Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

Chapter 9. Two-Sample Inferences

§ 9.1 Introduction

§ 9.2 Testing $H_0 : \mu_X = \mu_Y$

§ 9.3 Testing $H_0 : \sigma_X^2 = \sigma_Y^2$

§ 9.4 Binomial Data: Testing $H_0 : p_X = p_Y$

§ 9.5 Confidence Intervals for the Two-Sample Problem

Chapter 9. Two-Sample Inferences

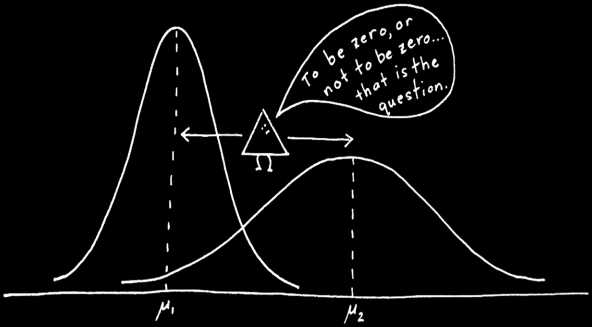
§ 9.1 Introduction

§ 9.2 Testing $H_0 : \mu_X = \mu_Y$

§ 9.3 Testing $H_0 : \sigma_X^2 = \sigma_Y^2$

§ 9.4 Binomial Data: Testing $H_0 : p_X = p_Y$

§ 9.5 Confidence Intervals for the Two-Sample Problem



Multilevel designs:

1. Two methods applied to two independent sets of similar subjects.

   E.g., comparing two products.

2. Same method applied to two different kinds of subjects.

   E.g., comparing bones of European kids and American kids.

Test for normal parameters (two sample test)

1. Let $X_1, \cdots, X_n$ be a random sample of size n from $N(\mu_X, \sigma_X^2)$.
2. Let $Y_1, \cdots, Y_m$ be a random sample of size m from $N(\mu_Y, \sigma_Y^2)$.

Prob. 1 Find a test statistic Λ in order to test $H_0 : \mu_X = \mu_Y$ v.s. $H_1 : \mu_X \neq \mu_Y$.

1-1 When σ_X^2 and σ_Y^2 are known

1-2 When $\sigma_X^2 = \sigma_Y^2$ is unknown

1-3 When $\sigma_X^2 \neq \sigma_Y^2$, both are unknown

Prob. 2 Find a test statistic Λ in order to test $H_0 : \sigma_X^2 = \sigma_Y^2$ v.s. $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

Prob. 1-1 Find a test statistic for $H_0 : \mu_X = \mu_Y$ v.s. $H_1 : \mu_X \neq \mu_Y$,

with σ_X^2 and σ_Y^2 known.

Sol.

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Test statistics: $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$.

Critical region $|z| \geq z_{\alpha/2}$. $\square$

Prob. 1-2 Find a test statistic for $H_0 : \mu_X = \mu_Y$ v.s. $H_1 : \mu_X \neq \mu_Y$,

with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ but unknown.

Sol. Composite-vs-composite test with:

$$\omega = \left\{(\mu_X, \mu_Y, \sigma^2) : \mu_X = \mu_Y \in \mathbb{R}, \quad \sigma^2 > 0\right\}$$

$$\Omega = \left\{(\mu_X, \mu_Y, \sigma^2) : \mu_X \in \mathbb{R},\ \mu_Y \in \mathbb{R},\ \sigma^2 > 0\right\}$$

The likelihood function

$$L(\omega) = \prod_{i=1}^{n} f_X(x_i) \prod_{j=1}^{m} f_Y(y_j)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{m+n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \mu_X)^2 + \sum_{j=1}^{m}(y_i - \mu_Y)^2\right]\right)$$

Under ω, the MLE $\omega_e = (\mu_{\omega_e}, \mu_{\omega_e}, \sigma^2_{\omega_e})$ is

$$\mu_{\omega_e} = \frac{\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j}{n+m}$$

$$\sigma^2_{\omega_e} = \frac{\sum_{i=1}^{n}(x_i - \mu_{\omega_e})^2 + \sum_{j=1}^{m}(y_j - \mu_{\omega_e})^2}{n+m}$$

Hence,

$$L(\omega_e) = \left(\frac{e^{-1}}{2\pi\sigma^2_{\omega_e}}\right)^{\frac{n+m}{2}}$$

