

STAB57: An Introduction to Statistics

Shahriar Shams

Week 8 (Comparing two populations)



Winter 2023

Recap of Week 7

- Idea of test of hypothesis and types of hypothesis
- Two approaches:
 - Critical value approach
 - p-value approach
- Type-1, Type-2 error and Power of a test.
- Test of hypothesis using Confidence Interval

Learning goals for this week

- Comparing two independent Normal Populations:
 - equality of two variances
 - equality of two means (variances known)
 - equality of two means (variances unknown)
- Comparing two population means (paired data)

These are selected topics [John A. Rice: Chap 11](#)

Section 1

Comparing two independent Normal Populations

Subsection 1

Equality of two variances

Testing equality of two variances

- Suppose we have two independent Normal samples

$$X_1, X_2, \dots, X_n \sim N(\mu_x, \sigma_x^2)$$

and

$$Y_1, Y_2, \dots, Y_m \sim N(\mu_y, \sigma_y^2)$$

- We want to test $H_0 : \sigma_x^2 = \sigma_y^2$ vs. $H_1 : \sigma_x^2 \neq \sigma_y^2$
- We can write

$$\frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi_{(n-1)}^2$$

and

$$\frac{(m-1)S_y^2}{\sigma_y^2} \sim \chi_{(m-1)}^2$$

Equality of two variances

Recall: If $A \sim \chi_p^2$ and $B \sim \chi_q^2$ then $\frac{A/p}{B/q} \sim F_{p,q}$

- Therefore we can write,

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n-1,m-1}$$

- Under H_0 we have, $\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1}$
- Rejection region: $(-\infty, F_{\alpha/2,(n-1,m-1)}) \cup (F_{1-\alpha/2,(n-1,m-1)}, \infty)$
- We reject H_0 if $\frac{s_x^2}{s_y^2}$ falls in the rejection region.

Related Questions

- Construct a γ -CI for σ_x^2/σ_y^2
- Construct a γ -CI for σ_y^2/σ_x^2
- At α level of significance, test $H_0 : \sigma_x^2 = \sigma_y^2$ vs. $H_1 : \sigma_x^2 > \sigma_y^2$
- At α level of significance, test $H_0 : \sigma_x^2 = \sigma_y^2$ vs. $H_1 : \sigma_x^2 < \sigma_y^2$

Subsection 2

Equality of two means (variances known)

Testing $\mu_x = \mu_y$ with known σ_x and σ_y

- Testing $H_0 : \mu_x = \mu_y$ is same as testing $H_0 : \mu_x - \mu_y = 0$
- We know \bar{X} is a sensible estimator of μ_x and \bar{Y} is sensible estimator of μ_y
- Then $\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$ and $\bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{m})$
- Therefore,

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right) \quad (1)$$

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

$$\text{under } H_0, \quad \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0, 1) \quad (3)$$

Testing $\mu_x = \mu_y$ with known σ_x and σ_y (cont...)

- **Approach-1:** Using equation (2) from the previous slide, construct a $(1 - \alpha)$ level confidence interval

$$(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

Since testing $H_0 : \mu_x - \mu_y = 0$, check if 0 is inside the interval or not.

- **Approach-2:** Using equation (3) of slide 21, we can construct the α level rejection region and calculate the test statistic or p-value etc...

Testing $\mu_x = \mu_y$ with known $\sigma_x = \sigma_y = \sigma$

- $\bar{X} \sim N(\mu_x, \frac{\sigma^2}{n})$ and $\bar{Y} \sim N(\mu_y, \frac{\sigma^2}{m})$
- Therefore,

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0, 1)$$

$$\text{under } H_0, \quad \frac{(\bar{X} - \bar{Y})}{\sigma\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0, 1)$$

The rest is the same...

