STATS 412

Twentieth Class Note

In Son Zeng

02 December, 2018

My Office Hour:

The office hours this week are on **16:30 - 18:00 Tuesday** and **13:30 - 15:00 Friday**, at **USB 2165**. You may check the campus map to get to my office. I am prepared for your questions, so please feel free to come to my office hours.

Exam 3 Updates:

I will be updating the revision for the final exam day by day, so check closely to the update for the final exam.

• The revision contents are various: concepts, test convention, sample questions and solutions. Hope you find the materials helpful for performing better in the final exam.

• If our class can improve the handwriting for the rest of the homeworks, I can have more time to write the revision day by day to improve the general quality. Also, when I have more time, I can be more efficiently respond questions on Piazza and provide better designed questions and solutions (with video).

Homework Grading Policy:

Please include the final answer for each homework question. If the final answer is not included, you will risk 0.5 points for each missing part.

Homework 11 Reminder:

Population mean difference hypothesis testing or confidence interval: Before reading this, please tell yourself do not get confused with the one-sample and two-sample hypothesis testing and confidence interval, which is stated on class notes 15, 16, 17, 18 and 19.

Now, let us go through the **Hypothesis testing for population mean difference**

• Our parameter of interest is μ_D, which is the **population mean difference between ..... (a paired sample)**.

• Hypothesis: In most cases we construct the hypothesis as the follows: $H_0 : \mu_D = \mu_0$ vs $H_1 : \mu_D \neq \mu_0$ (two tails), you may check your notes for the one-tail situation.

• If the paired samples are given random, you are good to go! If not, we can either explain by your own reasoning why you think the two paired samples are impacting each other (not independent) or not impacting each other (independent), or draw a scatterplot (here most of the time the samples are given). This is important because we need randomness of the paired samples to perform either t-test or z-test.

• If you use R, then you can look at the summary() and sd() to obtain the sample mean, quantiles and sample standard deviation for the population mean difference. For example, in base R we can calculate:

```
truckhot = c(4.56, 4.46, 6.49, 5.37, 6.25, 5.90, 4.12, 3.85, 4.15, 4.69)
truckchold = c(4.26, 4.08, 5.83, 4.96, 5.87, 5.32, 3.92, 3.69, 3.74, 4.19)
summary(truckhot - truckchold) # sample mean for population mean difference is 0.398
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3200  0.3950  0.3980  0.4775  0.6600
```

```
sd(truckhot - truckchold) # sample mean for population mean difference is approximately 0.156
```

```
## [1] 0.1558347
```

```
# You may say all observations are within 2 standard deviations from the mean, there is no outlier,
# so we can rely on the robustness of t-distribution and proceed with caution.

# Or you may use IQR, (Q3 + 1.5 * IQR = 0.478 + 1.5 * 0.1575 = 0.714 > 0.66) and
# (Q1 - 1.5 * IQR = 0.32 - 1.5 * 0.1575 = 0.084 < 0.16) to conclude that there is no outlier
```

• If the underlying distribution of the two paired samples are given normal, you are good to go! If not, we first check the sample size of the two paired samples (they should have the same sample size). If the sample sizes are large, i.e., $n \geq 30$, then we can use the CLT to say that the sampling distribution for the population mean difference is approximately normal.

• If the sample sizes are small, we cannot use CLT. We may use QQ-plot (or perform normality test) to look at the data to see if the samples seem to be normally distributed. In this case, most of the time the paired data are given. If the samples do not show terrible outliers (I know it is kind of subjective), then we rely on the robustness of the t procedures (and/or proceed with caution) against the mild violations of normality.

• Sample statistics: First we define the significance level α, which you can specify as 0.05, 0.01, 0.005 or the quantity given in the question. If we perform the t-test, the degree of freedom for population mean difference is $n-1$.

