Taller Iniciación Estadística Aplicada a la Investigación

Test de Hipótesis: Comparación de dos muestras

Hugo J. Bello

2024/04

Test de Hipótesis: Definición

Dada una muestra $x_1, \ldots, x_n$, Los tests (o contrastes) de hipótesis tratan usar la información disponible en los datos para decidir **entre dos posibilidades referentes a la distribución que genera los datos**:

- La **hipótesis nula (H_0)**. Es la hipótesis que queremos comprobar o refutar.
- La **hipótesis alternativa (H_1)**. La opuesta a la hipótesis anterior.

Ejemplos

Estas dos opciones pueden venir dadas de muchas maneras, por ejemplo

- Podemos ante un ensayo clínico de un medicamento preguntarnos si el tratamiento no ha funcionado (hipótesis nula) o si lo ha hecho (hipótesis alternativa).
- Podemos preguntarnos si la media de infartos es igual en varones que en mujeres (hipótesis nula) o si es mayor en varones (hipótesis alternativa)

Ejemplos

Veamos un ejemplo con mayor detalle:

Si conocemos el consumo de azucar en 100 hombres y 100 mujeres y queremos averiguar si hay o no un mayor consumo de azucar en hombres y mujeres. Imaginemos que podemos suponer que el consumo de azucar en hombres sigue una distribución normal $N(\mu_1, \sigma_1)$ y en mujeres una distribución $N(\mu_2, \sigma_2)$. Plantearíamos:

$$
\begin{aligned}
H_0 :& \mu_1 = \mu_2 \\
H_1 :& \mu_1 \neq \mu_2
\end{aligned}
$$

Veremos cómo atacar este problema.

Estructura

Los tests de hipótesis

- Usan un **estadístico de contraste** E, calculado a partir de los datos $(x_1, \ldots, x_N)$ y distinto en cada caso. La idea que bajo la hipótesis nula la distribución de E es una conocida (Normal, E de Snedecor, t de Student, χ^2...)
- Si el valor de E es *muy extremo o improbable* bajo la Hipótesis nula, concluiremos tenemos evidencias **en contra de** H_0, de lo contrario concluiremos que tenemos evidencias **A favor de** H_0.
- Al igual que los intervalos de confianza utilizan un **nivel de significación** α fijado de partida (probabilidad de rechazar H_0 siendo verdadera).

Podemos equivocarnos al decidirnos por aceptar o rechazar H_0, este error puede presentarse de dos formas

- H_0 puede rechazarse siendo verdadera. En este caso estamos ante un **error de tipo I**, y la probabilidad de que esto ocurra se denota por α. A este valor α se denomina el **nivel de significación** del test.
- H_0 puede ser aceptada siendo falsa. En este caso estamos ante un **error de tipo II**. La probabilidad de este hecho se denota por β

	H_0 es verdadera	H_0 es falsa
Aceptamos H_0	correcto	**Error tipo II**
Rechazamos H_0	**Error tipo I**	correcto

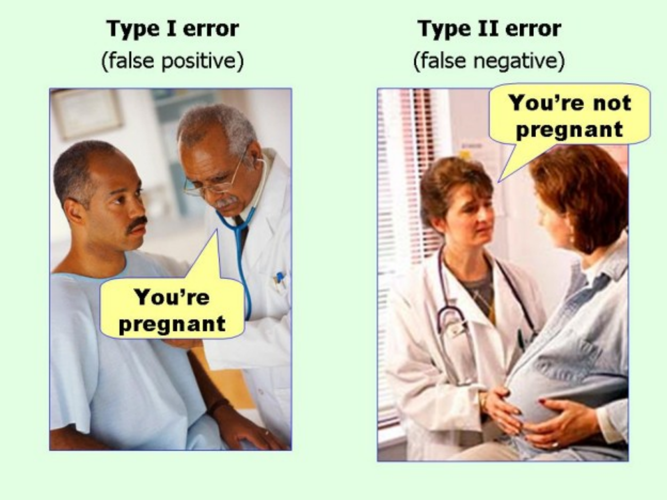


Figure: Errores

Elección de H_0

A menudo la hipótesis nula se toma como *la más conservadora* o la hipótesis que ya asumíamos correcta. En otras palabras **aquella que aceptar erróneamente conlleva un menor riesgo.**

La analogía común es imaginarlo en el contexto de un juicio.

