Técnicas de Inferencia Estadística II

Tema 1. Introducción a los contrastes de hipótesis

Conchi Ausín
Universidad Carlos III de Madrid

Grado en Estadística y Empresa
Curso 2014/15

Contenidos

Contraste de hipótesis

Hipótesis estadística

Una hipótesis estadística es una afirmación respecto a una característica de una población.

Esta hipótesis puede ser cierta o no. La mejor manera de averiguar si una hipótesis es cierta o no sería examinar toda la población. Como en general esto no es factible, se toma una muestra aleatoria de la población y se realiza un contraste de hipótesis.

Contraste o test de hipótesis

Un contraste de hipótesis es un procedimiento formal para rechazar o no una hipótesis estadística planteada sobre una población utilizando para ello una muestra de observaciones.

Ejemplo 1.1.

Se desea saber si el precio medio por comensal en el "Asador Felipe" es mayor de los 30 euros que se indica en la Guía del Ocio. Para ello, se toma una muestra del gasto de 10 clientes independientes que hayan comido en este asador en distintos días.

Hipótesis nula y alternativa

- La hipótesis nula, H_0, es la hipótesis por defecto.
  - H_0 se mantiene a no ser que los datos indiquen su falsedad.
  - H_0 nunca se considera probada.
  - H_0 se rechaza si la muestra de datos proporciona evidencias de que es falsa.
  - H_0 siempre tiene un sólo elemento, por ejemplo, $\mu = 7$.
- La hipótesis alternativa, H_1, es habitualmente la que el investigador quiere demostrar como cierta.
  - Cuando se rechaza H_0, se admite que H_1 es cierta.
  - Si no se especifica H_1 de manera explícita, se considera definida implícitamente como H_0 es falsa.
  - H_1 nula suele contener más de un elemento, por ejemplo, $\mu > 7$.

Hipótesis nula y alternativa

En un contraste de hipótesis no se trata de juzgar cuál de las dos hipótesis, H_0 ó H_1, es más verosímil, sino de decidir si la muestra proporciona o no evidencia suficiente para rechazar H_0. Cuando se rechaza H_0, se admite que H_1 es cierta.

Ejemplo 1.2.

Se desea saber si el precio medio por comensal en el "Asador Felipe" es mayor de los 30 euros que se indica en la Guía del Ocio. Para ello, se toma una muestra del gasto de 10 clientes independientes que hayan comido en este asador en distintos días. El objetivo es contrastar:

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

donde μ es la esperanza de la variable aleatoria, X, que representa el precio de la comida de un cliente tomado al azar en el "Asador Felipe".

Tipos de contrastes

Una test de hipótesis puede ser:

1.
   - Relativo a una única variable aleatoria, X, de la que se tiene una muestra $(X_1, X_2, \ldots, X_n)$.
   - Relativo a más de una variable, por ejemplo:
     - Dos variables independientes, X e Y, de las que se tienen dos muestras $(X_1, X_2, \ldots, X_n)$ e $(Y_1, Y_2, \ldots, Y_m)$.
     - Una variable bivariante, (X, Y), de la que se tiene una muestra emparejada: $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
2.
   - Paramétrico: Se asume un modelo paramétrico sobre la variable, por ejemplo, se asume que $X \sim N(\mu, \sigma^2)$.
   - No paramétrico: No se asume ningún modelo paramétrico sobre la distribución de X.

Ejemplo 1.3.

Plantear un test de hipótesis para contrastar si....

1. *El precio medio de alquiler un piso de dos habitaciones en un barrio es mayor a los 30 euros mensuales que pide un arrendador.*
2. *El gasto medio mensual de los clientes de una compañía telefónica es diferente en clientes de tarjeta que de contrato.*
3. *El gasto medio mensual por familia en comida es que mayor que en ropa y calzado.*
4. *En unas elecciones municipales, la proporción de votantes a un partido político es mayor del 25 %.*
5. *La distribución de los salarios en un país es diferente para hombres que para mujeres.*
6. *El ancho de los berberechos de la Ría de Arousa no sigue una distribución normal.*
7. *El sueldo de los empleados de una empresa depende de la edad.*

Errores de tipo I y de tipo II

Para cualquier decisión que se tome, hay cuatro posibilidades:

	H_0 es cierta	H_0 es falsa
Rechazar H_0	Error tipo I (α)	Decisión correcta
No rechazar H_0	Decisión correcta	Error tipo II (β)

Lo ideal sería encontrar un test que hiciese mínima las probabilidades de ambos errores. Sin embargo, esto no es posible ya que la reducción de la probabilidad de un tipo de error hace que aumente la probabilidad del otro tipo de error.

