Tests de hipótesis estadísticas

Test de hipótesis sobre la media de una población .

Introducción con un ejemplo.

Los tests de hipótesis estadísticas se emplean para muchos problemas, en particular para comparar las medias de dos o más poblaciones. Por ejemplo cuando se desea comparar los resultados de dos (o más) métodos de medición. En estos ejemplos hay que considerar dos o más muestras. Los problemas de comparar dos o más muestras los veremos en la clase siguiente. Por ahora consideremos un problema más simple, que es el de considerar una sola muestra y querer estudiar si es cierta o no una hipótesis sobre la media de la población.

Ejemplo 1: Consideremos el siguiente ejemplo. Para conocer la exactitud de un método de medición del contenido de níquel en un mineral, se hacen 10 determinaciones para una aleación "standard" preparada de modo que se conoce el verdadero valor del contenido de níquel (con una muy buena aproximación) que es de 4.44%. Se obtienen los siguientes valores:

4.32 4.31 4.50 4.12 4.43 4.36 4.48 4.28 4.18 4.42

Calculemos la media y la DS de estas 10 determinaciones:

DESCRIPTIVE STATISTICS

VARIABLE	N	MEAN	SD	MINIMUM	MAXIMUM
NIQUEL	10	4.3400	0.1243	4.1200	4.5000

La media de las 10 determinaciones es menor que el valor verdadero, pero esto puede deberse al azar. Aunque el método de medición no tuviese error sistemático (μ fuese igual al verdadero valor), sabemos que la media muestral ($\overline{X}$) no va a coincidir con el verdadero valor, simplemente porque $\overline{X}$ es una variable aleatoria.

La pregunta que nos formulamos es: ¿con estos 10 datos podemos afirmar que el método de medición tiene error sistemático?

Para poder hacer afirmaciones de este tipo, vamos a tener que aceptar una probabilidad de error. La teoría de tests de hipótesis forma parte de la teoría de INFERENCIA ESTADÍSTICA. Al querer extrapolar de una muestra a una población siempre hay una probabilidad de cometer error.

Una forma intuitiva de responder a la pregunta formulada, sería calcular un intervalo de confianza para la media de la población. Calculemos un IC al 95%, usando la expresión:

$$[\overline{X} - t_{n-1;\,\alpha/2} * s / \sqrt{n} \;\leq\; \mu \;\leq\; \overline{X} + t_{n-1;\,\alpha/2} * s / \sqrt{n}]$$

o simplemente con el Statistix:

DESCRIPTIVE STATISTICS

VARIABLE	N	LO 95% CI	MEAN	UP 95% CI	SD
NIQUEL	10	4.2511	4.3400	4.4289	0.1243

Vemos que el IC al 95% para μ es [4.25, 4.43] que no incluye al valor verdadero (4.44%). Basándonos en este IC podríamos decir que μ es menor que el verdadero valor y que el método de medición tiene un error sistemático negativo. ¿Existe la posibilidad de que nos equivoquemos con este procedimiento? Sí, porque el IC no es "seguro" pero tiene una confianza del 95%, o sea una probabilidad de error del 5%.

El procedimiento que hemos usado recién es calcular un IC para μ y observar si el valor propuesto en la hipótesis está o no incluído en el IC. Generalmente se usa otro procedimiento de cálculo, pero la conclusión a la que se llega es la misma.

Problemas que trata la teoría de tests de hipótesis.

El problema que hemos planteado es un ejemplo de un tipo de problemas que se trata en la teoría de **tests de hipótesis estadísticas.**

Los problemas que trata esta teoría son los que pueden plantearse del siguiente modo: observo una muestra y tengo una hipótesis (por ejemplo acerca de la media o de la diferencia de dos medias poblacionales) y quiero saber si esa hipótesis es cierta o no. Para ello, como en cualquier problema de inferencia estadística, vamos a plantear primero un modelo probabilístico:

$X_1, X_2, ..., X_n$ vs. as. con una distribución de la forma $F(x,\theta,\lambda,...)$ donde los parámetros θ,λ, etc. son desconocidos,

e interpretar la hipótesis como una hipótesis sobre uno de los parámetros (o sobre una función de los parámetros). Luego, en función de la muestra observada, se decide si aceptamos o no la hipótesis..

En la teoría de tests de hipótesis estadísticas no se plantea una sóla hipótesis sino dos hipótesis: una se llama hipótesis nula y la otra alternativa.

