MAT2017: Assignment 2 – 50 points total

Due: By midnight on Tuesday, May 23, 2017

The 2nd project for MAT 4017 this semester is to conduct an observational study on "Douglas wooden beam data". For this project, you work in a team of two people, with each team submitting only one *Report*.

1 Overview

A lumber company produces wooden beams in 360 different locations in Canada (210 in Eastern Canada and 150 locations in Western Canada). There are six properties related to the quality of wooden beams and these properties are specified in the catalog of the company. One of the important properties is "bending strength" and $50N/mm^2$ or higher bending strength indicates a good quality of wooden beams.

Last year the company received a customer rating of 2.3 out of 5 for the quality of wooden beams. In fact, more complaints were about wooden beams produced in Western Canada. The company recently released a new sales brochure that includes the following information of the quality improvement in their products:

- The average bending strength of the wooden beams is at least $50N/mm^2$.
- There is not difference between the average bending strength of wooden beams produced in 150 locations in Western Canada and the average bending strength of wooden beams produced in 210 locations in Eastern Canada.

Despite of the company's announcement, there are still many customer reviews indicating the bending strength of the wooden beams produced by the company is below $50N/mm^2$, and these reviews are mainly about the wooden beams from Western Canada.

You are wondering about the bending strength of their wooden beams, either it is below $50N/mm^2$ (if the customer reviews are plausible) or it is greater than equals to $50N/mm^2$ (if the company's report is believable). You are also wondering if there is a difference between wooden beams produced in Western Canada and wooden beams produced in Eastern Canada. You are to perform a observational study about the average value (mean) and the variance of the variable, "bending strength", on a set of sampled data and make a conclusion.

2 Questions Studied (Guidelines)

The population: Wooden beams produced a Canadian company

The variable: Bending strength of wooden beams

2.1 The Questions

When formulating the question you will study, form a belief about the average of your variable(s) and variance(s) and then state the question formally.

If you are more likely to believe the company's report, your questions are as follow:

1. The average bending strength of the wooden beams produced by the company is below $50N/mm^2$.

2. There is no difference between the average bending strength of wooden beams produced in Eastern Canada and that of wooden beams produced in Western Canada.

3. There is no difference between the variance of bending strength of wooden beams produced in Estern Canada and that of wooden beams produced in Western Canada.

If you are more likely to believe the customer reviews, your questions are as follows:

1. The average bending strength of the wooden beams produced by the company is greater than or equals to $50N/mm^2$

2. The average bending strength of wooden beams produced in Eastern Canada is lower than that of wooden beams produced in Western Canada.

3. The variance of bending strength of wooden beams produced in Eastern Canada is greater than that of wooden beams produced in Western Canada.

2.2 Sample Datasets

Four sample datasets of different sample sizes were drawn from the population using the simple random sampling method:

- douglas_data_50.csv - a sample of size 50 from the population of wooden beams produced in 360 locations, in Eastern and Western Canada. **Z test Normal distribution 따르기 때문에**

- douglas_data_20.csv - a sample of size 20 from the population of wooden beams produced in 360 locations, in Eastern and Western Canada. **T test**

- douglas_eastern_data_42.csv - a sample of size 42 from the population of wooden beams produced in 210 locations in Eastern Canada.

- douglas_western_data_30.csv - a sample of size 30 (rom the population of wooden beams produced in 150 locations in Western Canada.

2.3 Statistical Tests

Assuming that the population is normally distributed and you have no way of knowing the population standard deviation, you will perform a hypothesis test on the given sample dataset: state null hypothesis and alternative hypothesis, and compute the sample mean and sample standard deviation, calculate the p value of your hypothesis test, compare the p value to the significance level *alpha* and make a decision of rejection on the null hypothesis. Finally, you will report your findings.

