Student Guide

Discovery Projects in Introductory Statistics

Dianna J. Spence, Ph.D.
Sherry L. Hix, Ph.D.
Thomas E. Cooper, Ph.D.
Robb Sinn, Ph.D.

Table of Contents

Overview

The best way to understand something is to experience it for yourself. That's why we recommend conducting your own statistical study, instead of just studying statistics from a book, video, or website. This guide will take you through the steps of conducting your own independent statistical research, from start to finish.
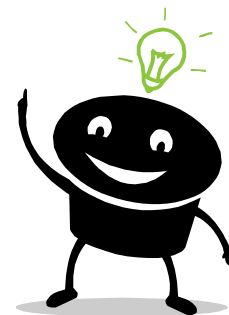
This guide covers two types of projects— a linear regression project and a comparison project. Your instructor may assign one or both of these projects to you during your statistics course.

In the sections below, we assume that you will define your research question, plan all the details of your project, and carry it out on your own. However, if there is not time for you to define and implement your own independent project, your instructor may assign a guided project instead. Guided project assignments are provided in Appendix E.

General Project Resources

The next several sections address tasks and issues that are relevant to both the linear regression and comparison projects.

Phases of a Statistics Project

Your project will consist of several phases, as described below. Each phase is addressed in detail in the sections that follow.

1. Choose your variable(s) and identify a research question
2. Devise a plan to collect your data
3. Submit a project proposal to the instructor and obtain approval
4. Once your research question and data collection plan are approved, carry out your research:
    a) Collect and organize data
    b) Conduct the appropriate analysis
    c) Write your results in a report
5. Submit your written report.
6. If your instructor assigns a presentation as part of the project, prepare to present your project to the class.

Selecting Variables and Identifying a Research Question

Your first order of business will be to decide what research topic you will investigate. Ultimately, you will need to: 1) identify the *variable(s)* you will study; and 2) articulate a specific *research question* about those variables. This process will also require that you identify the *population(s)* to which the variable applies.

Translating Informal Ideas into Formal Research Plans

One way to start the process is to pose "I wonder" statements. Below are two examples.
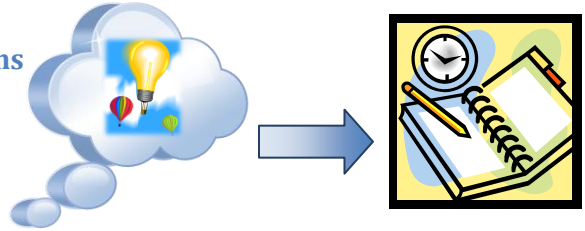
Example 1

"*I wonder if people in major cities tend to be more educated than people in small towns.*"

If you look carefully, there is a lot of information about a potential project bundled up in this sentence. It names two general categories, big cities and small towns. These would be the foundation for defining the populations for your research question, and they need to be well-defined: What counts as a big city? What counts as a small town? Should the populations be the cities and towns themselves? Or should the populations be the people who live in these big cities and small towns?

Also notice that the focal point of the question is level of education—this would be the basis for your variable. You would then need to define the variable explicitly, relating it to the specific population(s) you have identified. Here is one example of how the populations, variable, and research question could be defined for a *comparison project*:

The two **populations** are:
   a) Major U.S. cities, defined as cities in the U.S. with over 50,000 residents
   b) Small U.S. towns, defined as towns in the U.S. with fewer than 1000 residents

The **variable** of interest is "Percent of college-educated residents", defined as:
   The percent of residents 25 and over who hold a bachelor's degree or higher

The **research question** is:
   "On average, is the percent of residents with at least a bachelor's degree greater in U.S. big cities than in U.S. small towns?"

There are many potential ways we could have defined the variables and research question for this topic. Here are some important notes about the choices we made:

- The populations are the cities and towns themselves, because data about cities and towns are available on a "City Data" website (see Appendix A, Internet Data Sources.) If we defined the populations as the residents of the cities, we would then have to obtain data from individual residents in several cities and towns. This would be a difficult task.

- The variable is a characteristic of the populations we identified. For instance, we could **not** identify our variable simply as "education level" with 0 as "no college degree", 1 as "bachelors degree", and 2 as "graduate degree". Why? Because that variable would describe individual people, and we need a variable that describes entire cities and towns.

- The "City Data" website helped us refine many of our choices. We used their criteria for a major city and for a small town. We also used one of the measures of educational attainment that is reported on the website for every city. By examining what data were available, we narrowed down how to define our variables and research question.

Example 2

"*I wonder if people who are bigger perfectionists also tend to have more anxiety.*"
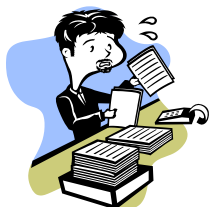
This statement describes two different characteristics people may have—level of perfectionism and level of anxiety. These could be the basis for two variables. The task remains to define each of these variables clearly: How would each variable be measured? How would meaningful numbers be used to represent the level of each one? What would be the possible values for each variable, and what would those values mean? If these variables can be quantified, the relationship between them can be explored.

You would also need to decide: To what population do these variables apply? The original statement says "people". Clearly, the variables (level of perfectionism and level of anxiety) apply to individual people. Should your population be "all people"? Is there reason to believe that you could collect data from a sample that would fairly represent all people everywhere? Is it safe to assume the results from a survey on your campus would be the same as results on another campus? Among older people? In another state? Another community or culture? Another country? You would use these considerations to help you select an appropriate population to identify in your research question.

