Inferential Statistics, Test Statistics Manuel

Tingfeng Xia

2019, Winter Term

Contents

1 Test for $\mu = \mu_0$, w/σ^2 known

Assume that $X_i \sim N(\mu, \sigma^2)$ are i.i.d, then the test statistic is

$$T(X) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

then under α significance level, we have the rejection region

$$R_\alpha(T) = (-\infty, z_{\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$$

2 Test for $\mu = \mu_0$, w/σ^2 unknown

Assume that $X_i \sim N(\mu, \sigma^2)$ are i.i.d, then the test statistic is

$$T(X) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

then under α significance level, we have the rejection region

$$R_\alpha(T) = (-\infty, t_{\frac{\alpha}{2}, df=n-1}) \cup (t_{1-\frac{\alpha}{2}, df=n-1}, \infty)$$

3 Test for $\sigma^2 = \sigma_0^2$

Assume that $X_i \sim N(\mu, \sigma^2)$ are i.i.d, then the test statistic is

$$T(X) = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{df=n-1}$$

and the α significance level rejection region is

$$R_\alpha(T) = (-\infty, \chi^2_{\frac{\alpha}{2}, df=n-1}) \cup (\chi^2_{1-\frac{\alpha}{2}, df=n-1}, \infty)$$

4 Equality of Variances $\sigma_x = \sigma_y$

If we have $X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_n \sim N(\mu_y, \sigma_y^2)$, then under our null hypothesis

$$T(X, Y) = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \frac{S_x^2}{S_y^2} \sim F_{(n-1)(m-1)}$$

With α significance level, we then have the rejection region

$$R_\alpha(T) = \Big(-\infty, F_{\frac{\alpha}{2}(n-1)(m-1)}\Big) \cup \Big(F_{1-\frac{\alpha}{2}(n-1)(m-1)}, \infty\Big)$$

5 Equality of $\mu_x = \mu_y$, w/σ_x, σ_y known

If we have $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ and $\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right)$, then

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$$

6 Equality of $\mu_x = \mu_y$, w/$\sigma = \sigma_x = \sigma_y$ known

If this is the case, we can pull the $\sigma_x = \sigma_y = \sigma$ out from the above equation, we will have

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim N(0,1)$$

7 Equality of $\mu_x = \mu_y$, w/$\sigma = \sigma_x = \sigma_y$ unknown

If we have $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ and $\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$, then

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$$

where, S_p is the polled sample standard deviation, the square root of of the polled sample variance, defined as

$$S_p^2 := \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

8 Equality of $\mu_x = \mu_y$, w/σ_x, σ_y unknown

In this case, we have a messy formula for the degrees of freedom, the test statistics that we use stays the same as above.

9 Equality of $\mu_x = \mu_y$ for paired data

We set the $H_0 : \mu_x - \mu_y = 0$, define $D = X - Y$, then $\mu_d = \mu_x - \mu_y$. Notice that $\mu_d = 0 \iff \mu_x = \mu_y$, then our test statistic is

$$T(D) = \frac{\bar{D}}{s_d/\sqrt{n}} \sim t_{n-1}$$

10 Likelihood Ratio Test

Define

$$\Lambda := \frac{max_{\theta\in\Omega_0}[L(\theta)]}{L(\hat{\theta})} = \frac{\text{Restricted Likelihood}}{\text{Unrestricted Likelihood}}$$

Denoting $p = \dim\Omega =$ number of free var in the whole space, and $d = \dim\Omega_0 =$ number of free var under our null hypothesis, we have

$$T(X) = -2\ln\Lambda \xrightarrow{D} \chi^2_{df=p-d}$$

11 Likelihood Ratio Test for Equality of $\mu_x = \mu_y$ for Normally Distributed Random Variables

Consider i.i.d $X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ and i.i.d $Y_1, \ldots, Y_m \sim N(\mu_y, \sigma_y^2)$. Notice that we have $p - d = 2 - 1 = 1$ in this case, and the likelihood is

$$L(\mu_x, \mu_y) = \left\{ \left(2\pi\sigma_x^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_x^2}\sum_i (X_i-\mu_x)^2} \right\} \left\{ \left(2\pi\sigma_y^2\right)^{-\frac{m}{2}} e^{-\frac{1}{2\sigma_y^2}\sum_i (Y_i-\mu_y)^2} \right\}$$

by re-writing with $H_0 : \mu = \mu_x = \mu_y$, we have

$$L(\mu) = \left\{ \left(2\pi\sigma_x^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_x^2}\sum_i (X_i-\mu)^2} \right\} \left\{ \left(2\pi\sigma_y^2\right)^{-\frac{m}{2}} e^{-\frac{1}{2\sigma_y^2}\sum_i (Y_i-\mu)^2} \right\}$$

Our test statistic is then

$$T(X, Y) = -2\ln\Lambda = -2\ln\frac{L(\hat{\mu})}{L(\hat{\mu}_x, \hat{\mu}_y)} \sim \chi^2_{df=1}$$

where we remind ourselves that $\hat{\mu}_x = \bar{x}$ and $\hat{\mu}_y = \bar{y}$ here and $\hat{\mu}$ is some wright-ed average of $\bar{x}$ and $\bar{y}$ as below

$$\hat{\mu} = \left(\frac{\frac{1}{\sigma_x^2/n}}{\frac{1}{\sigma_x^2/n} + \frac{1}{\sigma_y^2/m}} \right) \bar{x} + \left(\frac{\frac{1}{\sigma_y^2/m}}{\frac{1}{\sigma_x^2/n} + \frac{1}{\sigma_y^2/m}} \right) \bar{y}$$

12 Chi-Square Test of Goodness of Fit

12.1 Known categorical probabilities

[1] Suppose, $X_1, X_2, \ldots, X_k$ are the observed counts of category $1, 2, \ldots, k$ respectively. Then

$$(X_1, X_2, \ldots, X_k) \sim \text{Mult}(n, p_1, p_2, \ldots, p_k) \quad \text{where } E[X_i] = np_i(\geq 1), \forall i$$

and our test statistic will be, in this case

$$T(X) = X^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \xrightarrow{D} \chi^2_{(df=k-1)}$$

[1] This is usually used to test parameters.