Como se considera que el error más grave es el de tipo I, se le impondrá una cota, α, llamada nivel de significación, (normalmente 0.05 ó 0.01):

$$\Pr(\text{Rechazar } H_0 \mid H_0 \text{ cierta}) = \Pr(\text{Error tipo I}) \leq \boldsymbol{\alpha}$$

Potencia de un contraste

La probabilidad de cometer un error de tipo II se denota por:

$$\beta = \text{Pr(Error tipo II)} = \text{Pr(No rechazar } H_0 \mid H_0 \text{ falsa)}$$

La potencia de un contraste es la probabilidad de NO cometer un error de tipo II:

$$\text{Potencia} = 1 - \beta = \text{Pr(Rechazar } H_0 \mid H_0 \text{ falsa)}$$

- La potencia de un contraste disminuye cuando α disminuye.
- La potencia de un contraste aumenta cuanto mayor es la diferencia entre H_0 y la realidad.
- La potencia de un contraste aumenta cuanto mayor es el tamaño de la muestra.

Recapitulando...

- Un contraste de hipótesis puede rechazar la hipótesis nula.
- Un contraste de hipótesis no puede probar la hipótesis nula.
- Si rechazamos H_0, es porque se está razonablemente seguro de que es falsa ya que:
$$\Pr(\text{Rechazar } H_0 \mid H_0) \leq \alpha,$$
y estamos aceptando implícitamente la hipótesis alternativa.
- Si por el contrario, no rechazamos H_0 debe interpretarse como que las observaciones no han aportado evidencia para descartarla.
- Se puede tomar α tan pequeña como se quiera, pero esto hará que aumente la probabilidad de error de tipo II y disminuya la potencia del contraste.

Metodología del contraste

- Es necesario desarrollar una regla de decisión para rechazar H_0.
- La primera regla consiste en definir una región de rechazo, R, tal que:
    - Si $\{X_1, \dots, X_n\} \in R$, se rechaza H_0.
    - Si $\{X_1, \dots, X_n\} \notin R$, no se rechaza H_0.
- La región R tiene que considerar el signo de H_1 que lleva a definir regiones de rechazo en una o dos direcciones.
- La región R tiene que ser tal que:

$$\Pr(R \mid H_0) \leq \alpha$$

Por tanto, es necesario conocer la probabilidad de R cuando H_0 es cierta. Para ello, se puede asumir un modelo paramétrico sobre los datos.

Ejemplo 1.4.

Consideramos de nuevo el siguiente contraste sobre el precio medio de un restaurante.

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

- *Parece natural rechazar H_0 si $\bar{X}$ es mucho mayor que 30 euros.*
- *Queremos poder calcular $\Pr(R \mid H_0)$ y determinar R para que esta probabilidad sea menor que α.*
- *Tiene sentido asumir que el precio de la comida en el asador, X, sigue una distribución normal, $X \sim N(\mu, \sigma^2)$.*

Metodología del contraste

- En realidad, lo que hemos hecho en el ejemplo anterior es considerar una medida de discrepancia entre las observaciones y la hipótesis nula, de modo que si la discrepancia es grande, se rechaza H_0.
- En contrastes paramétricos, es habitual tomar medidas de discrepacia basadas en un estimador del parámetro de interés:

$$Discrepancia = \frac{Estimador - Parámetro}{Desviación\ típica\ del\ estimador}$$

- La medida de discrepancia se denomina estadístico de contraste y tiene que ser tal que su distribución es conocida si H_0 es cierta.

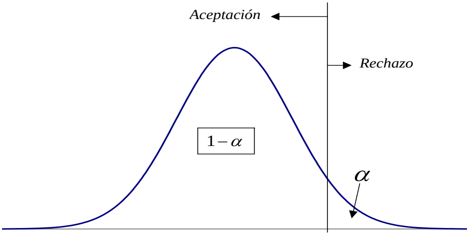
Ejemplo 1.5.