• Then, first compute the difference between two paired samples $(d_1, ..., d_n)$ and take the mean $\bar{D}$. If we use R, it is simply the sd function. If you wish to compute by hand, we compute the sample standard deviation of the differences s_D. That is,

$$\boxed{s_D = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}} \tag{1}$$

• Sample statistics: Then, we derive the t-statistics by:

$$\boxed{t = \frac{\bar{D} - \mu_0}{s_D/\sqrt{n}}} \tag{2}$$

Upon finding the t-value we refer back to the t-table and find the range of the corresponding p-value. Note that if you use the pt function in R, we should specify the t quantity and degree of freedom. The codes are as follows:

```
# p-value for t>2 with 20 degree of freedom
p_value = 1 - pt(2,20); p_value # 0.0296
```

```
## [1] 0.02963277
```

```
# p-value for t<-2.3 with 45 degree of freedom
p_value1 = pt(-2.3,45); p_value1 # 0.0131
```

```
## [1] 0.01307008
```

• Since the standard student-t distribution is symmetric (taught in chapter 4), if the hypothesis testing is one-tail, we just use that p-value; if the hypothesis testing is two-tail, we need to multiple the p-value by 2.

Formal Statement for Conclusion for two-sample hypothesis testing:

• With a p-value lower than the significant level (such as 0.05, 0.01), we reject H_0, the null hypothesis. And we say: There is sufficient evidence to suggest that the **population mean difference of A and B** is (different from/greater than/smaller than).

• With a p-value greater than the significant level (such as 0.05, 0.01), we fail to reject H_0, the null hypothesis. And we say: There is not sufficient evidence to suggest that the **population mean difference of A and B** is (different from/greater than/smaller than).

Procedure of deriving Population mean difference (Confidence interval):

• The first step is the define the parameter correctly! The parameter should be the **population mean difference in A and B**, denoted μ_D. If you mention sample mean or mean, you will be taken points off.

• To check the conditions, first, we have to check whether the two paired samples are random. If it is told, great! Just move on! If not, we can either explain by your own reasoning why you think the two paired samples are impacting each other (not independent) or not impacting each other (independent), or draw a scatterplot (here most of the time the samples are given). This is important because we need randomness of the paired samples to perform either t-test or z-test.

Then, we check

• 1. Whether the underlying population distributons of two paired samples are normal, we check if we are told that the distributions of the two populations from which the measurements are taken are normal.

• 2. If not 1, then we check whether the sample sizes are large enough $n \geq 30$ to employ the central limit theorem, which claims that the sampling distributions of the population mean difference of the measurements is approximately normal. Note that now the two samples should have same sample size because they are paired samples.

• 3. If not 1 and not 2, we may use QQ-plot (or perform normality test) to look at the data to see if the samples seem to be normally distributed. In this case, most of the time the paired data are given. If the samples do not show terrible outliers (I know it is kind of subjective), then we rely on the robustness of the t procedures (and/or proceed with caution) against the mild violations of normality.

Then you can construct confidence interval as follows:

• Here our degree of freedom is $n-1$ for the paired sample.

• Then, first compute the difference between two paired samples $(d_1, ..., d_n)$ and take the mean $\bar{D}$. Then, we compute the sample standard deviation of the differences s_D. That is,

$$\boxed{s_D = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}} \tag{3}$$

• Find the t-score or z-score by referencing the table, given the specified significance level. For two-tails (keyword: **between**,$\pm$) , we find $t_{v,\frac{\alpha}{2}}$ or $z_{\frac{\alpha}{2}}$, and the confidence interval is (most often)

$$\boxed{\bar{D} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}} \tag{4}$$

• For one-tail (keyword: **no greater than, no less than**), we find the confidence upper bound (CUB) or confidence lower bound (CLB) by finding $t_{n-1,\alpha}$ or z_α. The confidence upper bound is $\bar{D} + t_{n-1,\alpha} \cdot \frac{s_d}{\sqrt{n}}$ and the confidence lower bound is $\bar{D} - t_{n-1,\alpha} \cdot \frac{s_d}{\sqrt{n}}$

Example of Formal Statement for Conclusion:

• We are approximately (95%/99%) confident that the **population mean difference** in . . . .what question say. . . . . is (between /no greater than/no smaller than) . . . the result. . . .

Revision: Steps of constructing confidence interval for difference in population means:

• The first step is the define the parameter correctly! The parameter should be the **difference between the population/true mean of A and B**. If you mention sample mean or mean, you will be taken points off.

• To check the conditions, first, we have to check whether the two samples are random. If it is told, great! Just move on! If not, we can either explain by your own reasoning why you think the two samples are random or not, or draw a scatterplot if the samples are given (if you have great sense of calculation you can explain clearly by plain words). This is important because we need randomness of samples to perform either t-test or z-test.

