

STAT 217: INTRODUCTION TO STATISTICAL CONCEPTS AND METHODS
California Polytechnic State University, San Luis Obispo
vskip10pt PROJECT

Phase	Points	Percent
1: Team selection	2	3%
2a: Project proposal	5	8%
2b: Project proposal meeting	2	3%
2c: Project proposal revision	3	5%
3: Data collection	5	8%
4: Preliminary report	11	18%
5a: Final report	30	50%
5b: Peer evaluation	2	3%
	60	100%

Seek help

If you have **any** questions or need **any** clarification at **any** point, please (1) come to office hours, (2) schedule an appointment to meet with me, or (3) post on the course discussion forum.

Work flow

Watch: [Getting Started on the Project \(2:06\)](#)

Download the `ProjectTemplate.Rmd` file from PolyLearn. Update this file for each phase of your project submission, and **submit the resulting rendered .html file** to PolyLearn (not the `.Rmd` file). When submitting a phase, you may simply update information for that phase and leave the remaining sections blank. Only one student per group needs to submit a phase.

After you have collected your data - if you want to share the `.Rmd` file between group members, (1) all sharing group members will need a copy of the data set saved to their computer, and (2) you will need to change the code used to import the data set to match the location of the data set on your computer.

Revisions

Revisions may be requested for any phase of the project. If a revision is requested you do not need to keep the old text that you submitted - you may simply delete older work and submit newer work.

Writing guidelines

- Rmarkdown does not automatically indicate misspelled words like Microsoft Word does. In order to execute a spell check in RStudio, submit `F7`. Please make sure you proof-read your writing carefully.
- Avoid using the word *correlation* unless you are specifically referring to the relationship between two quantitative variables. If you are not referring to two quantitative variables, you may discuss the relationship or the association between the two variables.
- When writing a paragraph or summary, do not refer to R variable names; instead, refer to the meaning of the variable. For example, I might have a variable in data set called `prev_stats` which indicates if a student has had previous experience with statistics prior to STAT 217.
 - It would be *incorrect* to write:
The percent of students with `prev_stats` is 35%.
 - It would be *correct* to write:
The percent of students with previous experience in statistics is 35%.
- If you start a sentence with a number, you must spell it out.
 - It would be *incorrect* to write:
64 people participated in the study.
 - It would be *correct* to write:
Sixty-four people participated in the study.
- When writing, round numbers (including p -values) to an appropriate number of decimal places. (You don't need to worry about rounding in your R output.)
 - It would be *incorrect* to write:
The average weight is 130.2384638 pounds.
 - It would be *correct* to write:
The average weight is 130.2 pounds.
 - An appropriate number of decimals to round a p -value is between 2 to 4 decimals. For really small p -values you may write “the p -value is less than 0.01”, if appropriate.
- Your projects will also be evaluated on
 - overall quality: coherent sentences, few grammatical mistakes, few typos, results discussed in context of research question
 - overall correctness: calculating appropriate descriptive statistics, choosing correct statistical methods, correct interpretation of results

Phase 1: Team selection

- **Identify** 3 to 4 classmates to work with. Include their full name and e-mail address in the project submission. If you need help finding a team, please let me know. Include this information in your Phase 1 submission. In addition, create your group on PolyLearn by clicking on [Select groups for project](#) in the Project content area of PolyLearn.
- **Designate** one team member to be the primary contact person.
 - This person is responsible for facilitating communication between team members.
 - This person is my primary contact in case I have questions for your team.
- Create your team **contract**: state three things that you think makes a group work well together that you all agree to do.

Phase 2a: Project proposal

[Watch: Example Study Design \(2:16\)](#)

[Watch: The Island \(3:26\)](#)

• Study Design Options

Your team may choose to do either an observational or an experimental study in the real world or on *The Island*. Selection one of the following four options:

1. Observational study in the real world
2. Experimental study in the real world
3. Observational study on *The Island*
4. Experimental study on *The Island*

Be sure to state in your project proposal which option your team chose.