- H_0 sería la hipótesis de que el acusado es inocente
- H_1 de que es culpable

El **error de tipo I** sería decir que hemos condenado al acusado siendo inocente, y el **de tipo II** que hemos dejado libre al culpable. En la realidad se prefiere cometer el primer error que el segundo.

Elección de H_0

Por ejemplo al realizar un ensayo clínico se suele tomar como la hipótesis nula *que el tratamiento no funciona*.
Este tipo de elecciones se justifican porque lo que se suele intentar minimizar en los tests de hipótesis es el error de tipo I, ya que **minimizar a la vez el error de tipo I y II es imposible**.

El p-valor

El p-valor se define como la probabilidad de encontrar de, una vez realizado un experimento, encontrar resultados tan extremos o más que los observados, suponiendo que la hipótesis nula es cierta.

Desentrañemos esto:

- Es una **probabilidad**, por lo tanto un número entre 0 y 1.
- Se usa como manera de medir la **evidencia a favor de la hipótesis nula**: Cercano a 0 poca evidencia, cercano a 1 mucha.
- Si el p-valor es bajo, la hipótesis nula es menos probable y la **significación estadística es baja**. Debemos rechazar la teoría planteada en favor de la alternativa. El valor a partir del cual se decide esto es el nivel de significación α.

Procedimiento general

El **procedimiento de un test de hipótesis** es el siguiente:

1. Se elige el nivel de significación α, habitualmente 0.05.
2. Se eligen las hipótesis H_0 y H_1.
3. Se elige el tipo de test (t-test, chi-cuadrado, Wilcoxon..., veremos cómo) y se lleva a cabo.
4. Se calcula el p-valor.
    - Si el p-valor es menor que el nivel de significación α se **rechaza la hipótesis nula** H_0. Consecuentemente se acepta la alternativa H_1.
    - Si es mayor se **acepta la hipótesis nula** H_0.

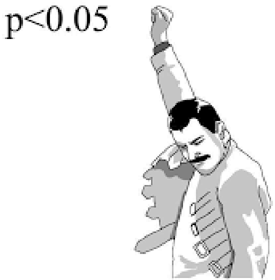


Figure: a por ese p-valor bueno

Problemas con el p-valor

- El p-valor mide en parte la probabilidad de que los resultados que hemos obtenido *no respalden la hipótesis* por culpa de una *mala suerte* a la hora de recoger los datos. No indican realmente la probabilidad de la hipótesis nula.
- Si observamos 100 variables es probable que encontremos un p-valor bajo para alguna de ellas **de forma casual**. Si solo incluimos esa variable en el estudio parecerá que hemos conseguido demostrar algo que no es real.
- El p-valor **no es un indicativo de la importancia científica del efecto observado en la variable**. Por ejemplo un test puede demostrar que el grupo placebo tiene medio grado menos de temperatura corporal, pero esto no ser un efecto médicamente realista referente a la enfermedad estudiada.

Problemas con el p-valor

Figure: Boromir lo sabe

Tests para comparar dos muestras

La situación más habitual que trataremos es aquella en que tenemos dos muestras:

$$x_1, \ldots, x_n$$

$$y_1, \ldots, y_n$$

Que queremos comparar. Habitualmente querremos saber si estas dos muestras son **lo suficientemente distintas** en algún sentido. Por ejemplo si sus medias son significativamente distintas.

Como veremos esta es la clave para averiguar, mediante tests de hipótesis, cosas como: si un tratamiento ha funcionado, si una enfermedad afecta más o más duramente a un grupo de pacientes que a otros..