Under Ω, the MLE $\omega_e = (\mu_{X_e}, \mu_{Y_e}, \sigma^2_{\Omega_e})$ is

$$\mu_{X_e} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \mu_{Y_e} = \frac{1}{m}\sum_{j=1}^{m} y_j$$

$$\sigma^2_{\Omega_e} = \frac{\sum_{i=1}^{n}(x_i - \mu_{X_e})^2 + \sum_{j=1}^{m}(y_j - \mu_{Y_e})^2}{n+m}$$

Hence,

$$L(\Omega_e) = \left(\frac{e^{-1}}{2\pi\sigma^2_{\Omega_e}}\right)^{\frac{n+m}{2}}$$

$$\lambda = \frac{L(\omega_e)}{L(\Omega_e)} = \left(\frac{\sigma^2_{\Omega_e}}{\sigma^2_{\omega_e}}\right)^{\frac{m+n}{2}}$$

$$\lambda^{\frac{2}{n+m}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2}{\sum_{i=1}^{n}\left(x_i - \frac{n\bar{x}+m\bar{y}}{m+n}\right)^2 + \sum_{j=1}^{n}\left(y_j - \frac{n\bar{x}+m\bar{y}}{m+n}\right)^2}$$

$$\sum_{i=1}^{n}\left(x_i - \frac{n\bar{x} + m\bar{y}}{m+n}\right)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{m^2 n}{(m+n)^2}(\bar{x} - \bar{y})^2$$

$$\sum_{j=1}^{m}\left(y_j - \frac{n\bar{x} + m\bar{y}}{m+n}\right)^2 = \sum_{j=1}^{m}(y_i - \bar{y})^2 + \frac{mn^2}{(m+n)^2}(\bar{x} - \bar{y})^2$$

$$\Downarrow$$

$$\sum_{i=1}^{n}\left(x_i - \frac{n\bar{x} + m\bar{y}}{m+n}\right)^2 + \sum_{j=1}^{n}\left(y_j - \frac{n\bar{x} + m\bar{y}}{m+n}\right)^2$$

$$||$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{m}(y_i - \bar{y})^2 + \frac{mn}{m+n}(\bar{x} - \bar{y})^2$$

$$\lambda^{\frac{2}{m+n}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_i-\bar{y})^2 + \frac{mn}{m+n}(\bar{x}-\bar{y})^2}$$

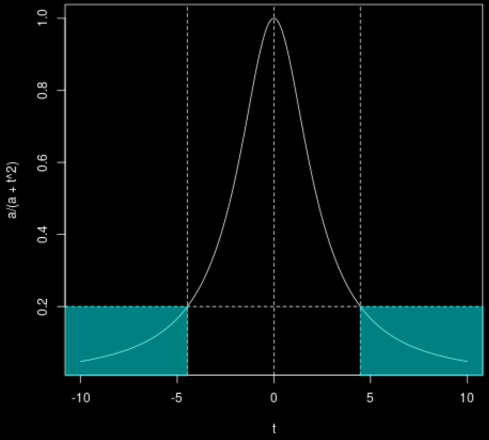
$$= \frac{1}{1 + \frac{(\bar{x}-\bar{y})^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_i-\bar{y})^2\right]\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

$$= \frac{n+m-2}{n+m-2 + \frac{(\bar{x}-\bar{y})^2}{\frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_i-\bar{y})^2\right]\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

$$= \frac{n+m-2}{n+m-2 + \frac{(\bar{x}-\bar{y})^2}{s_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}} = \frac{n+m-2}{n+m-2+t^2}.$$

$$\boxed{t := \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}}$$

$$t \mapsto \frac{a}{a + t^2}$$



One can use the following statistic

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where S_p^2 is called the *pooled sample variance*

$$S_p^2 = \frac{1}{n+m-2} \left[\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 + \sum_{i=1}^{m} \left(Y_j - \overline{Y} \right)^2 \right]$$

$$= \frac{1}{n+m-2} \left[(n-1) S_X^2 + (m-1) S_Y^2 \right]$$

Three observations:

1. $\mathbb{E}[\overline{X} - \overline{Y}] = 0$ and

$$\text{Var}(\overline{X} - \overline{Y}) = \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} = \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)$$

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

2. $\frac{n+m-2}{\sigma^2} S_p^2 = \sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \sum_{j=1}^m \left(\frac{Y_j - \overline{Y}}{\sigma}\right)^2 \sim$ Chi square$(n+m-2)$

3. $\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \perp \frac{n+m-2}{\sigma^2} S_p^2$

$$\Longrightarrow T = \frac{\dfrac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\dfrac{n+m-2}{\sigma^2} S_p^2 \times \dfrac{1}{n+m-2}}} = \frac{\overline{X}-\overline{Y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim \text{t distr.}(n+m-2)$$

Finally,

Test statistics: $t = \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$

Critical region: $|t| \geq t_{\alpha/2,n+m-2}$. $\square$

Prob. 1-3 Find a test statistic for $H_0 : \mu_X = \mu_Y$ v.s. $H_1 : \mu_X \neq \mu_Y$,

with $\sigma_X^2 \neq \sigma_Y^2$, both unknown.