Real world vs *The Island*

- In the real world, you don't have to conduct your study on people, but of course you may. Your observational units could be things other than people, like text books or cows. However, if you do conduct your study on people, make sure it is an ethical and reasonable study. Don't ask people to do anything or answer anything that might make them uncomfortable. It may be easier to conduct an observational study than an experiment in the real world.
 - *The Island* is an online environment in which you can survey fictional people. Be aware that the islanders do mimic real people - they give birth, die, get sick, refuse to answer your questions, and even sleep at night. It may be easier to conduct an experiment on *The Island* compared to the real world. For more details about *The Island*, see the end of this document.
- **Research question**: State a broad research question that your group would like to address. Motivate this question as to why it interests you.

- **Comparison groups:** Regardless of whether you do an observational or an experimental study, you are required to compare **two** groups. State the two groups that you are comparing.
- **Population:** Describe your population of interest.
- **Sample:** Describe how you will obtain sample from the population.
 - *Observational studies:* include a plan for you will *randomly select* participants from your population.
 - *Experimental studies:* include a plan for (1) how study participants will be identified (it is OK if they are not randomly selected), and (2) how you will *randomly assign* participants to the two experimental groups.
- **Sample size:** At a minimum, each team should have at least 30 participants in each of their two groups (for a total sample size of at least 60). You may choose to collect more data than that. State your target sample size per group.
- **Variables:** All teams must identify **three** variables related to their research question (one explanatory variable and two response variables). You should specifically state what variables you want to collect data on and how you will measure the variables.
 1. Categorical explanatory variable - this variable defines your two comparison groups
 2. Quantitative response variable - this variable represents something that you are measuring to compare between your two comparison groups
 3. Categorical response variable - this variable represents something that you are classifying to compare between your two comparison groups

Be sure you state how your variables will be measured:

- For categorical variables, what values will it take on? (e.g., little/some/a lot, or on campus/off campus)
- For quantitative variables, what are the measurement units? (e.g., If you are measuring time, are you doing it in hours, minutes, or seconds? If you are measuring performance on a memory task, is that measured in terms of a score or time until completion?)
- **Data collection:** How do you intend to collect your data? Will you be marking data on paper, a tablet, a laptop, or an online survey? Will team members interview individuals or will individuals read all questions themselves? Describe your data collection protocol in as much detail as possible.
- **Other protocol:** There may be some other details you need to address. Any other study details should go here.
- **Predictions:** Discuss what predictions you have for what you will discover. In your predictions, address both:
 1. the relationship between your categorical explanatory variable and your quantitative response variable
 2. the relationship between your categorical explanatory variable and your categorical response variable

Note: In this project, you will **not** examine the relationship between your quantitative response variable and your categorical response variable.

Note: Posting a survey on a Facebook page (or other form of social media) is not an acceptable study design.

Phase 2b: Project proposal meeting

- Select an appointment time to meet with Dr. Pileggi to discuss your project proposal. The more group members that can attend the better, but not all group members are required to attend.

Phase 2c: Project proposal revision

- Re-submit your project proposal according to any comments discussed with Dr. Pileggi.
- Make sure any changes you make are reflected throughout your study design. For example:
 - If you change a variable that you study, that should be reflected in your research question and in your predictions.
 - If you change the way you obtain your sample, then you may need to update your Data Collection.

Phase 3: Data collection

- **Wait** to do data collection until your project proposal has been approved.
- Import the data into R include the import code in an R chunk.
- Provide a summary of the data set with the command `summary(mydata)` in an R chunk.
- **Tips:**
 - Store your data in a *single* spreadsheet.
 - Your spreadsheet should have three columns for three variables:
 - Variable 1: categorical explanatory variable
 - Variable 2: quantitative response variable
 - Variable 3: categorical response variable
 - Save your spreadsheet as a `csv` file to import into R (you can import other file types, but we will primarily be using `csvs` throughout the quarter).
 - Keep the name of your data set short (one word) so it is easy to work with.
 - Keep the names of your variables short (one word) so they are easy to work with.
 - Be consistent in your data entry. For example, if you are entering gender make sure you agree that it should be entered as `male` and `female`. Otherwise, different students may enter `M`, `Male`, `male`, `m` which would then require you to do *data cleaning* before you analyze your data.

