas probabilidades.

5. *Determinar la(s) potencia(s).* Como vimos en el ejemplo de la moneda, una vez que se establece la región de rechazo, podemos determinar la potencia del test en varios valores de la hipótesis alternativa.

El p-valor

En la práctica, al igual que en los test de permutaciones, a menudo se especifica el nivel de significación (la probabilidad umbral) y se rechaza o no H_0 utilizando el p-valor.

Recordar que el p-valor es la probabilidad, asumiendo la hipótesis nula, de observar datos tanto o más extremos que los datos observados. Lo que "tanto o más extremo" significa depende del contexto. Si el p-valor es menor que el nivel de significación α, entonces se rechaza H_0. De lo contrario, no rechazamos H_0.

Veamos por qué esto es coherente con la regla de decisión basada en la región de rechazo. Supongamos por simplicidad que la región de rechazo es $I = [c, +\infty)$. Denotemos por $p(x|H_0)$ la distribución nula y supongamos que es estrictamente creciente, también por simplicidad.

De la definición de nivel de significación tenemos que $\mathbf{P}(I|H_0) = \alpha$, o lo que es lo mismo, $\mathbf{P}(X \geq c|H_0) = \alpha$. Llamemos X_{obs} al valor observado de X. En este caso (a una cola) el

p-valor es $p = \mathbf{P}(X \geq X_{\text{obs}}|H_0)$, y usando la monotonía estricta de la probabilidad, vemos que

$$X_{\text{obs}} \geq c \text{ si, y solo si } p = \mathbf{P}(X \geq X_{\text{obs}}|H_0) \leq \mathbf{P}(X \geq c|H_0) = \alpha.$$

Es decir, rechazamos H_0 si, y solo si $p \leq \alpha$.

3. El test z

En general, un test z es cualquier test para el cual la distribución del estadístico bajo la hipótesis nula es normal. Este supuesto de normalidad es bastante usual, y se cumple con buena aproximación en caso de muestras grandes.

Nosotros nos limitaremos al caso en que las observaciones se pueden suponer normales de media desconocida, pero de varianza conocida. Este supuesto sí es un poco extraño ¿cómo puede uno conocer la varianza y no la media? Se da en situaciones particulares, como por ejemplo en control de calidad, en donde la media puede depender de la calibración de una máquina, pero la variabilidad intrínseca del procedimiento es fija. Esta variabilidad se puede conocer por mediciones previas independientes (o por argumentos teóricos), y solo se desea controlar la calibración de la máquina.

El test z se usa comúnmente cuando (i) se desea decidir sobre el valor de la media de una población normal, o (ii) cuando se desea comparar las medias de dos poblaciones normales; en ambos casos se asumen las varianzas conocidas.

Veamos un ejemplo concreto. Supongamos que fabricas máquinas de café. El cliente inserta \$50 pesos y la máquina de café entrega 150 ml de café premium. La máquina debe entregar "exactamente" 150 ml de café. Si entrega más de 150 ml (como cuando queda chorreando por unos segundos más de lo esperado), no le gustará a el propietario de la máquina ya que afectará sus márgenes de ganancia. Si entrega menos de 150 ml, los clientes que usan la máquina se sentirán estafados.

Supongamos que el contenido de líquido vertido por la máquina es $L = \mu + X$ en donde X es normal de esperanza nula. Asumimos que la variabilidad de líquido vertido por la máquina es intrínseca e igual a $\sigma = 5$ ml. Para controlar la máquina se toma una muestra de 9 vertidos $L_1, \ldots, L_9$. Los valores observados son

144 154 156 144 150 157 144 143 142

Vamos a diseñar un TdH para controlar si la máquina está funcionando correctamente. Para esto seguiremos los pasos sugeridos en la sección anterior.