Remark: 1. Known as the *Behrens-Fisher problem*.

2. No exact solutions!

3. We will derive a widely used approximation by

Bernard Lewis Welch (1911–1989)

Sol.

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \Bigg/ \frac{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

$$U := \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

$$\frac{V}{\nu} := \frac{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}$$

!! **Assumption/Approximation:**

Assume that V follows Chi Square(ν) and assume that $V \perp U$.

$\Longrightarrow$ Then, $W \sim$ Student's t-distribution of freedom ν.

? It remains to estimate ν: Suppose we have

$$\nu = \frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^4}{n^2(n-1)} + \frac{\sigma_Y^4}{m^2(m-1)}} = \frac{\left(\theta + \frac{n}{m}\right)^2}{\frac{1}{n-1}\theta^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2}, \quad \theta = \frac{\sigma_X^2}{\sigma_Y^2}.$$

!! Still need to know $\theta = \sigma_X^2/\sigma_Y^2$... Another approximation $\hat{\theta} = S_X^2/S_Y^2$, i.e.,

$$\nu \approx \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}} = \frac{\left(\hat{\theta} + \frac{n}{m}\right)^2}{\frac{1}{n-1}\hat{\theta}^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2}, \quad \hat{\theta} = \frac{s_X^2}{s_Y^2}.$$

In summary:

$$W = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim \text{Student's t of freedom } \nu$$

$$\nu = \left[\frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}} \right] = \left[\frac{\left(\hat{\theta} + \frac{n}{m}\right)^2}{\frac{1}{n-1}\hat{\theta}^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2} \right], \quad \hat{\theta} = \frac{s_X^2}{s_Y^2}.$$

Test statistic: $t = \frac{\bar{x} - \bar{y} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$

Critical region: $|t| \geq t_{\alpha/2,\nu}$. □

Remark If $\nu \geq 100$, replace the t-score, e.g., $t_{\alpha/2,\nu}$ by the z-score, e.g., $z_{\alpha/2}$.

Thm The moment estimate for ν

$$\nu = \frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^4}{n^2(n-1)} + \frac{\sigma_Y^4}{m^2(m-1)} + \frac{\sigma_X^2\sigma_Y^2}{mn}}$$

$$\approx \frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^4}{n^2(n-1)} + \frac{\sigma_Y^4}{m^2(m-1)}} = \frac{\left(\theta + \frac{n}{m}\right)^2}{\frac{1}{n-1}\theta^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2}, \quad \theta = \frac{\sigma_X^2}{\sigma_Y^2}.$$

Proof.

$$\frac{V}{\nu}\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right) = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

$$(n-1)S_X^2/\sigma_X^2 \sim \text{Chi Sqr}(n-1) \Longrightarrow \mathbb{E}(S_X^2) = \sigma_X^2. \text{ Similarly, } \mathbb{E}(S_Y^2) = \sigma_Y^2.$$

First moment gives identity. Need to consider second moment.

Second moments for Chi sqr(r) is $2r$. Hence, $\mathbb{E}(S_X^4) = \frac{\sigma_X^4}{n-1}$.

$$\frac{2\nu}{\nu^2}\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2 = 2\frac{\sigma_X^4}{n^2(n-1)} + 2\frac{\sigma_Y^4}{m^2(m-1)} + 2\frac{\sigma_X^2\sigma_Y^2}{mn}$$

... $\square$

Remark Welch (1938) approximation is more involved, which actually assumes that V follows the *Type III Pearson distribution*.

`https://en.wikipedia.org/wiki/Behrens-Fisher_problem`

Prob. 2 Find a test statistic Λ in order to test $H_0 : \sigma_X^2 = \sigma_Y^2$ v.s. $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

Sol.

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \text{F-disribution } (n-1, m-1)$$

Test statistic: $f = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{s_X^2}{s_Y^2}$

Critical regions: $f \leq F_{\alpha/2,n-1,m-1}$ or $f \geq F_{1-\alpha/2,n-1,m-1}$. $\square$