Phase 4: Preliminary report

- Produce a figure to examine the distribution of your quantitative response variable.
- Produce at least one additional figure that compares your two groups.
- Calculate appropriate descriptive statistics to compare your two groups. Be sure to address both (1) the relationship between your categorical explanatory variable and your quantitative response variable, and (2) the relationship between your categorical explanatory variable and your categorical response variable.
- Write brief paragraph describing your findings. This paragraph should include, at a minimum, the sample size achieved in each group, specific statistics per group, and commentary on what the figures show.
- **Tips:**
 - If you need help with the R commands, please review Lab 3! You learned all of the commands necessary to complete this phase in Lab 3.
 - Please note that formal statistical tests should NOT be submitted with this phase (there should not be a p -value or a confidence interval). This phase is only about describing your sample data.

Phase 5a: Final Report

- **Statistical methods:**
 1. State the statistical methods you will use to evaluate the relationship between
 - (a) your categorical explanatory variable and your quantitative response variable
 - (b) your categorical explanatory variable and your categorical response variable
 2. State the conditions necessary for each statistical method *in the context* of your research question (e.g., simply stating “normal” is not sufficient.)
 3. Evaluate if the conditions are satisfied. If a condition has an “or” in it, evaluate both parts of the condition even if one part is already satisfied. Comment on your findings.

Note: If your conditions are not satisfied you may discuss this in the Limitations section.
- **Statistical results:**
 1. Execute your statistical analysis.
 2. Write at least one paragraph summarizing your results in the context of your data. This should include discussion of *both* the evidence and strength of association.
- **Limitations:** State any limitations of your study. Pay particular attention to any measurement error that may have occurred, include discussion of *all* possible forms of bias, whether or not you can draw cause and effect conclusions, and if any conditions for statistical inference were violated. Justify your statements.

- **Conclusion:** Discuss the overall findings of your study and how they relate to your broad research question. Do your findings match your predictions? Suggest reasons for what you've observed (e.g., why do you think these groups differ? or are not very different?). Provide recommendations based on your analyses (recommendations may not apply to all research questions). Discuss what you might do differently next time. What related research questions could a future team investigate to build on your results?
- **Project title:** Modify the header of your Rmarkdown document to include an overall title for your project.
- **STOP!** Review writing guidelines prior to submission.

Phase 6B: Peer evaluation

- All team members are expected to contribute equally to the project.
 - Two points of your *individual* project grade comes from *completing* the peer evaluation form on PolyLearn.
 - After peer evaluations are assessed by the instructor, an *individual's* project grade may be adjusted by $\pm 20\%$ from the overall group grade.
- Here is how you will assess **yourself** and your peers:

Rate the individual on the following attributes according to the scale

1 = Strongly disagree, 2 = Disagree, 3 = Agree, 4 = Strongly agree

- * Communicated well with group members (electronically or in person)
- * Willingly volunteered for or accepted assigned tasks
- * Contributed positively to group discussions with useful ideas
- * Completed work on time or made alternative arrangements
- * Did work accurately and completely
- * Contributed a fair share to the project
- * Overall was a valuable member of the team

More about *The Island*

- All students have been added to the Island via their CalPoly email address.
- The Island has 38 villages. Each of these villages contains households of living Islanders as well as a cemetery where users can view details of Islanders who have died in the village. The combined population (living and dead) is 14,771 Islanders.
- To obtain information from an islander, use the map to click on the Village, and then the town, and then the resident, then an Islander. Once a person is selected, you will see some family information and his or her history. Click the link for "Obtain consent" to determine whether this person is willing to participate in your study. Then you can "set a task" which include "complete my survey" or you can impose a treatment or take a measurement. You should also add the person to your contacts so you can follow up with them later (e.g., which treatment group did you assign the person to?!).

- To survey the Islanders, follow the link for Survey and then edit your survey. There are about 50 survey questions to choose from. Longer surveys take longer for the Islanders to complete so dont ask too many questions. Also be aware that some Islanders may choose to lie on a survey.
- You can then view the results for a contact by either clicking on the person and view task history or in the contact list click on the results link. You have the option of collecting data daily or every thirty seconds
- I encourage you to play around with the Island to help you formulate your research questions.