

Math 1401 Elementary Statistics Project

Summer 2020 – Professor Pruenok

This project is worth 15% of your overall course grade.

The deliverables for the project will consist of written documentation of your work, submitted electronically in the iCollege assignment folders by 11:59 p.m. on the below Deadline dates. File formats should be PDF or anything compatible with Microsoft Office software such as Word or Excel. (Please do not submit .PAGES files from a Mac, my PC won't open them!) The checkpoints and final report will receive half credit if late, but only if they are submitted by the Late Deadline. Work submitted after the Late Deadline will receive no credit.

	<u>Deadlines</u>	<u>Points (100 total)</u>
Checkpoint 1: Framework	Monday, June 22	20
Checkpoint 2: Data Collection	Monday, July 6	20
Rough Draft (optional but encouraged)	Monday, July 20	0
Written Report	Monday, July 27	60
Late Deadline	Friday, July 31	Half credit

Checkpoint 1: Framework (20 points)

Submit written answers to these questions in the iCollege assignment folder by Monday, June 22.

1. What is the population you're going to study? (5 points)

If you're gathering data from people, then the population is the type of people you'll be gathering data from (Chapter 1). The individuals in the population should be the same type of individuals accessible by your sampling technique, so you have to be specific.

For example, suppose you want to study the impact of the coronavirus pandemic lockdown on the amount of time spent watching video content by individuals ages 16-25, and you intend to gather your data with an online survey of your Facebook friends. Any conclusions drawn could not be extended to hunter gatherer tribespeople in the Amazon ...right? A phrasing "people ages 16-25" is not specific enough as it would include individuals from that group and they aren't relevant to your study. It would be reasonable, however, to make such inferences for "Facebook users aged 16-25 in the United States" (or whatever country is predominant in your Facebook friends...provided that nation implemented a lockdown), because *that's* the type of person in your sample.

2. What data do you intend to collect? (10 points)

Here is where you define your variables (Chapter 2). Part of the project is on bivariate statistics (Chapter 4), so you'll need to choose TWO quantitative variables, i.e. that have numerical values you can "measure" regarding the individuals in your sample. Optionally you can include a third variable which is qualitative (like gender or eye color), if you're interested in that information.

Continuing the previous example, the variables could be stated as follows:

Variable 1: Average daily number of hours spent watching video content in February 2020

Variable 2: Average daily number of hours spent watching video content in April 2020

The variables are clearly stated, easy to understand, include units, and capture timeframes before and during the lockdowns. The importance of specifying units is explained in Checkpoint 2 below. If you'd like to refine your study, perhaps by exploring whether gender had any effect on changes in time spent watching video content, you could add:

Variable 3 (optional): Gender

3. What is a hypothesis that you'd like to test? (5 points)

A hypothesis is a claim that you think might be true or false about your population. It's essentially a guess about what results you're going to find, so it must be stated **BEFORE** you collect data and do statistical calculations. There are three types of hypothesis tests that will work for this project. Choose ONE of the following:

A. Quantitative variable: This is a statement about the mean of one of your variables.

Examples:

"The average daily number of hours spent watching video content by Facebook users aged 16-25 in the United States in during the coronavirus lockdown in April 2020 was 3 hours."

"The average daily number of hours spent watching video content by Facebook users aged 16-25 in the United States in during the coronavirus lockdown in April 2020 was greater than 3 hours."

B. Qualitative variable: This is a statement about a proportion that's identifiable in the data.

Examples:

"The population of Facebook users aged 16-25 in the United States is 55% female."

"The population of Facebook users aged 16-25 in the United States is more than 50% female."

“Most Facebook users aged 16-25 in the United States watched video content for at least 3 hours per day on average during April 2020.” *(This statement categorizes the quantitative variable into two qualitative groups, “at least 3 hours per day” and its complement “less than 3 hours per day”, thereby defining a qualitative variable from quantitative data, with “most” meaning more than 50%, i.e. more than half the population falls into “at least 3 hours per day”, which is a proportion.)*

- C. **Comparison between two groups:** If your population naturally breaks down into two groups, what effect do you think that might have on the value of one of your other variables? This comparison can be between two means or two proportions.

Examples:

Comparison of means: “The average daily number of hours spent watching video content by Facebook users aged 16-25 in the United States increased in April 2020 compared with February 2020.”

Comparison of means: “The average daily number of hours spent watching video content by Facebook users aged 16-25 in the United States increased more for males than for females in April 2020 compared with February 2020.”

