STATS 412

Fifteenth Class Note

In Son Zeng

07 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 8 Reminder:

Random: Using scatter plots is fine in checking randomness. Remember that if the samples show patterns such as 1) Monotone increasing/decreasing, or 2) U-shape (inverse-U-shape), then the sample is not random. This parts leave you great room for performing various checkings.

Steps of Hypothesis Testing: Professor Miller mentioned in Piazza that for the conditions, we have to check first, whether the sample is random. If it is told, great! 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).

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.

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 $P(\bar{X} > t) \approx P(Z > \frac{t-\mu}{\sigma})$.

If the concepts are not clear, please raise your questions so that we know what are possible misunderstandings way before the exams.

Homework 7 Short Version Comments:

• Make sure you are applying the correct formula for the MSE: $MSE(x) = [Bias(x)]^2 + Var(x)$. Additionally, there is always a tradeoff between increasing the bias while reducing the variance, or decreasing the bias while increasing the variance. In the realm of statistics, depending on the context, we may select an estimator which minimizes the variance, or is unbiased, or minimizes the MSE.

• Bias can be either positive, 0 or negative. If the bias is positive, it means the estimator overestimates the parameter. If the bias is 0, it means the estimator is unbiased. If the bias is negative, it means the estimator underestimates the parameter.

• In question 2 and 3, a majority of submissions show no work in checking that the MLE is indeed the maximum by taking the second derivative. This is a big problem since based on calculus, when the first derivative of a parameter equals to 0, then the parameter achieves **local maximum** if the second derivative (curvature) is negative, or the parameter achieves **local minimum** if the second derivative is positive.

• For question 4 - 8, a vast majority of students mistakenly did two things. The first one is the distribution of the sample mean $\bar{X}$ and the summation of observations S_n should be approximately normal, but NOT exactly normal. As a result, the second thing, the probability should be approximately equal, but not equal to the probability given by the z-score. That is the trickiest part of CLT which worth practices. See Homework 7 comments and the solution and come to my office hours for detail.

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:

Test Reminder: There will be the standard normal cumulative distribution function table provided for reference during the test. This can save your time in finding the correct probability or z-score to answer the problems such as: (1) Central Limit Theorem related approximation problems, (2) Normal distribution problems, (3) Deriving p-values for hypothesis testing.

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.

p-value: We officially introduced the definition of p-value as follows: A p-value is the (conditional) of observing a test statistics as extreme (or weird) as or more extreme than what we observed, assuming that the null hypothesis (H_0, the hypothesis that we would like to find evidence to reject) is true.

Report p-values It is always great to report the p-values because everyone may define a small p-value differently. Usually, people choose either $\alpha = 0.05$ or $\alpha = 0.01$. This "selected" level is called **significance level**.

Formal Statement for Conclusion:

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

Derive Confidence Interval: We typically find $z_{\frac{\alpha}{2}}$ such that we have $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, then set $z_{\frac{\alpha}{2}} = \frac{c}{\sigma/\sqrt{n}}$, so that we have derived the confidence interval for the population mean μ as $[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$. The number c is called the **margin of error**. The confidence interval gives us a plausible coverage of the true population parameter, and the confidence interval becomes more accurate when sample size increases.

How to calculate the quantile or the probability in t-distribution: Student's t distribution introduces the concept **degree of freedom** so that the R code for the t distribution involves assigning the degree of freedom. The R-codes to compute the probablity of $P(t < a)$ with sample size n is given by the format pt(a, n). The R-codes to compute the $(100 \times a)$ th quantile with sample size n is given by the format qt(a, n).

```
pt(2,200) # The probability of P(t<2) if the sample size is 200.
```

```
## [1] 0.9765734
```

```
qt(0.95,2000) # The 95-th quantile of Student's t distribution when sample size is 2000
```

```
## [1] 1.645616
```

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!