En el ejemplo 1 el químico quiere decidise entre estas dos hipótesis:

1) el método de medición no tiene error sistemático
2) el método de medición tiene error sistemático

Para este ejemplo, podemos plantear el siguiente modelo probabilístico:

$$X_1, X_2, ..., X_n \text{ vs. as. i.i.d } N(\mu,\sigma^2) \qquad (14)$$

donde n=10 y Xi es la i-ésima determinación de níquel . Con este modelo las dos hipótesis se escriben

$$\mu = 4.44$$

$$\mu \neq 4.44$$

Es costumbre denotar H_0 a la hipótesis **hipótesis nula** y H_1 a la **hipótesis alternativa**. En el ejemplo conviene elegir

$$H_0 : \mu = 4.44$$
$$H_1 : \mu \neq 4.44$$

Un test de hipótesis es una regla de decisión que en función de los datos de una muestra $X_1, X_2, ..., X_n$ nos permite decidirnos por H_0 o por H_1 (mejor diremos "aceptamos H_0" o "rechazamos H_0"). Esta es la definición de test.

DEFINICIÓN: Un test es una regla de decisión que, en función de los datos de una muestra $X_1, X_2, ..., X_n$, permite rechazar o aceptar la hipótesis nula.

Derivemos un test para el ejemplo que estamos considerando. Es intuitivamente razonable que si $\overline{X}$ "se parece" a 4.44 vamos a aceptar H_0 y que si $\overline{X}$ "está lejos" del valor 4.44 vamos a rechazarla.
Sabemos que , bajo el modelo (14)

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2 / n}} \sim t_{n-1}$$

Esta variable T no se puede calcular, porque no conozco μ . Si H_0 fuese cierta $\mu = 4.44$, por lo tanto:

$$\text{si } H_0 \text{ es cierta} \qquad T = \frac{\overline{X} - 4.44}{\sqrt{s^2 / n}} \sim t_{n-1}$$

Este valor de T se puede calcular, en base a los datos de la muestra y resulta

$$T = \frac{4.34 - 4.44}{0.1243 / \sqrt{10}} = -2.54$$

La idea ahora es la siguiente: si H_0 fuese cierta se espera que $\overline{X}$ se parezca a 4.44 y que por lo tanto el valor de T recién calculado "esté cerca" del valor cero. Por lo tanto, si $\overline{X}$ "está lejos" de 4.44 o, lo que es lo mismo, si el valor calculado de T "está lejos" de cero, pensaríamos que es difícil que H_0 sea cierta, y estaríamos dispuestos a RECHAZAR H_0.

Tenemos que definir que queremos decir con "está cerca" o "está lejos" de cero.

Cuando H_0 es cierta, el cociente T tiene distribución aproximadamente t_{n-1}, lo que equivale a decir que si sacásemos muchas muestras (en la práctica sólo se saca una!) y graficásemos el histograma de estos cocientes, el histograma sería parecido a la curva de densidad t_{n-1}.

Se procede así: suponiendo H_0 cierta, se calcula la probabilidad de que ocurra un valor de T como el observado o aún más "lejos" de cero, o sea la probabilidad de que $|T| \geq -2.54$ (que es el área bajo las dos colas de la curva de la cola de la curva t_{n-1} a partir del valor –2.54).

Esta área puede calcularse usando el StatistiX (Statistics, Probability Functions, T2-tail, x=-2.54, DF=9). Resulta ser 0.03171. Esta probabilidad se llama "valor P" del test.

Entonces, si H_0 fuese cierta, la probabilidad de que ocurra una media muestral $\overline{X}$ como la observada o más alejada del valor propuesto en H0 es BAJA (p=0.032), ¿cual seria la conclusión entonces? : **SE RECHAZA H_0**

Se rechaza H_0 cuando el valor de P es pequeño. El valor de corte es arbitrario, pero casi siempre se usa 0.05, o sea **se rechaza H_0 cuando P<0.05.**

¿Puedo equivocarme? Sí, si H_0 fuese cierta podría darme un valor de T en las colas y rechazar, pero ¿cual es esta probabilidad? Es precisamente 0.05.

Es práctica común decir "la diferencia es estadísticamente significativa" como sinónimo de "se rechaza H_0". En el ejemplo la conclusión podría redactarse así:

La media de las 10 determinaciones es $\overline{X}$=4.34. La diferencia entre esta media y el valor verdadero 4.44 es estadísticamente significativa (P=0.031).

