STA255 Week-10 (day-2)

Shahriar Shams

17/03/2020

Review of week-10 (day-1)

- Testing population proportion, $p = p_0$ (Chapter 9.3)
- Idea of p-value. (Chapter 9.4)

Learning goals

- Z test and confidence interval for difference between two population means
- Two sample t-test and confidence interval
- Analysis of paired data

Z test and confidence interval for difference between two population means

- So far what we have learned only talks about one population.
  - For example, our population is all UofT students and we wanted to know the average height of them.
- Assuming height of a single student follows Normal distribution with mean μ and variance σ^2, we learned
  - how to calculate a point and interval estimate of μ
  - how to test a hypothesis like $H_0 : \mu = \mu_0$

- In this lecture we will learn how to deal with two population both both independently following Normal distribution.

- Suppose we have two **independent** Normal samples

$$X_1, X_2, ...X_m \sim N(\mu_x, \sigma_x^2)$$

and

$$Y_1, Y_2, ..., Y_n \sim N(\mu_y, \sigma_y^2)$$

Constructing confidence interval for $\mu_x - \mu_y$

- Since $X_1, X_2, ...X_m \sim N(\mu_x, \sigma_x^2)$ we can write

$$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{m})$$

- Since $Y_1, Y_2, ..., Y_n \sim N(\mu_y, \sigma_y^2)$ we can write

$$\bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n})$$

- If we consider the random variable $\bar{X} - \bar{Y}$ which is linear combination of two independent Normal distributions, we can write

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

- Standardizing this new variable we can write

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

- If we want to construct a confidence interval for $(\mu_x - \mu_y)$ we can do it using the same idea that we used for single population/single parameter case.
- Using the same idea used in week-8, we can write, $100 * (1 - \alpha)\%$- CI for $(\mu_x - \mu_y)$

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

Testing $H_0 : \mu_x - \mu_y = \Delta_0$

- Here Δ_0 represents a numeric value.
- Typically we test $H_0 : \mu_x - \mu_y = 0$, which is same testing whether the two population means are equal or not.

- From previous section, we already know,

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

- Under null hypothesis, we can write

$$Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \sim N(0,1)$$

- If the sample observations are given, the value of $\Delta_0, \sigma_x^2 \ and \ \sigma_y^2$ are given we can calculate the value of the test statistic and can either check whether it falls in the rejection region or not or can calculate the associated p-value.

Example: 10.1 on page 487

Two sample t-test and confidence interval

- In the previous section we assumed σ_x^2 and σ_y^2 are known.
- When they are unknown, we don't use the standard normal distribution rather we use a t-distribution.
- We can write
$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}} \to t_{(df=v)}$$
- Here, S_x^2 and S_y^2 are the two sample variances.
- The t-distribution that we use here is an approximation. Hence, we didn't use the sign $\sim$ rather used the sign $\to$.
- The degrees of freedom of this t-distribution is calculated using a rather complex formula. Deriving this formula is out of the scope of this course. So we will just use it as

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

- v is then rounded down to the nearest integer.

- $100 * (1 - \alpha)\%$- CI for $(\mu_x - \mu_y)$ then can be written as

$$(\bar{X} - \bar{Y}) \pm t_{1-\frac{\alpha}{2},v}\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}$$

- For testing, $H_0 : \mu_x - \mu_y = \Delta_0$ we can use the test statistic,

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}} \to t_{(df=v)}$$

Example 10.6 on page 501

Example 10.7 on page 502

Pooled t-procedure

- In the previous section we assumed σ_x^2 and σ_y^2 are unknown.
- If we assume the two population curves have the same spread out calculation becomes a bit simpler.
- Under the assumption $\sigma_x^2 = \sigma_y^2$, there is a simpler test statistic which follows ($\sim$) a t-distribution.
- We can write

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

- Here, S_p^2 is called the pooled sample variance which can be seen as a weighted average of the two sample variances calculated as

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

- $100 * (1-\alpha)\%$- CI for $(\mu_x - \mu_y)$ then can be written as

$$(\bar{X} - \bar{Y}) \pm t_{1-\frac{\alpha}{2}, m+n-2}\sqrt{S_p^2(\frac{1}{m} + \frac{1}{n})}$$

- For testing, $H_0 : \mu_x - \mu_y = \Delta_0$ we can use the test statistic,

$$T = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{S_p^2(\frac{1}{m} + \frac{1}{n})}} \sim t_{df=m+n-2}$$

Analysis of paired data

- In previous sections we assumed the two samples are independent.
- In many practical settings 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.

- 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 sections 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
$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \sim t_{(n-1)}$$
- Now the problem is like one sample t-test that we learned last week.

Numeric 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.

Chapter 10.4 not needed

I recommend reading chapter 10.4 for future courses. I believe you will be able to follow.

Chapter 10.5 and 10.6 are not needed.

Homework

Chapter 10.1

1(a,b), 5(a,b), 6(a), 10(a)

Chapter 10.2

20, 24, 26, 29(b,c,d), 31, 33

Chapter 10.3

39, 41, 42(b), 43