- Q1 (the average of bending strength of wooden beams)
  - z-test on large sample data (douglas_data_50.csv)
  - t-test on small sample data (douglas_data_20.csv)
- Q2 (the difference between the averages of bending strength of Eastern wooden beams and Western wooden beams)
  - z-test on sample datasets from two populations (douglas_eastern_data_42.csv and douglas_weatern_data_30.csv)
- Q3 (the difference between the variances of bending strength of Eastern wooden beams and Western wooden beams)
  - F-test on sample datasets from two populations (douglas_eastern_data_42.csv and douglas_weatern_data_30.csv)

A statistical test consists of the following steps:

1. **The significant level:** Statement of the significance level α (you can either choose $\alpha = 0.05$ or $\alpha = 0.01$)
2. **The null and alternate hypotheses:** Formal statement of the Null hypothesis and Alternate hypotheses of your study. Basically, the alternate hypothesis is the claim we are seeking evidence for and the null hypothesis is the one we seek evidence against. You may think that these are opposites, but they are sometimes not. For example, we would have $H_0 : \mu = 10$ and $H_a : \mu > 10$.
3. **Performing the hypothesis test:** Calculate the sample mean and sample standard deviation. (You do not need to show me the calculations – just report the output from the software you are using.) Find the p-value of the hypothesis test. (Again, you don't need to show me the actual calculation – just the output of your software.) If the p value of your test is smaller than your significance level α, then you may conclude that you have statistically significant evidence at the α level against your null hypothesis in favor of your alternate hypothesis. Basically, this is like evidence in favor of your alternate hypothesis. Hence you reject the null hypothesis. If your p value is greater than α, then your study does not provide evidence at the α level against H_0 in favor of H_a. Basically, it provides no or only weak evidence that your alternate hypothesis may be true. Hence you fail to reject the null hypothesis.
4. **Reporting your findings:** Report in one or two sentences whether or not you reject the null hypothesis and what this means about your population.

3 Final Report

After you perform the theoretical computations, report the results in the final report. Begin the final report with an introduction section restating the population of interest, variable of interest and sample datasets. Present the sample data distribution using graphical tools. Then provide four hypothesis tests followed by your conclusion on the population.

- **Question studied** - 5 points
- **Sample data distribution** - 5 points
  For each of four sample datasets, you are to provide (1) graphical analysis of the sample data using a histogram and a boxplot, and (2) discussion of the shape of your data (describe any skewness or outliers present or state that the data is symmetrical with no outliers).
- **Hypotheses test results and your findings** - 40 points

4 Understanding Hypothesis Results

Here are some notes for understanding the results from hypothesis tests:

- If you study fails to reject the null hypothesis in favor of the alternate hypothesis, this does not mean that your study was a failure. It just means that there is not strong evidence that your alternate hypothesis was true. In some cases this is good news. For example if we are studying the average number of deadly earthquakes in California each decade, finding no evidence that the average is increasing is actually good news. Here is another example, if you are opposed to seat belt laws then you would consider it a success to be able to claim, "Our data provide no strong evidence that stricter seat belt laws reduce society's medical costs associated with traffic accidents."
- After you have finished doing your hypothesis test and establishing whether or not you reject the null hypothesis, go back and see whether or not you have confidence that the z-test we used are reasonably justifiable. Think about things like, "Does my stem-and-leaf plot look like the sample came from a normally distributed population?" If the population doesn't look normal, ask yourself if your sample is large enough to make you relatively confident in your results anyway. Report your answers in the last section of the report.

Notes for computation of t distribution and F distribution
The t-test and F-test are really no harder than the z-test, but you just have to be sure to use the correct distribution in any statistical software you are using. I will also provide an Python notebook example to get the correct distributions.

For the computations of t statistic, you can either be using table $A5$ in the back of the book, or if you have experience with statistical software, you can use that. If your number of degrees of freedom is not listed, then you could use statistical software that does have the t-distribution for your number of degrees of freedom. Or, you could use the line in Table $A5$ for the next smaller number of degrees of freedom listed.