Comentario:
Todas las cuentas que hicimos pueden hacerse automáticamente con el StatistiX. Para ello vamos a "Statistics", "One, Two and Multiple Sample Tests", "One sample T test" , ponemos en el casillero "sample variables" el nombre de la variable que estamos estudiando (en este ejemplo Niquel) y en el casillero "null hypothesis" el valor propuesto en H_0 (en este caso 4.44) y obtenemos

```
ONE-SAMPLE T TEST FOR NIQUEL

NULL HYPOTHESIS: MU =  4.44
ALTERNATIVE HYP: MU <> 4.44

MEAN          4.3400
STD ERROR     0.0393
MEAN - H0    -0.1000
LO 95% CI    -0.1889
UP 95% CI    -0.0111
T             -2.54
DF                9
P             0.0315

CASES INCLUDED 10
```

Comentario: Aunque generalmente se usa 0.05 como punto de corte para P, también podría usarse otro (0.01 o 0.10). Llamemos α a ese punto de corte.

Errores tipo I y tipo II.

En todo problema de test de hipótesis se plantean dos hipótesis y, una vez observada la muestra se RECHAZA H_0 o NO. Entonces puede ocurrir alguna de estas cuatro situaciones:

	REALIDAD	
Se aplica el test y	**H_0 es cierta**	**H_1 es cierta**
Se acepta H_0	Bien!	Error tipo II
Se rechaza H_0	Error tipo I	Bien!

Como se aprecia en la tabla anterior, pueden cometerse dos tipos de errores, que se los distingue con los nombre de error tipo I y tipo II.

En el ejemplo 1, dijimos que si H_0 fuese cierta podría dar un valor de T en las colas y rechazar, pero que este evento tiene probabilidad 0.05. Si hubiésemos usado otro punto de corte (α) para el valor P ese sería la probabilidad de error tipo I.

Para cualquier test: **la probabilidad de error tipo I es ≤ α (donde α es el valor de corte que se elija para el valor de P). Este valor α se suele llamar "nivel de significación" del test.**

La probabilidad de error tipo II se suele llamar β y es más difícil de calcular.

Entonces al elegir el punto de corte para el valor P (generalmente 0.05) estamos eligiendo la probabilidad de error tipo I. La probabilidad de error tipo II es más difícil de calcular y puede ser grande si el tamaño de muestra (n) es pequeño.

Si el tamaño de muestra aumenta, la probabilidad de error tipo I se mantiene en el 5% (porque yo la fijo así). Es intuitivamente esperable (y así ocurre) que, cuando el tamaño de la muestra aumenta, la probabilidad de error tipo II disminuye y se acerca a cero cuando la muestra es muy grande. Como consecuencia de esto, puede calcularse un tamaño de muestra para lograr que la probabilidad de error tipo II sea la deseada. Veremos más adelante algún ejemplo de cálculo de probabilidad de error tipo II y de cálculo de tamaño de muestra.

La idea del test que aplicamos para el ejemplo de las mediciones de níquel es válida para cualquier hipótesis sobre la media de una población, basado en una muestra normal. Generalicemos entonces.

Tests acerca de la media basado en una muestra normal con varianza desconocida.

Hipótesis bilaterales:

Problema: quiero decidir entre dos hipótesis

$$H_0 : \mu = \mu_o$$
$$H_1 : \mu \neq \mu_o$$

donde μ_o es un valor propuesto (antes de observar la muestra).

Elijo un valor de corte para P que llamaremos α (generalmente α=0.05). Observo una muestra, calculo $\overline{X}$ y s y aplico el siguiente test:

Test:

1er. paso. Calculo

$$T = \frac{\overline{X} - \mu_o}{s/\sqrt{n}}$$

Comentario: si H_0 es cierta, T tiene distribución t de Student con n-1 grados de libertad.

2do. paso. Calculo el valor P que es el área bajo las dos colas de la función de densidad t_{n-1} a partir del valor de T calculado en el paso anterior.

3er. paso. Si $P < \alpha$ rechazo H_0 o equivalentemente afirmo que la diferencia es estadísticamente significativa.

Comentario: el valor de T que se calcula en el primer paso se llama "el estadístico del test".