Comparison of means: “Gender had no effect on the change in the average daily number of hours spent watching video content by Facebook users aged 16-25 in the United States in April 2020 compared with February 2020.”

Comparison of proportions: “For Facebook users aged 16-25 in the United States, more males increased their average daily time spent watching video content in April 2020 compared with February 2020 than did females.” *(This divides the population into groups by gender and the proportion with an increase in time watching video content (vs. decrease or no change in time) is compared for each gender group.)*

Checkpoint 2: Data Collection (20 points)

Submit written answers to these in the iCollege assignment folder by Monday, July 6.

1. Describe your sampling technique. (5 points)

Don't just say “simple random sample”, “systematic sample”, “Facebook survey”, etc...give details on exactly how you selected the individuals for your sample. Your sample should contain **at least 35** individuals.

2. Describe any biases that may be present. (5 points)

How might your sample not be representative of the population you're studying, and why? Biases will limit how well you can generalize your findings to the population, and it's very important to understand these limitations when interpreting and communicating your

findings to others. This is especially true if you used convenience sampling, like surveying your Facebook friends.

3. Describe how you collected your data. (5 points)

For example, if you're using a Facebook survey, explain how it was distributed and to whom, as well as the response rate (non-responders can't be included in the data set unless "respond to the survey" a qualitative variable you're measuring...), and include the survey itself.

For your quantitative variables, remember that it's important to specify the units. If your survey includes an open question (Chapter 1) like this:

Survey Question: "How much time per day on average did you spend watching video content in February 2020 before the coronavirus pandemic lockdown?"

Everyone knows what "how much time" means but, if you asked a question like that, you'd end up getting responses like this:

Responses: "A lot", "2 hours a day", "3-5 hours", "about an hour", "sometimes 3 hours but sometimes none at all", "30-45 minutes", "I'm not sure, but I binged a couple Netflix shows"

If you modify the question to replace "how much time" with "how many hours", that will take care of most of the problem but not eliminate possible responses like "3-5 hours" if the question is open-ended. A way to fix that would be to make it a closed question with answer choices like

- A. 0 hours
- B. 1 hour
- C. 2 hours
- D. 3 hours

Etc.

If your survey contains ill-phrased questions that result in bad data that you can't use, then you'll have to do your data collection all over again.

4. List the data you gathered in a table. (5 points)

If gathering data from people, it's typically reported anonymously. For example:

Individual	Average daily # hours watching video content in February 2020	Average daily # hours watching video content in April 2020	Gender
Person 1	1	2	Female
Person 2	8	7	Male
Person 3	0	1	Female
Etc....			

Written Report (60 points)

Submit written report in the iCollege assignment folder by Monday, July 6.

Submitting a rough draft for feedback from me at least one week prior to the final report deadline is optional but strongly encouraged. Also, if no final report is submitted, then the rough draft may be graded in its place.

The report should be written in **expository essay** format. (See [here](#) for a good description, but note that this is a generic structure and not specific to a statistical report). Remember that the point is to communicate your findings to other people, so ask yourself, “is this comfortable for another person to read and does it make sense?” You might want to test this with draft readers. If something makes sense to you but is confusing for them, like if they can’t tell what a graph is trying to say, then that suggests revisions are in order.

Your report is communicating mathematical findings and will need to include **graphical representations of data** inserted in the report as tables, figures, and charts, such as histograms, box plots, bar graphs, pie charts, scatter diagrams, etc. (See [here](#) for examples of good figures vs. bad figures. It also has links at the bottom for further reading.) The types of graphs to include will depend on the data they’re describing and what information is trying to be conveyed.

Of course, in all written communication it’s important to **use correct grammar, punctuation, spelling and capitalization.**

Outline

The written report should contain all the information this outline. PLEASE follow the order below otherwise it will be very difficult for me to grade it.

I. Introductory Paragraph (10 points)

Summarize the information in Checkpoint 1.

II. Data Collection (10 points)

Summarize the information in Checkpoint 2. If your table of data takes up a lot of space, it may be better to place it as an appendix at the end of your report. Think about the experience of your readers.

III. Univariate Descriptive Statistics (10 points)

Describe each variable separately so that information about different variables isn’t mixed together. I.e. complete all of part A for Variable 1, then all of part A for variable 2, etc.

