STA261: Probability and Statistics II

Assignment

Winter 2019

Instructions related to grading: You do not need to hand in this assignment. There will be a quiz (7-8 multiple choice questions) based on this assignment. The quiz will be open on Quercus on Mar 28 and remain open until Mar 29. The quiz will be worth 2% of your final grade.

You will get the other 1% by filling out a survey on Quercus (named "STA261 survey") which relates to STA130 and how taking that has helped you in learning different materials of this current course. Read the questions carefully and try to answer using the best answer that suits your condition. You will not get any point for partially filling out the survey. There are 5 questions. You will only get that 1% for filling out the entire survey.

Instructions on completing the assignment The numerical calculations involved in this assignment are simple and you are already familiar with them (hopefully). The goal of this work is to help you "see" some of the theorems and concepts we have learned or used in this course using empirical data. Calculations are mostly repetitive in nature! I suggest using R (or any other programming language that you are comfortable with). All the R codes that you might need were already given to you in the .R files used during lectures. If you don't want to use R, my advice is at least use Excel. If you need any help with R or Excel feel free to come to my office hour.

Q1. Suppose you have a population of size 5 [i.e. N=5]. You measure some quantity (X) and the corresponding numbers are:

$$21, 22, 23, 24, 25$$

a) Calculate the population mean (μ)

b) Calculate the population variance (σ^2) using the formula $\sigma^2 = \frac{\sum_{j=1}^{N}(X_j - \mu)^2}{N}$

Q2. Imagine you are taking samples (of size n = 3) from this population with replacement. Recall: "sampling WITH replacement" ensures independence.

a) Write down **every possible** way that you could have a sample of size 3 **with replacement** from this population. (hint: there will 5*5*5 = 125 possible combinations)

Help: if you are struggling with figuring out the combinations try this code in R:

```
expand.grid( c(21:25), c(21:25), c(21:25) )
```

b) For each of these samples of size 3, calculate the sample mean and record it (either as a new object in R or as a new column in excel). Lets call this new column "X_bar". So you should have 125 values in this column.

Q3. You should have noticed that the values in the "X_bar" column are repetitive. For example, 21.3333333 will show up 3 times.

a) Construct a frequency table based on the column "X_bar". [i.e. write down which values showed up how many times]. Now using the frequencies (also known as counts) calculate proportion of each of those repeated values. [For example: proportion of 21.3333333 will be 3/125]

b) Plot these proportions against the values and connect the points using a non-linear line. (it will look like a density plot). Does the shape of this plot look like any known distribution?

c) Using the table of proportions or otherwise, calculate the mean of these 125 numbers and compare it to your answer of 1(a).

d) Using the table of proportions or otherwise, calculate the variance of these 125 numbers. Use the population variance formula (i.e. divide by 125 not 124). What is the relationship of this answer to your answer of 1(b)?

e) Which theorem did you demonstrate empirically in part b, c and d?

Q4. For each of these sample of size 3, calculate the sample variance using the following two formulas

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Assume the population variance, $\sigma^2 = 2$.

a. By calculating (numerically using the 125 different values) $Bias[S^2]$ and $Bias[\hat{\sigma}^2]$ check the unbiasedness of these two estimators.

b. By calculating all three components separately check the following identity

$$MSE[\hat{\sigma}^2] = var[\hat{\sigma}^2] + (Bias[\hat{\sigma}^2])^2$$

Q5. Even though we need sample size n to be large to apply central limit theorem, but let's apply it anyway. Suppose you know that the population variance, $\sigma^2 = 2$.

a. For each of these 125 cases, calculate a 95% confidence interval and finally calculate the proportion of the intervals that includes $\mu = 23$.

b. Suppose someone observes only one of these 125 combinations (23,24,25). If that person is testing the null hypothesis $H_0 : \mu = 23$, based on this observed sample calculate the p-value that the person will get using central limit theorem.

c. Calculate the p-value numerically using the 125 $\bar{X}$ values that you calculated in part 2(b) (do not use CLT here).