Hipótesis unilaterales:

$$H_0 : \mu = \mu_o$$
$$H_1 : \mu < \mu_o$$

ó

$$H_0 : \mu = \mu_o$$
$$H_1 : \mu > \mu_o$$

Si la hipótesis alternativa es unilateral, todos los cálculos son similares, salvo que el valor P es el área bajo una cola (la cola de la izquierda si la hipótesis alternativa es "$\mu < \mu_o$", la de la derecha si la hipótesis alternativa es "$\mu > \mu_o$").

Advertencia: La elección de aplicar un test a 1 o 2 colas tienen que ser hecha **antes** de observar los datos. Los tests a dos colas son más usados y tienen la ventaja de que se puede informar que existe diferencia significativa, tanto cuando la media muestral observada es menor que la propuesta en la hipótesis nula, como cuando es mayor.

Hemos presentado hasta ahora solamente el test sobre la media basado en una muestra de una distribución normal. Hay muchos otros tests. El estadístico del test y la distribución que se usa para calcular el valor P son diferentes para cada caso. La elección del test depende del modelo probabilístico que propondremos según el tipo de datos que estamos analizando y de las hipótesis H_0 y H_1.

Pero todos los tests tienen muchas cosas en común. Siempre se plantean dos hipótesis. Se pueden cometer dos tipos de errores. Se fija (generalmente en 5%) la probabilidad de error tipo I, la probabilidad de error tipo II suele ser difícil de calcular y puede ser muy grande para muestras pequeñas. Si el tamaño de muestra aumenta, la probabilidad de error tipo II disminuye y tiende a cero cuando n. El valor de P siempre puede interpretarse como la probabilidad de observar nuestra muestra o una muestra aún mas alejada de H_0, si H_0 fuese cierta.

Como la probabilidad de error tipo I esta controlada (=5%) mientras que la de tipo II no es tan fácil de controlar y puede ser grande para muestras pequeñas, **rechazar H_0** (y por lo tanto elegir H_1) **es una afirmación más fuerte que aceptar H_0**. Por lo tanto "lo que se quiere demostrar" conviene (si se puede) ponerlo en H_1. No siempre se puede hacer esta elección; éste es el problema que hace que el test de normalidad de Shapiro-Wilk (es un test donde se pone como H_0 que la distribución es gaussiana, ver Statistix, Statistics, Normality Tests) no sea muy satisfactorio: para muestras pequeñas puede tener mucha probabilidad de error tipo II y por lo tanto ser poco "potente" para detectar falta de normalidad.

Tests acerca de la media basado en una muestra normal con varianza conocida.

El modelo normal con varianza conocida es más simple que el anterior (hay un sólo parámetro desconocido que es μ) pero menos usado en la práctica. Podría usarse en el ejemplo de las mediciones de níquel si, por la experiencia previa, suponemos conocida la precisión del método de medición: (la desviación estándar σ). Lo único que no sabemos es si el método es exacto o tiene error sistemático.

Modelo: $X_1, X_2, ..., X_n$ vs. as. i.i.d $N(\mu,\sigma^2)$ con σ conocido

Hipótesis:

Bilateral:

$$H_0 : \mu = \mu_o \qquad H_1 : \mu \neq \mu_o$$

Unilaterales:

$$H_0 : \mu = \mu_o \qquad H_1 : \mu < \mu_o$$

ó

$$H_0 : \mu = \mu_o \qquad H_1 : \mu > \mu_o$$

Test para la hipótesis bilateral:

1er. paso. Calculo el estadísticos del test:

$$Z = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}$$

Comentario: si H_0 es cierta, Z tiene distribución N(0.1).

2do. paso. Calculo el valor P que es el area de las dos colas de la función de densidad N(0,1) a partir del valor de Z calculado en el paso anterior.

3er. paso. Si el valor $P < \alpha$ rechazo H_0 o equivalentemente afirmo que la diferencia es estadísticamente significativa.

Pensar: ¿Qué hay que cambiar en el test si la hipótesis es **unilateral**?

Ejemplo 2: (ejemplo de test para la media de una muestra normal con varianza conocida):
Supongamos ahora que por alguna medición previa ya sospechábamos que el método de medición de níquel tenía error sistemático negativo (estaba subestimando la cantidad de níquel). Además sabíamos por haber hecho muchas determinaciones del mismo material (aunque no supiesemos el verdadero contenido de niquel) que la DS del método es $\sigma = 0.12$.
Es con este conocimiento previo que realizamos las 10 mediciones de un material que sabemos que tiene 4.44% de níquel.
Planteamos las siguientes hipótesis:

$$H_o : \mu = 4.44 \qquad H_1 : \mu < 4.44$$

que se interpretan como “el método de medición no tiene error sistemático” y “el método tiene error sistemático negativo” respectivamente.

