Probabilidad y Estadística

Agustín G. Bonifacio

UNSL

Estadística: Test de Hipótesis y Regresión Lineal

Consideremos la siguiente situación:

Una compañía farmacéutica está fermentando un tanque de antibiótico. Se toman muestras del tanque para *estimar* la potencia media μ para todo el tanque. Pero el real interés de la empresa es que el antibiótico cumpla la potencia mínima reglamentaria. Por consiguiente lo que realmente la farmacéutica es decidir entre las siguientes dos posibilidades:

1. La potencia media μ no excede la potencia aceptable mínima.
2. La potencia media μ sí excede la potencia aceptable mínima.

Definición

Una prueba (test) de hipótesis consta de cinco partes:

1. La hipótesis nula, H_0,
2. La hipótesis alternativa, H_a,
3. El estadístico de prueba y su valor p,
4. La región de rechazo, y
5. La conclusión.

Observación

Las dos hipótesis en competencia son la *hipótesis alternativa,* que en general se quiere apoyar, y la *hipótesis nula*, que la contradice.

Ejemplo (prueba de dos colas)

Se desea mostrar que el salario promedio por hora de los obreros de la construcción en California es distinto de 14. Entonces:

$$H_a: \ \mu \neq 14$$

y

$$H_0: \ \mu = 14.$$

Ejemplo (prueba de una cola)

Un proceso de laminado produce en promedio $3\,\%$ de piezas defectuosas. Queremos demostrar que un ajuste en una máquina disminuirá la proporción de piezas defectuosas, p. Entonces:

$$H_a: \ p < 0,03$$

y

$$H_0: \ p = 0,03.$$

La decisión de rechazar o aceptar la hipótesis nula se basa en la información que contiene una muestra de la población de interés. Esta información viene dada por:

1. El estadístico de prueba: en general un estimador muestral.
2. El valor p: una *probabilidad* calculada mediante el estadístico de prueba.

Prueba estadística con muestras grandes para μ (Prueba de una cola)

1. Hipótesis nula: $H_0: \ \mu = \mu_0$
2. Hipótesis alternativa: $H_a: \ \mu > \mu_0$ (o $H_a: \ \mu < \mu_0$)
3. Estadístico de prueba: $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ (si no se conoce σ usar s)
4. Región de rechazo: dependiendo el nivel de significación α

$$\text{si } z > z_\alpha \text{ (o } z < -z_\alpha \text{ cuando } H_a: \ \mu < \mu_0).$$

Prueba estadística con muestras grandes para μ (Prueba de dos colas)

1. Hipótesis nula: $H_0 : \ \mu = \mu_0$
2. Hipótesis alternativa: $H_a : \ \mu \neq \mu_0$
3. Estadístico de prueba: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ (si no se conoce σ usar s)
4. Región de rechazo: dependiendo el nivel de significación α

$$\text{si } z > z_{\alpha/2} \text{ o } z < -z_{\alpha/2}.$$

Definición

El valor p o nivel de significación observado es el valor más pequeño de α con el cual puede rechazarse H_0.

Proposición

- Si $p < 0,01$, se rechaza H_0. Los resultados son muy significativos.
- Si $0,01 \leq p < 0,05$, se rechaza H_0. Los resultados son estadísticamente significativos.
- Si $0,05 \leq p < 0,1$, generalment no se rechaza H_0. Los resultados sólo tienden a la significación estadística.
- Si $p > 0,1$, H_0 no se rechaza y los resultados no son estadísticamente significativos.

- ¿Cómo se puede construir un modelo **poblacional** para describir la relación entre una v.a. y una variable independiente relacionada?
- Postulamos un modelo probabilístico a través dela siguiente ecuación:

$$y = \alpha + \beta x + \epsilon$$

donde los valores ϵ son términos de error que cumplen los siguientes supuestos:

  - son independientes,
  - $E(\epsilon) = 0$ y $Var(\epsilon) = \sigma^2$,
  - $\epsilon \sim N(0, \sigma^2)$.

- Podemos utilizar información de una muestra de los valores de x e y para estimar los valores de α y β.
- Una forma de hacer esto consiste en ajustar una recta que minimice la suma de los cuadrados de los errores.

Dado el modelo probabilístico $y = \alpha + \beta x + \epsilon$ y una muestra $(x_i, y_i)_{i=1}^n$, la recta de mejor ajuste

$$\hat{y} = a + bx$$

se obtiene minimizando la suma de los cuadrados del error (SSE), esto es, resolviendo el problema de minimización

$$\min_{a,b} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Los valores de a y b que resuelven el problema son:

$$b = \frac{S_{xy}}{S_{xx}} \quad \text{y} \quad a = \bar{y} - b\bar{x}$$

donde

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

y

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

Observación

A la hora de hacer una regresión lineal, es útil realizar una tabla con los siguientes valores para hacer los cálculos de a y b :

$$y_i, x_i, x_i^2, x_i y_i, y_i^2$$

con sus respectivas sumas al pie.

- ¿Qué tan bien ajusta el modelo de regresión a los datos?
- Para responder podemos utilizar una medida relacionada al *coeficiente de correlación (muestral).*

Coeficiente de Correlación, r:

$$r = \frac{S_x y}{\sqrt{S_{xx} S_{yy}}}$$

Definición

El coeficiente de determinación, r^2, es el cuadrado del coeficiente de correlación r y se puede interpretar como la reducción porcentual en la variación total en el experimento obtenida al usar la recta de regresión $\hat{y} = a + bx$.

Observación

- $r^2 \to 1$: no hay variación aleatoria y todos los puntos caen en la recta de regresión.
- $r^2 \to 0$: los puntos están dispersos de manera aleatoria y la regresión no explica los datos.