1. *Elegir la hipótesis nula H_0*. Que la máquina funcione correctamente es que $\mu = 150$, por lo que $H_0 : \mu = 150$.
2. *Decidir si H_A es a una o a dos colas*. En cualquiera de las dos situaciones, tanto si la máquina vierte menos o más líquido del necesario, estaríamos en problemas. Además, no tenemos conocimientos previos que nos permitan descartar una alternativa. Así que nos interesa hacer un test a dos colas con $H_A : \mu \neq 150$.
3. *Elegir un estadístico*. El estadístico natural es elegir el promedio de las mediciones. Sin embargo, tomaremos el promedio estandarizado (pues con el tiempo uno se va

familiarizando con los valores estandarizados de un estadístico)

$$Z = \frac{\overline{L}_9 - 150}{5/\sqrt{9}}.$$

Notar que el valor observado de Z es

$$Z_{\text{obs}} = \frac{3(148.22 - 150)}{5} = -1.067.$$

4. *Elegir un nivel de significación y determinar la región de rechazo*. Tomemos $\alpha = 0.05$ que es un valor razonablemente seguro (es 1 en 20). Deseamos rechazar H_0 si el valor observado del estadístico Z está "lejos" de 0. Consideremos entonces una región de rechazo de la forma

$$I = (-\infty, -c] \cup [c, +\infty),$$

y calculemos c para que el nivel de significación sea α. La distribución nula de Z es la normal estándar, por lo que

$$\mathbf{P}(Z \in I | H_0) = 2(1 - \Phi(c)) = 0.05.$$

De aquí vemos que

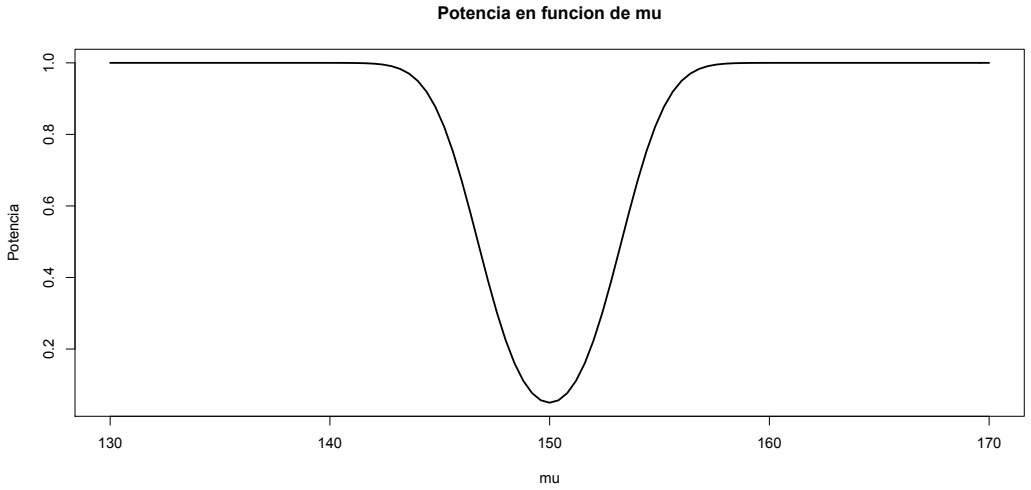
$$\Phi(c) = 1 - \frac{0.05}{2} = 0.975$$

De la tabla calculamos el valor crítico es $c = 1.96$.

5. *Determinar la(s) potencia(s)*. Para esto, tomemos $\mu \neq 150$. Si H_A es verdadera con el valor de μ, entonces la distribución de Z es normal de esperanza $3(\mu - 150)/5$. Por lo tanto

$$\begin{aligned}\mathbf{P}(Z \in I | \mu) &= \mathbf{P}(Z \leq -c | \mu) + \mathbf{P}(Z \geq c | \mu) \\ &= \Phi\left(-1.96 - \frac{3(\mu - 150)}{5}\right) + 1 - \Phi\left(1.96 - \frac{3(\mu - 150)}{5}\right).\end{aligned}$$

En la figura siguiente se muestra el gráfico de la potencia en función de μ.



