STATS 412

Sixteenth Class Note

In Son Zeng

14 November, 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.

Calculus Review:

• To compute the Maximum Likelihood Estimator (MLE), you may encounter the difficulty for partial differentiation. If you have studied MATH 215 or the equivalent class before, you may review the notes. If you do not know how to perform partial differentiation, you may refer to the following websites for reference:

• http://tutorial.math.lamar.edu/Classes/CalcIII/PartialDerivsIntro.aspx

• https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-article
a/introduction-to-partial-derivatives

These are great practices to prepare you with essential calculus skills and knowledge of distributions for the subsequent homework and the exam 2.

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 9 Reminder:

Steps of constructing confidence interval for population mean: Professor Miller mentioned in Piazza that for the conditions of performing hypothesis testing. Now we change a little bit to fit for the process of constructing the confidence interval.

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

Then, we check • 1. Whether the underlying population distributon is normal, we check if we are told that the distribution of the population from which the measurements are taken is normal.

• 2. If not 1, then we check whether the sample size is large enough $n \geq 30$ to employ the central limit theorem, which claims that the sampling distribution of the sample mean of the measurements is 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 population seems to be normally distributed.

• 4. If the data is not given or the sample size is 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:

• If σ is given, skip this part. Otherwise, compute σ, which is the square root of the sample variance.

• Find the t-score or z-score by referencing the table, given the specified significance level. For two-tails (keyword: **between,**$\pm$) , we find $t_{\frac{\alpha}{2}}$ or $z_{\frac{\alpha}{2}}$; 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_α or z_α

Example of Formal Statement for Conclusion for confidence interval:

• We are approximately (95%/99%) confident that the population mean of .....what question say........... is (between /no greater than/no smaller than) ...the confidence interval....

• 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 of .....what question say........... is (different from/greater than/smaller than) ...the number....

Clarification for the caveat: The caveat 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.

Using t-score or z-score: During the office hours there are concerns about using the z-score and t-score. My explanation is that, after checking the randomness:

• If the sample size is small ($n < 30$), use t-test.

• If the population standard deviation σ is not known, use t-test

• If the underlying distribution is given normal and the population standard deviation is known, use z-test.

• If the sample size is large ($n \geq 30$) and the population standard deviation is known, we employ Central Limit Theorem, and use z-test with approximated probability $P(\bar{X} > t) \approx P(Z > \frac{t-\mu}{\sigma})$.

Formal Statement for Conclusion for 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 of .....$\underline{\text{what question say}}$........... is ($\underline{\text{different from/greater than/smaller than}}$) ...$\underline{\text{the number}}$....

• 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 of .....$\underline{\text{what question say}}$........... is ($\underline{\text{different from/greater than/smaller than}}$) ...$\underline{\text{the number}}$....

• Confidence interval contains all the points that we fail to reject the null hypothesis at the desinated significance level. For example, a 99% confidence interval contains all the points that we fail to reject the null hypothesis at the significance level $\alpha = 0.01$.

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

Extra Example for hypothesis testing

1. Suppose the nutrition label of apple cider says that apple cider contains an average concentration of sugar 90g/liter, with standard deviation 10g/liter. Now we want to know whether the claim is true. So we bought 50 liters of apple ciders as sample and found that 5050g sugar are contained in total. Based on this information, perform a hypothesis testing to answer whether the claim in the nutrition label is true, at significance level $\alpha = 0.05$.
2. With the same setting, now we only know that the average concentration of sugar in apple cier is 90g/liter. Now we bought 10 liters of apple ciders instead, and found the concentrations of sugar for each liter of apple cider as follows: 95.3, 101.2, 92.3, 90.1, 96.4, 99.2, 110.3, 103.4, 91.2, 89.9. Based on this information, perform a hypothesis testing to answer whether the claim in the nutrition label is true, at significance level $\alpha = 0.05$.

Key Points during Lecture:

Critical values for t and z: With the same significance level, the t-value goes down as the sample size increases, but in comparison with the z-value corresponding to the same significance level, the t-value is always slightly larger. Particularly, when the sample size goes to infinity, then the t-value converges to the z-value. In other words, the standard normal distribution is the limit case of the t-distribution.

How are p-value related to the critical values? If you use an $\alpha = 0.05$ significance level for an alternative hypothesis of $H_1 : \mu \neq \mu_0$, we are allowing 0.025 in each tail. If the p-value is less than that, we can reject the null hypothesis (the alternative hypothesis is “differ from”, so we apply two-tails case and assign 0.025 to the right tail and 0.025 to the left tail.

t-distribution: To use the t-table, round the degree of freedom down. In the table, the α represents the area to the right hand side. A t statistics (in absolute value) greater than the corresponding value for $\alpha = 0.01$ in the table, for example, indicates that the p-value is less than 0.01.

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!