12.2 Unknown categorical probabilities

[2] In this case, we have $X_1, \dots, X_k \sim \text{Mult}(n, p_1(\theta), \dots, p_k(\theta))$ and the test statistic will then be

$$T(X) = X^2 = \sum_{i=1}^{k} \frac{\left(X_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})} \xrightarrow{D} \chi^2_{(df=k-1-\dim\Omega)}$$

where $\dim\Omega$ is the number of params that need to be estimated to calculate $p_1, \dots, p_k$.

13 Chi-square Test of Independence

Suppose that we have two categorical random variables X, Y. Let $i = 1, \dots, a$ and $j = 1, \dots, b$ represent the categories of X and Y respectively. Let f_{ij} represent the number of samples corresponding to the i-th of X and j-th of Y, notice that $\sum_{ij} f_{ij} = n$. Let F_{ij} represent the random variable corresponding to the cell at position (i, j). Let $\theta_{ij} = P(X = i, Y = j)$, then we have To access the null hypothesis, $H_0 : X \perp Y$, we have $P(X = i, Y = j) = P(X = i)P(Y = j)$, so $\theta_{ij} = \theta_{i.}\theta_{.j}$[3], and our statistics follows

$$F_{11}, F_{12}, \dots, F_{ab} \sim \text{Mult}(n, \theta_{1.}\theta_{.1}, \theta_{1.}\theta_{.2}, \dots, \theta_{a.}\theta_{.b})$$

By using the MLEs

$$\hat{\theta}_{i.} = \sum_{j=1}^{b} f_{ij}/n$$
$$\hat{\theta}_{.j} = \sum_{i=1}^{a} f_{ij}/n$$

we have our test statistic

$$T(X, Y) = X^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(f_{ij} - n\hat{\theta}_{.j}\hat{\theta}_{i.}\right)^2}{n\hat{\theta}_{.j}\hat{\theta}_{i.}} \xrightarrow{D} \chi^2_{df=(a-1)\times(b-1)}$$

Note. In performing such test, we need to calculate the expected count of each slot and there is a neat formula for this

$$E_{ij} = \frac{i\text{-th row total} * j\text{-th column total}}{\text{grand total of the table}}$$

[2] This is usually used to test the fitting of a model. For example we wish to test if the out come of a dice(with unknown probabilities) is binomial, we can estimate, using MLE, the probabilities and the proceed as the previous case.

[3] The "." here is a wildcard.

14 Chi-square Test of Homogeneity

Let n_i be the marginal total of $X = i$ category, then we have $\sum_i n_i = n$. Notice that this is different from the test of independence above in the sense that we are fixing marginal totals of all categories of X before hand. We wish to test the hypothesis $H_0 : \theta_{j|X=1} = \theta_{j|X=2} = \ldots = \theta_{j|X=a} = \theta_j$. Using the MLE

$$\hat{\theta}_j = \sum_{i=1}^{a} f_{ij}/n$$

we have our test statistic

$$T(X,Y) = X^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(f_{ij} - n_i\hat{\theta}_j\right)^2}{n_i\hat{\theta}_j} \xrightarrow{D} \chi^2_{df=(a-1)\times(b-1)}$$

Again, we calculate the E_{ij} using the formula above. (and this is a coincidence.)

15 Discrepancy Statistic for Normal R.V.s

Consider $X_1, \ldots, X_n \sim N(\hat{\mu}, \sigma_0^2)$ where σ_0^2 is known. Define $R = X_i - \bar{X}$, where $R \sim N(0, \sigma_0^2(1 - \frac{1}{n}))$ then, the descrepancy statistic is defined as

$$D(R) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} R_i^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

16 Simple Linear Regression

Consider the data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we want to find the line that fits the best for our set of data. Let the hypothetical line be $y = b_1 + b_2x$, then we define

$$res_i = (y_i - b_1 - b_2x_i) \text{ is the deviation of } y_i \text{ from the line}$$

and to minimize $\sum_{i=1}^{n} res_i^2$ we need

$$b_1 = \overline{y} - b_2\overline{x}$$

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

17 Likelihood Method of Simple Linear Regression under Normal Distribution

First, there are several assumptions:

- The conditional distribution of Y is assumed to be Normal, $(Y|X = x) \sim N(\beta_1 + \beta_2 x, \sigma^2)$
- The mean of Y is a linear function of X, $E[Y_i|X_1 = x_i] = \beta_1 + \beta_2 x_i$
- The variance σ^2 is a constant, $var[Y_i|X_i = x_i] = \sigma^2$
- Let $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ be observed data of X, Y respectively
- Assume also that $y_i's$ are independent

then the likelihood function is

$$L(\beta_1, \beta_2, \sigma^2|data) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2}$$

which is maximized at

$$\hat{\beta}_1 = b_1 = \overline{y} - b_2\overline{x}$$

$$\hat{\beta}_2 = b_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$