Ejemplo: Datos

Regresemos al ejemplo:

paciente	sexo	edad	grupo	días ingreso	marcador en sangre
1	H	36.0	Placebo	10.0	0.44
2	H	43.0	Placebo	9.0	2.27
3	M	48.0	Placebo	10.0	2.1
4	M	57.0	Placebo	10.0	3.9
⋮	⋮	⋮	⋮	⋮	⋮
77	H	19.0	Medicamento	3.0	9.58
78	M	47.0	Medicamento	1.0	2.54
79	H	38.0	Medicamento	5.0	3.12
80	H	23.0	Medicamento	6.0	5.48

Ejemplo: Planteamiento

Queremos ver si el medicamento reduce los días de ingreso. Para ello **nos planteamos si la media de los pacientes que han tomado el medicamento es menor**.
Si llamamos

$$\mu_1 = \text{Média de los días de ingreso del grupo placebo}$$
$$\mu_2 = \text{Média de los días de ingreso del grupo experimental}$$

Planteamos el test de hipótesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Observa cómo hemos elegido como hipótesis nula la hipótesis equivalente a *el tratamiento no ha funcionado*. Veamos distintos tests de hipótesis con

Test de la t de Student

Desarrollado por William Sealy Gosset (bajo el pseudónimo de Student).

El **test de la t de Student** sobre dos muestras $x_1, \ldots, x_n$ y $y_1, \ldots y_n$ puede utilizarse para testar las hipótesis

$$
\begin{aligned}
H_0 :& \mu_1 = \mu_2 \\
H_1 :& \mu_1 \neq \mu_2
\end{aligned}
$$

Conocido como el test de **dos colas** o

$$
\begin{aligned}
H_0 :& \mu_1 \leq \mu_2 \\
H_1 :& \mu_1 > \mu_2
\end{aligned}
$$

asume que las siguientes condiciones se cumplen:

- los datos son de una variable continua, tomados aleatoriamente de una población.
- La varianza es la misma para las dos muestras (aunque esto puede arreglarse usando una variante llamad el test de Welch)
- Los datos **siguen una distribución normal**

Utiliza el estadístico de contraste

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}},$$

que sigue una distribución T de student con $n - 1$ grados de libertad.

Para el ejemplo

Veamos **si se cumplen las condiciones** vara la variable dias de ingreso y los dos grupos estudiados:

- es evidente que son datos continuos tomados aleatoriamente ✓
- la varianza del grupo placebo es 4.14 y la del medicamento es 4.13, así que son prácticamente iguales. ✓
- Estudiamos la gráfica qqplot para verificar la normalidad ✓

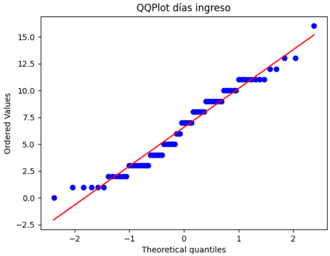


Figure: qqplot

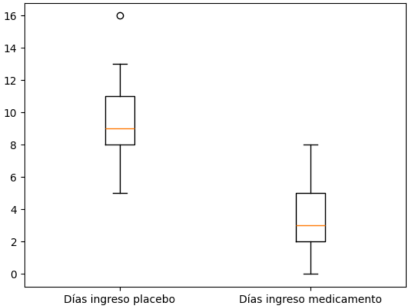


Figure: boxplot

Procedemos con el **test de la t de Student**:

- Plateamos el test de dos colas en el que la hipótesis nula es que las medias coinciden (medicamento no ha funcionado)

$$
\begin{aligned}
H_0 : & \mu_1 = \mu_2 \\
H_1 : & \mu_1 \neq \mu_2
\end{aligned}
$$

- **Fijamos el nivel de significación** habitual $\alpha = 0.05$ y llevamos a cabo el test (habitualmente con un software estadístico como R o SPSS)
- **el resultado es un p-valor** de 0.00001 ($0.00001 < 0.05$). Por lo tanto **rechazamos la hipótesis nula** y concluimos que hay evidencia estadística para afirmar que el tratamiento ha funcionado.

¿Qué pasa si los datos no siguen la distribución normal?

En estos casos debemos usar métodos **no paraméticos**. Se trata de métodos que no asumen que los datos siguen una particular distribución

Test de Wilcoxon / Mann-Whitney

Se puede utilizar como **alternativa no-paramétrica al test de la t de student** ya que no asume normalidad, en su lugar exige:

- Las observaciones de ambos grupos son independientes, además de ser variables ordinales o continuas.
- Bajo la hipótesis nula, la distribución de partida de ambos grupos es la misma: $P(X > Y) = P(Y > X)$. Es decir testa algo más exigente que el test de la t de student.