Then, we check

• 1. Whether the underlying population distributons of two samples are normal, we check if we are told that the distributions of the two populations from which the measurements are taken are normal.

• 2. If not 1, then we check whether the sample sizes are large enough $n \geq 30$ to employ the central limit theorem, which claims that the sampling distributions of the sample means of the measurements are approximately normal.

• 3. If not 1 and not 2, we may use QQ-plot (or perform normality test) to look at the data to see if the samples seem to be normally distributed. (It is a good way to practice in homework)

• 4. If the data are not given or the sample size are small $n < 30$, then we rely on the robustness of the t procedures against violations of normality. In short, we should check (not PROVE) the randomness and normality.

Then you can construct confidence interval as follows:

• We compute the degree of freedom v using the same formula above for the hypothesis testing.

• If σ_X and σ_Y are given, compute $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$. Otherwise, compute s_X and s_Y, which are the square root of the sample variances. Then, compute $\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$.

• Find the t-score or z-score by referencing the table, given the specified significance level. For two-tails (keyword: **between**,$\pm$) , we find $t_{v,\frac{\alpha}{2}}$ or $z_{\frac{\alpha}{2}}$, and the confidence interval is (most often)

$$\boxed{(\bar{x} - \bar{y}) \pm t_{v,\frac{\alpha}{2}} \cdot \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \tag{5}$$

• For one-tail (keyword: **no greater than, no less than**), we find the confidence upper bound (CUB) or confidence lower bound (CLB) by finding $t_{v,\alpha}$ or z_α. The confidence upper bound is $(\bar{x} - \bar{y}) + t_{v,\alpha} \cdot \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$ and the confidence lower bound is $(\bar{x} - \bar{y}) - t_{v,\alpha} \cdot \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$

Example of Formal Statement for Conclusion for confidence interval:

• We are approximately (95%/99%) confident that the **difference between population means** of ....what question say..... is (between /no greater than/no smaller than) ...the result....

Revision: Using t-score or z-score (two samples): During the class 17 (Monday) there are concerns about using the z-score and t-score. Professor's explanation is that, after checking the randomness:

• If one of the sample size is small ($n < 30$), use t-test. Also, you could plot a normal Q-Q plot to see whether both samples are approximately normal. If yes, you are good to use either z-test or t-test. In short, using t-test guarantees correctness.

• If the population standard deviation σ_X or σ_Y (for either sample) is not known, use t-test.

• If the two samples have large sample size, are given underlying distribution normal and the population standard deviations are known, use z-test.

• If the two samples have large sample size ($n \geq 30$) and the population standard deviations are known, but with unknown underlying distributions, we still need to employ Central Limit Theorem, and use z-test with approximated probability

$$P(\bar{X} - \bar{Y} > t) \approx P(Z > \frac{t - \mu_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_y}}})$$

.

Important: Difference between parameter and statistics:

A statistic is defined as a numerical value obtained from a sample. Therefore, a statistic represents just a fraction of the population. We typically use statistics to estimate the parameter.

So what is a parameter? A parameter is a fixed numerical value, or a true value of the population; it reflects the aggregate of all population members under consideration. The difference between these two are described in the following website in detail.

Reference: https://keydifferences.com/difference-between-statistic-and-parameter.html

Clarification for the caveat: The caveat in chapter 4 says t-distribution tolerates the violation of normality due to the small sample size, so that we can perform t-distribution whenever the sample is random. However, the caveat does not indicate that the t-distribution also allows the outliers which may totally deviate the sample from the (approximately) normal distribution.

Pooling: If we are assuming the same variance for two samples, we can compute the pooled standard deviation:

$$\boxed{s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}} \tag{6}$$

In this case, the degree of freedom becomes $n_X + n_Y - 2$, which is an integer now. However, if we assume same variance (or standard deviation) between two samples when they are not in reality, our inferences can be really wrong (e.g., incorrect p-values).

Important: Population mean difference vs Difference in population means: Population mean difference is not the same as the difference in population means. That is, μ_D is not $\mu_X - \mu_Y$. The way to evaluate the population mean difference is more similar to the inference part 1. Also, paired samples/procedures have less variability than their independent sample/procedure counterparts.

Last Comment:

Please inform me to fix the typos and grammatical mistakes if they exist. It is a great practice of writing and I appreciate your help!