En nuestro caso, como $Z = -1.067 \notin I$, concluimos que no hay evidencia para rechazar H_0, o lo que es lo mismo, no hay evidencia suficiente que justifique que la máquina está funcionando incorrectamente. Notar que el p-valor es en este caso $\mathbf{P}(|Z| > 1.067 | H_0) = 0.286 > \alpha$.

4. El test z para muestras grandes

Como vimos, un test z es cualquier test en el cual el estadístico elegido tiene distribución normal estándar bajo la hipótesis nula. Este es claramente el caso del promedio estandarizado cuando los datos son normales y la varianza es conocida.

Cuando el estadístico es asintóticamente normal (bajo la hipótesis nula), se puede definir la región de rechazo usando los valores críticos de la distribución normal estándar, al menos cuando el tamaño muestral n es razonablemente grande. En la práctica $n \geq 30$ suele ser suficiente en la mayor parte de las aplicaciones en las que se usa el promedio como estadístico. Esta es una de las consecuencias importantes del TCL.

Ejemplo 1

Un político afirma que más del 90 % de la población uruguaya está a favor de la legalización de las drogas. Una encuesta de 100 personas muestra que solo 79 de ellas están a favor de la legalización.

¿Qué podemos decir sobre la afirmación del político?

Antes de diseñar un TdH para analizar los datos de la encuesta, es conveniente ponernos de acuerdo en ciertos supuestos que harán el análisis posible.

Llamemos p a la proporción de uruguayos que están a favor de la legalización. Podemos suponer que la muestra de encuestados puede modelarse por variables

$$X_i = \begin{cases} 1 & \text{si la } i\text{-ésima persona encuestada está a favor;} \\ 0 & \text{si la } i\text{-ésima persona encuestada está en contra.} \end{cases}$$

De este modo, si las personas son elegidas al azar, cada X_i tiene distribución Bernoulli de parámetro p. Además, vamos a suponer que las $X_i's$ son independientes. Esto puede no ser muy creíble, pero para una primera aproximación es suficiente.

De este modo, la proporción de personas en la muestra que están a favor de la legalización es

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n}$$

cuyo valor observado es 0.79. El político afirma que $p > 0.9$, y vamos a diseñar un TdH para comprobar si los datos respaldan esta afirmación. Como siempre, seguiremos los cinco pasos sugeridos en la clase anterior.

1. *Elegir la hipótesis nula H_0*. Estamos interesados en comprobar si $p > 0.9$ o no. Luego, la hipótesis nula es $H_0 : p > 0.9$, que es la posición defendida por el político.
2. *Decidir si H_A es a una o a dos colas*. Valores grandes de la proporción observada estarían a favor del político (a favor de H_0), y valores pequeños en contra. Luego, la hipótesis alternativa es a una cola $H_A : p < 0.9$.
3. *Elegir un estadístico*. La distribución de S_n es binomial de parámetro p, y es tentador usar la suma estandarizada
$$Z = \frac{S_n - pn}{\sqrt{np(1-p)}}$$

como estadístico. Sin embargo, la hipótesis nula que hemos elegido no permite calcular el valor de Z_{obs}; es decir, el valor observado de S_n es 79, pero ¿cómo usamos $p > 0.9$ en la fórmula?

Usaremos entonces

$$Z = \frac{S_n - 0.9n}{\sqrt{n0.9 \times 0.1}} = \frac{S_{100} - 90}{3}$$

como estadístico. Por lo tanto $Z_{\text{obs}} = -11/3 = -3.67$.

4. *Elegir un nivel de significación y determinar la región de rechazo.* Usaremos $\alpha = 0.05$ como nivel de significación. Como H_0 es compuesta, el nivel de significación corresponde al máximo del nivel de significación variando en $0.9 < p \leq 1$.

   Rechazaremos H_0 si Z_{obs} es menor que c para un cierto valor crítico c. En palabras, esto quiere decir que la proporción observada es demasiado pequeña para ser compatible con H_0.

   Para cualquier p, como $n = 100$ es razonablemente grande, podemos suponer que S_{100} tiene distribución normal $N(100p, 100p(1-p))$.