A. Quantitative variables

- i. Include a histogram and boxplot appropriately labeled as figures in the report.
- ii. Describe the five number summary (min, Q1, M, Q3, max).
- iii. Describe the shape of the distribution (symmetric, skewed, bimodal, etc.).
- iv. Describe the measures of central tendency (mean, median, mode(s)). Which you think is best for describing your data and why?

- v. Describe the measures of dispersion (standard deviation, interquartile range). Which you think is best for describing your data and why?
- vi. Calculate the fences and identify any outliers in your data.
- B. Qualitative variables (if used)
 - i. Include a bar graph or pie chart appropriately labeled as a figure in your report.

IV. Bivariate Descriptive Statistics (10 points)

- A. For your two quantitative variables:
 - i. Include a scatter diagram of the data, appropriately labeled as a figure in the report. The least-squares regression line should be shown on the graph too.
 - ii. What is the equation describing the least-squares regression line? What do the slope and y-intercept mean in a real-world context?
 - iii. Calculate and interpret the linear correlation coefficient and coefficient of determination.
 - iv. Is there a linear association between your variables? (Reference the table in the back of your textbook stating the strength of correlation needed for the size of your sample.)
 - v. If your sample data has outliers, how do they affect the linear association between your variables? (Calculate the correlation coefficient and least-squares regression equation without the outliers and compare them to the ones with the outliers. Do the outliers make a significant difference? If so, do you think you would be justified in removing the outliers when making inferences about the population?)
 - vi. Based on whether your study is an observational study or a designed experiment, can you say whether the relationship between your variables is one of cause and effect?
 - vii. If your variables are linearly associated, discuss any potential lurking variables that could influence the variables and account for the relationship.
- B. If your hypothesis involves a comparison between two groups, include a graphical representation of the data that you think displays the comparison well (pie charts, side-by-side bar graph, etc.).

V. Hypothesis Testing (Inferential Statistics) (10 points)

- A. State both your null hypothesis H_0 and alternative hypothesis H_1 in words using appropriate statistical language. (If the hypothesis in your introduction is a statement of equality then it's the null hypothesis, otherwise it's the alternative hypothesis. If it's the null hypothesis, then you'll need to decide whether to conduct a one-tailed test (greater than, less than) or two-tailed test (different than).)
- B. Test your hypothesis at the $\alpha = 5\%$ significance level using **ONE** of the following options based on the type of hypothesis you chose in Checkpoint 1. (quantitative, qualitative, or comparison of two groups)

1) quantitative variable

- A. If your data has outliers, remove them before proceeding.
- B. Include a normal probability plot and discuss whether the population is normally distributed or not. If not normally distributed, that's why we started with a sample size of 35 so that removing outliers still (probably) leaves at least 30 in the smaller dataset.

- C. How many degrees of freedom are needed in the Student's t -distribution?
- D. Test the claim with a T-Test using either the
 - i. (Classical Approach) What is the test statistic t_0 and how does it compare to the critical value?
 - ii. (P-Value Approach) What is the P-value and how does it compare to the significance level?
- E. Do you reject or not reject your hypothesis? State your real-world conclusion using appropriate statistical language.

2) qualitative variable

- A. Check to see whether the assumptions $np_0(1 - p_0) \leq 10$ and $n \leq 0.05N$ are satisfied.
- B. Test the claim with a 1-Proportion Z-Test using either the
 - iii. (Classical Approach) What is the test statistic z_0 and how does it compare to the critical value?
 - iv. (P-Value Approach) What is the P-value and how does it compare to the significance level?
- C. Do you reject or not reject your hypothesis? State your real-world conclusion using appropriate statistical language.

3) comparison between two groups

- D. The test you use depends on the type of comparison you're making. It will be either
 - 1) T-Test comparing difference of means with matched pairs data
 - 2) 2-Sample T-Test using Welch's T-distribution comparing the means of two independent samples.
 - 3) 2-Proportion Z-Test comparing proportions of individuals in two groups
- E. Test the claim using either the
 - v. (Classical Approach) What is the test statistic and how does it compare to the critical value?
 - vi. (P-Value Approach) What is the P-value and how does it compare to the significance level?
- F. Do you reject or not reject your hypothesis? State your real-world conclusion using appropriate statistical language.

VI. Conclusions (10 points)

Summarize the relationships you found and the results of testing your claim. Did the results agree with what you thought would be the case before you studied them? Be sure to mention how any biases might affect your ability to generalize the results to the population and any sources of error that might be present.