Aplicamos el test correspondiente a este modelo y estas hipótesis:

1er. paso.

$$Z = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} = \frac{4.34 - 4.44}{0.12 / \sqrt{10}} = -2.63$$

2do. paso. Calculo el valor P que es el area de la cola "a izquierda" bajo la curva N(0,1). Esto lo podemos hacer con la tabla normal o con Statistix, "Statistics", "Probability functions", "Z1-Tail", X=-2.63. El resultado es 0.00427

3er. paso. Como P=0.004 < 0.05 rechazo H_0 .

Concluímos que el método de medición tiene error sistemático negativo.

Otra forma equivalente de plantear los tests.

Veremos una forma equivalente de plantear los tests. Pensemos por ejemplo en el test sobre μ para una muestra normal con σ conocido, a dos colas .
Hemos rechazado cuando $P<\alpha$. Como

$P < 0.05 \Leftrightarrow$ Z está en alguna de las dos "colas" de la curva N(0,1) que tienen área 0.05
$\Leftrightarrow |Z| > 1.96$

o en general

$P < \alpha \Leftrightarrow$ Z está en alguna de las dos "colas" de la curva N(0,1) que tienen área α
$\Leftrightarrow |Z| > z_{\alpha/2}$

Entonces otra forma de describir el test es:

Test sobre la media de una muestra normal con σ conocido:

$$H_0 : \mu = \mu_o$$

Test:
1er. paso. Calculo el estadístico del test:

$$Z = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}$$

2do. paso. Según el test sea a una o dos colas

- Para el caso $H_1 : \mu \neq \mu_o$ (test a dos colas)
  Rechazo H_o si $|Z| > z_{\alpha/2}$

- Para el caso $H_1 : \mu < \mu_o$
  Rechazo H_o si $Z < -z_{\alpha}$

- Para el caso $H_1: \mu > \mu_o$
  Rechazo H_o si $Z > z_\alpha$

Región de rechazo de un test:

Se llama así al conjunto de valores tal que si el estadístico del test pertenece a ese conjunto, se rechaza H_o.
Por ejemplo en el test anterior, para el caso de de la hipótesis unilateral $H_1: \mu < \mu_o$, la región de rechazo es la semirrecta $(-\infty, -z_\alpha)$. Para la hipótesis bilateral $H_1: \mu \neq \mu_o$ la región de rechazo son las dos semirrectas (las dos "colas"): $(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$

Un ejemplo de cálculo de probabilidad de error tipo II.

Continuemos con el ejemplo 2. Si H_o es cierta (el método de medición no tiene error sistemático) la probabilidad de equivocarnos y decir que lo tiene (probabilidad de error tipo I) es 0.05
¿Cuánto vale la probabilidad de error tipo II? Es la probabilidad de aceptar H_0 cuando es falsa, pero ¿que quiere decir que H_0 sea falsa? Quiere decir que $\mu < 4.44$. Esta no es una hipótesis "puntual" y la probabilidad de error tipo II depende de cuál sea el verdadero valor de μ. Intuitivamente ¿vale más si μ está cerca o lejos de 4.44?

Calcular lo siguiente:
a) probabilidad de error tipo II si el verdadero valor de $\mu = 4.34$

b) una expresión que permita calcular la probabilidad de error tipo II para cualquier valor de $\mu < 4.44$ (esta función del verdadero valor de μ se suele notar $\beta(\mu)$).

c) probabilidad de error tipo II si el verdadero valor de $\mu = 4.40$

Saber calcular la probabilidad de error tipo II de un test permite también determinar el tamaño de la muestra en la etapa del diseño del experimento. Por ejemplo:
d) ¿Cuanto debe valer n para que la probabilidad de error tipo II, si el verdadero valor de μ es 4.40, sea menor o igual que 0.10?

Respuestas:
a) 0.16
b) $1 - \Phi\left(-z_\alpha + \frac{\mu_o - \mu}{\sigma/\sqrt{n}}\right)$)
c) 0.72
d) n=77