   Supongamos que $p > 0.9$. Entonces

$$\begin{aligned}
\mathbf{P}(Z \leq c|p) &= \mathbf{P}\left(\frac{S_{100} - 90}{3} \leq c \middle| p\right) = \mathbf{P}(S_{100} \leq 3c + 90|p) \\
&= \mathbf{P}\left(\frac{S_{100} - 100p}{\sqrt{100p(1-p)}} \leq \frac{3c + 90 - 100p}{\sqrt{100p(1-p)}} \middle| p\right) \\
&= \Phi\left(\frac{3c + 90 - 100p}{\sqrt{100p(1-p)}}\right)
\end{aligned}$$

   El nivel de significación es el máximo de estas probabilidades con $0.9 < p \leq 1$. En ese rango de valores de p, la función

$$\frac{3c + 90 - 100p}{\sqrt{100p(1-p)}}$$

   es decreciente, y Φ es creciente por ser una f.d.a., por lo que su máximo se da en $p = 0.9$. Luego, el nivel de significación es

$$\sup_{0.9 < p \leq 1} \mathbf{P}(Z \leq c|p) = \Phi(c) = \alpha.$$

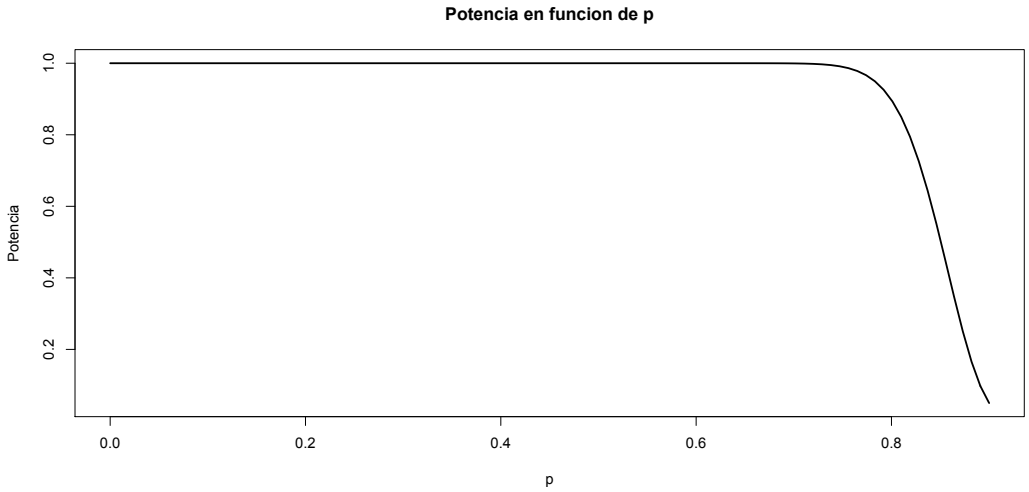
   Luego, c es el valor crítico $z_{1-\alpha} = z_{0.95}$ de la distribución normal estándar. De la tabla vemos que $c = -1.645$. Dicho de otro modo, la región de rechazo es

$$I = (-\infty, -1.645].$$

5. *Determinar la(s) potencia(s).* Para cada $0 \leq p < 0.9$ el mismo cálculo que hicimos arriba muestra que

$$\mathbf{P}(Z \leq c|p) = \Phi\left(\frac{3(-1.645) + 90 - 100p}{\sqrt{100p(1-p)}}\right).$$

Esta es la potencia si la verdadera proporción fuera p.



La conclusión entonces es que rechazamos H_0 ya que $Z_{\text{obs}} = -3.67 \leq -1.645$. También podemos razonar usando p-valores. En este caso de H_0 compuesta, el p-valor es el máximo p-valor para cada hipótesis simple en H_0. Del mismo modo que hicimos antes, se ve que

$$\sup_{0.9<p\leq 1} \mathbf{P}(Z \leq Z_{\text{obs}}|p) = \mathbf{P}(Z \leq Z_{\text{obs}}|p = 0.9) = \Phi(-3.67) = 0.0001.$$

Como el p-valor es menor que α rechazamos H_0. ■