Subsection 3

equality of two means (variances unknown)

Testing $\mu_x = \mu_y$ with unknown $\sigma_x = \sigma_y = \sigma$

- $\bar{X} \sim N(\mu_x, \frac{\sigma^2}{n})$ and $\bar{Y} \sim N(\mu_y, \frac{\sigma^2}{m})$
- Therefore,

$$\frac{(\bar{X} - \bar{Y})}{\sigma \sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim N(0, 1) \quad (4)$$

- Using the χ^2 properties

$$\begin{aligned} \frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2} &\sim \chi_{(n-1)}^2 + \chi_{(m-1)}^2 = \chi_{(n+m-2)}^2 \\ \implies \frac{1}{\sigma^2} [(n-1)S_x^2 + (m-1)S_y^2] &\sim \chi_{(n+m-2)}^2 \end{aligned}$$

Testing $\mu_x = \mu_y$ with **unknown** $\sigma_x = \sigma_y = \sigma$ (cont...)

Using the definition of a t-distribution

$$\begin{aligned} & \frac{\frac{(\bar{X} - \bar{Y})}{\sigma \sqrt{(\frac{1}{n} + \frac{1}{m})}}}{\sqrt{\frac{\frac{1}{\sigma^2} [(n-1)S_x^2 + (m-1)S_y^2]}{n+m-2}}} \sim t_{n+m-2} \\ \Rightarrow & \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}} \sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim t_{(n+m-2)} \\ \Rightarrow & \frac{(\bar{X} - \bar{Y})}{S_p \sqrt{(\frac{1}{n} + \frac{1}{m})}} \sim t_{(n+m-2)} \end{aligned}$$

where $S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$ is called the pooled sample variance.

Numerical example

Example A, Rice page-423

Example B, Rice page-425

Testing $\mu_x = \mu_y$ with unknown variances & $\sigma_x \neq \sigma_y$

- Suppose the two population variances are unknown.
- We test $H_0 : \sigma_x^2 = \sigma_y^2$ and H_0 is rejected.
- We can't use the test statistic that we have derived in slide 14 and 15.
- We use $\frac{S_x^2}{n} + \frac{S_y^2}{m}$ as an estimator of $\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$
- We use the test statistic,

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

- The distribution of the test statistic provided in the previous page is **approximated** using a t-distribution with

$$df = \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}$$

- In some textbook it is advised to round the df to nearest integer.
- The procedure for testing hypothesis and constructing CI remains exactly the same as before.
- We will learn doing this test using Likelihood Ratio Test(LRT) next week.

Section 2

Comparing two population means (paired data)

- In previous sections we assumed the two samples are independent.
- In many practical setting the samples are paired.
- For example, we want to test whether a new drink changes blood sugar level or not.
- We would measure the blood sugar level of the participants before drinking and measure again (say) 30 min after drinking.
- These two set of measurements are coming from same set of individuals.
- Hence the observations are not independent any more rather dependent.

Paired data (cont...)

- Let X represent the measurement before the drink and
- Y represent the measurement after the drink
- we want to test $H_0 : \mu_X - \mu_y = 0$ vs $H_1 : \mu_x - \mu_y \neq 0$
- We can still use $\bar{X} - \bar{Y}$ as we did in previous slides but $var(\bar{X} - \bar{Y})$ will contain a covariance term now.
- To simplify the problem, let's define $D = X - Y \implies \mu_d = \mu_x - \mu_y$
- testing $H_0 : \mu_X - \mu_y = 0$ is same as testing $H_0 : \mu_d = 0$
- We can use

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

- Now the problem is like one sample t-test learned in Week-7.

Numerical examples

- Example A: Rice-p446.
- Another example: Let X Y represent the before and after measurements of 10 participants. Check whether the drink changes the blood sugar level or not.

x	10.19	7.92	6.67	12.22	8.21	8.26	13.06	8.20	9.83	5.94
y	7.00	7.53	6.45	1.31	5.42	2.81	6.60	0.55	3.13	5.00
d	3.19	0.39	0.22	10.91	2.79	5.45	6.46	7.65	6.70	0.94

- 1 $\bar{d} = 4.47$ and $s_d = 3.545106$
- 2 Test-statistic, $T = \frac{4.47}{3.545106/\sqrt{10}} = 3.987294$
- 3 $t_{0.975, df=9} = 2.262$
- 4 Rejection region: $(-\infty, -2.262) \cup (2.262, \infty)$
- 5 Reject $H_0 \implies$ The drink changes blood sugar level.

Homework(Non-credit)

Evans and Rosenthal

Example 10.4.4 (p-585)

John A. Rice

Exercise 11: 1, 15, 16, 32, 21(a, d), 33, 35, 36, 39(b)