En el ejemplo anterior, si asumimos que $X \sim N(\mu, \sigma^2)$ *y si* H_0 *es cierta:*

$$\frac{\bar{X} - 30}{\sigma/\sqrt{n}} \sim_{H_0} N(0,1)$$

Así, para que el error de tipo I sea igual a un nivel de significación, α*, fijado previamente, la región de rechazo más razonable es:*

$$R = \left\{ \frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_\alpha \right\}.$$



Ejemplo 1.6.

Se desea saber si el precio medio por comensal en el "Asador Felipe" es mayor de los 30 euros que se indica en la Guía del Ocio. Para ello, se toma una muestra del gasto de 10 clientes independientes que hayan comido en este asador en distintos días dando lugar a una media muestral de $\bar{x} = 33.3$ *euros. Suponiendo que el precio de la comida en el asador,* X*, sigue una distribución normal,* $X \sim N(\mu, \sigma^2)$*, y que la desviación típica es conocida (no realista) e igual a 5 euros.*

Resolver el siguiente contraste al nivel $\alpha = 0.05$*:*

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

Metodología del contraste

Recapitulando, los pasos para resolver un contraste de hipótesis son:

1. Plantear las hipótesis nula, H_0, y alternativa, H_1.
2. Definir un estadístico de contraste cuya distribución sea conocida cuando H_0 es cierta.
3. Fijar un nivel de significación, α, (normalmente 0.05 ó 0.01) y determinar la región de rechazo, R, tal que $\Pr(R \mid H_0) \leq \alpha$.
4. Tomar una muestra de datos y calcular el valor del estadístico de contraste. Si éste valor está en R, se rechaza H_0.

Nivel crítico o p-valor

El procedimiento para obtener la región de rechazo usando el nivel de significación tiene dos incovenientes:

1. El resultado del test puede depender mucho del valor de α.
2. Dar sólo el resultado del test no permite diferenciar el grado de evidencia que la muestra indica a favor o en contra de H_0.

Ejemplo 1.7.

Resolver el contraste del ejemplo anterior al nivel $\alpha = 0.01$*:*

$$\begin{aligned} H_0 &: \mu = 30 \\ H_1 &: \mu > 30 \end{aligned}$$

Nivel crítico o p-valor

El nivel crítico p o p-valor es es la probabilidad de encontrar una discrepancia mayor o igual que la observada en los datos, cuando H_0 es cierta.

- El p-valor no se fija a priori, sino que depende de los datos.
- Usando el p-valor podemos resolver el contraste para cualquier α:
  - Si $\alpha >$ p-valor, se rechaza H_0 al nivel α.
  - Si $\alpha <$ p-valor, no se rechaza H_0 al nivel α.
- El p-valor es el nivel de significación más pequeño para el que la muestra obtenida obligaría a rechazar la hipótesis nula.

El p-valor se puede interpretar como una medida de la evidencia a favor de la hipótesis nula.

Nivel crítico o p-valor

Ejemplo 1.8.

Se desea saber si el precio medio por comensal en el "Asador Felipe" es mayor de los 30 euros que se indica en la Guía del Ocio. Para ello, se toma una muestra del gasto de 10 clientes independientes que hayan comido en este asador en distintos días dando lugar a una media muestral de $\bar{x} = 33.3$ *euros. Suponiendo que el precio de la comida en el asador,* X*, sigue una distribución normal,* $X \sim N(\mu, \sigma^2)$*, y que la desviación típica es conocida (no realista) e igual a 5 euros.*

Resolver el siguiente contraste para cualquier nivel usando el p-valor.

$$
\begin{aligned}
H_0 &: \mu = 30 \\
H_1 &: \mu > 30
\end{aligned}
$$

Nivel crítico o p-valor

Recapitulando, los pasos para resolver un contraste de hipótesis utilizando el p-valor son:

1. Plantear la hipótesis nula, H_0, y alternativa, H_1.
2. Definir un estadístico de contraste cuya distribución sea conocida si H_0 es cierta.
3. Fijar un nivel de significación, α.
4. Calcular el p-valor. Si este es suficientemente pequeño (menor que α), rechazar H_0.