Test de Wilcoxon / Mann-Whitney, funcionamiento

Algunas ideas del funcionamiento:

- dadas dos muestra $x_1, \ldots, x_{n_1}$ $y_1, \ldots, y_{n_2}$ las juntamos y ordenamos en una única $z_1, \ldots, z_{n_1+n_2}$.
- A los valores de z_i ordenados se les asigna su rango r_i (al primero rango 1, al segundo rango 2...)
- El estadístico de contraste es

$$U = \min(U_1, U_2)$$

  Donde $U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$, $U_2 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$. Siendo R_1, R_2 son la suma de los rangos correspondientes a $x_1, \ldots, x_{n_1}$ y $y_1, \ldots, y_{n_2}$ respectivamente.
- El estadístico si el estadístico U tiene una distribución conocida que puede consultarse en tablas o paquetes estadísticos, un valor extremo de U nos permitirá rechazar la hipótesis nula.

Ejemplo

Regresemos al ejemplo, fijémonos ahora en la variable marcador en sangre

paciente	sexo	edad	grupo	días ingreso	marcador en sangre
1	H	36.0	Placebo	10.0	0.44
2	H	43.0	Placebo	9.0	2.27
3	M	48.0	Placebo	10.0	2.1
4	M	57.0	Placebo	10.0	3.9
⋮	⋮	⋮	⋮	⋮	⋮
77	H	19.0	Medicamento	3.0	9.58
78	M	47.0	Medicamento	1.0	2.54
79	H	38.0	Medicamento	5.0	3.12
80	H	23.0	Medicamento	6.0	5.48

Ejemplo

Queremos repetir el análisis que hemos realizado antes y ver **si el tratamiento ha impactado en los niveles de marcador en sangre**
El boxplot parece suregirlo:

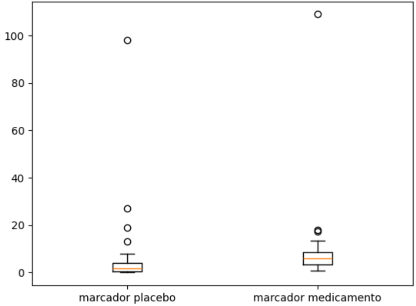


Figure: boxplot

Ejemplo

Esta variable es que no sigue la distribución normal:

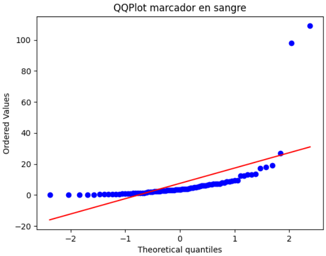


Figure: qqplot

No podemos realizar un test de la t de Student pero **si podemos realizar un test de Wilcoxon**

Procedemos con el test de Wilcoxon/Mann Whitney:

- Plateamos el test

$$H_0 : \text{Placebo y medicamento tienen la misma distribución}$$
$$H_1 : \text{Tienen distinta distribución}$$

- **Fijamos el nivel de significación** habitual $\alpha = 0.05$ y llevamos a cabo el test de Wilcoxon (habitualmente con un software estadístico como R o SPSS)
- **el resultado es un p-valor** de 0.0001. Por lo tanto **rechazamos la hipótesis nula** ($0.0001 < 0.05$) y concluimos que hay evidencia estadística para afirmar hay diferencias significativas entre las distribuciones. En vista de esto y del gráfico boxplot, si podemos concluir que el tratamiento puede haber funcionado.

Tests para datos cualitativos: χ^2 test

Si tenemos datos *cualitativos* no podemos relizar tests como los anteriores.

Se trata de situaciones como las de este estudio:

	enf. cardiovascular	enf. mental	sanos	total
deportistas	11	9	30	50
no deportistas	41	47	33	121
total	52	56	63	171

¿Afecta el deporte a la salud mental y cardiovascular de acuerdo a los datos? Este tipo de tablas se llaman **tablas de contingencia**

Podemos realizar un **test de la χ^2 de independencia para testar esto**.

χ^2 test

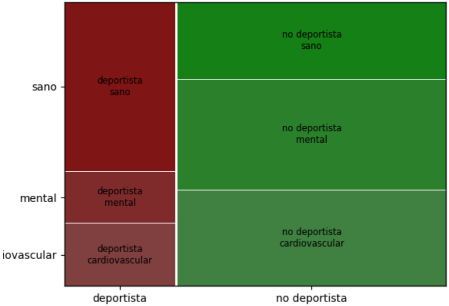
- Se utiliza para testar si variables reflejadas en una tabla de contingencia son estadísticamente independientes. Esto pueden entenderse en el ejemplo anterior como el hecho de que el ser deportista o no no afecta la proporción de pacientes que tienen o no una u otra enfermedad.

$$H_0: \text{la variable } \textit{deporte} \text{ y la variable } \textit{enfermedad} \text{ son independientes}$$
$$H_1: \text{no los on}$$

- si el p-valor es menor que el nivel de significación **rechazamos H_0 y nos inclinamos a decir que las variables tienen dist. dependientes y por lo tanto ser deportista si afecta a las proporiciones de cada enfermedad**.

Un ejemplo

Para la tabla de contingencia anterior obtenemos un **p-valor de 0.0002**, con lo cual rechazamos y deducimos que hay indicios para afirmar que ser deportista afecta. Las tablas de contingencia pueden representarse mediante gráficas mosaico:



Otro ejemplo

Ejemplo real: *Aspirin for the Primary Prevention of Cardiovascular Events in Women and Men: A Sex-Specific Meta-analysis of Randomized Controlled Trials. JAMA, 295(3):306-313*

	Aspirin	Control/Placebo
Ischemic stroke	176	230
No stroke	21035	21018

En este caso

$$H_0 : \text{El efecto de tomar o no tratamiento es el mismo}$$
$$H_1 : \text{El efecto es distinto}$$

El p-valor resulta 0.008, luego rechazamos H_0 y concluimos que el efecto de tomar aspirina es significativamente distinto.

Alternativa a los test de hipótesis: Tamaño del efecto

El **tamaño del efecto** es una manera de medir la fuerza de la relación entre las variables estudiadas. En el contexto de por ejemplo un ensayo clínico sería medir **cuánto o con qué fuerza ha funcionado el medicamento** al comparar el grupo experimental y placebo.

Una medida estandarizada del tamaño de efecto es la **d de Cohen**. Si tenemos dos muestras $x_1, \ldots, x_{n_1}$, $y_1, \ldots y_{n_1}$ (que podrían corresponder las variables del grupo experimental y placebo por ejemplo), se define como:

$$d = \frac{\bar{x} - \bar{y}}{s}.$$

donde

$$s = \sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}}$$

Interpretación

La fuerza del efecto se medir mediante la tabla

Tamaño del efecto	d de Cohen	referencia/fuente
Very small	$\pm$ 0.01	[1]
Small	$\pm$ 0.20	[2]
Medium	$\pm$ 0.50	[2]
Large	$\pm$ 0.80	[2]
Very large	$\pm$ 1.20	[1]
Huge	$\pm$ 2.0	[1]

El signo indica la dirección del efecto.

[1] Sawilowsky, S (2009). New effect size rules of thumb
[2] Cohen, Jacob (1988). Statistical Power Analysis for the Behavioral Sciences. Routledge

Ventajas de su uso

- No se fundamenta en niveles de significación, p-valores o hipótesis nula/alternativa. Por lo tanto **vale para cualquier distribución**.
- **Complementa** a los test de hipótesis, es considerado buena práctica su uso como acompañamiento de estos.

Figure: Boromir usa la d de Cohen

En los ejemplos anteriores

Para la variable **días de ingreso**, tomando

$$X = \text{días de ingreso en grupo placebo}$$
$$Y = \text{días de ingreso en grupo experimental}$$

Obtenemos una d de Cohen de 3.05. Lo cual se puede interpretar como que el medicamento tiene un efecto pequeño en los días de ingreso y el valor de la media días de ingreso en placebo es mayor (esto nos lo da el signo $+$).

En los ejemplos anteriores

Para la variable **marcador en sangre**, tomando

$$X = \text{niveles de marcador en sangre en el grupo placebo}$$
$$Y = \text{niveles de marcador en sangre en el grupo experimental}$$

Obtenemos una d de Cohen de -0.2. Lo cual se puede interpretar como que el medicamento tiene un efecto pequeño el nivel del marcador en sangre y el valor de la media este marcador en experimental es mayor (esto nos lo da el signo $-$).