Here is one example of how the population, variables, and research question could be defined for a *linear regression project*:

The **population** is:
>All students at Joliet Junior College

The **variables** of interest are:
a) "Level of perfectionism" defined as an individual's score on the "Almost Perfect" perfectionism scale, a 23-item survey. A person answers each item on a scale of 1 to 7. The score is the sum of all responses. Scores range from 23 to 161, with higher values indicating a greater degree of perfectionism.
b) "Level of anxiety" defined as an individual's score on the $DASS_{21}$ anxiety scale, a 7-item survey. A person answers each item on a scale of 0 to 3. The score is the sum of all responses. Scores range from 0 to 21, with higher values indicating a greater level of anxiety.

The **research question** is:
>"Among students at Joliet Junior College, is there a positive correlation between perfectionism and anxiety level?"

There are many potential ways we could have defined the variables and research question for this topic. Here are some important notes about the choices we made:

- The population was restricted to a particular college campus, because that is where we are able to collect data by administering surveys. It is most reasonable to assume that data collected exclusively on one college campus may represent students at that college, but not necessarily any other group of people.

- Surveys exist to measure both of these variables (see Appendix C, Collection of Surveys.) Therefore, the surveys and their scoring guidelines helped us to define how each variable would be measured.

Constructs

Example 2 above illustrates an important concept. Some variables are best measured with a collection of questions that result in a single numeric "score." These are sometimes called *constructs*. Surveys often combine multiple items to form or "construct" one reasonable overall measurement. For instance, some students are tempted to collect data with survey questions such as "On a scale of 1 to 10, how much anxiety do you have right now?" However, the 7 combined items on the anxiety survey in Example 2 above will almost certainly give a better measurement of a person's anxiety level.

You don't need to rely on a survey someone else has written to use this strategy. You can define a construct of your own.

For example, one group of students defined the construct of "dental hygiene" for their project. The team recorded the number of times per day participants reported brushing their teeth, flossing, and using mouthwash; then they settled on the following formula:

Dental Hygiene Score = (# of Brushings) + (# of Mouthwashes) + 3 × (# of Flossings)

The team gave more weight to flossing based on the reasoning that typical recommendations include flossing once daily, but brushing multiple times a day. A number of other reasonable formulas are possible; this example simply demonstrates how multiple pieces of data can be combined into a single construct that you can use as a project variable.

Additional examples of constructs are given in Appendix B, Variables and Constructs.

Tips for Selecting a Topic

The possible research topics are endless, so how should you go about picking a topic? Try some brainstorming, taking the following resources into account:

- You can find data on a vast number of different topics. Review the section Where to Find Data, later in this guide, for ideas that may help you think of a creative research topic.

- Look in Appendix D at some sample student projects. Although you should try to define your own project idea rather than duplicate a project someone else has already done, it may help spark your creativity to look at projects that other students have completed.

- Above all, choose a topic which holds some interest for you personally. Almost every career, hobby, activity, etc., has some data associated with it.

That said, you should approach your selection with care, as your choice will dictate how easily you can carry out the project, as well as the quality of your results. Here are some considerations that should guide your choice:

a) **You must be able to address your research question using the analysis required for the project, and your variables must lend themselves to the designated analysis.** For instance, you may be interested in investigating whether males and females differ in their choice of car color. However, if you are assigned a linear regression project, this question cannot be addressed using linear regression.

b) **You must be able to collect the data for the variable(s) you identify.** For instance, you might find it interesting to research the average level of education of various prison inmates. However, if you don't have a way to access this information, you won't be able to complete such a project successfully.

Where to Find Data

There are three primary sources to gather data. You may:
1) Use the Internet to obtain existing data
2) Collect and record data using measurement and/or direct observation
3) Create and administer surveys to gather data.

Each data source requires informed decisions by you and your instructor.

Using the Internet

The categories of data available via the Internet are almost endless. They include (but are not limited to) government and community records, restaurant nutrition information, sports statistics, and retail and consumer information. Appendix A contains several websites with large volumes of data that are readily available and current, as well as links to similar sites that may be of interest. These sites may provide all the data you seek, or they may simply provide a beginning point for you to find other data of interest.

It is important to gauge the reliability of a website before deciding whether to use data acquired there. For example, sites considered reliable include those sponsored by an official branch of the government; official sports leagues (e.g., NFL, MLB); network news agencies; publishers of respected journals or newspapers; restaurants, manufacturers, and other corporations providing product specifications; and reputable consumer research groups (e.g., Consumer Reports.) On the other hand, other websites that are not as trustworthy include personal web pages; sites that were constructed for student assignments or personal projects; networking sites that allow posts from the general public; blogs; and so on.

When it comes time to write your report, be prepared to give the name of the website where you obtained your data. Also be prepared to discuss whether or not data obtained from that website should be considered reliable.

Measuring and Recording Data Directly

The most direct form of data collection is to measure or observe something for yourself and write it down. Here are some examples of each of these.

Examples of measurement:
- Use a tape measure or ruler to measure length (or height or circumference); for instance:
  - Appendages (e.g., arms, legs, fingers)
  - Items found in nature (e.g., leaves, flowers, insects)
  - Merchandise (e.g., cameras, calculators, furnishings)

- Use a scale to find weights (e.g., animals, books, loaves of bread)

- Use a stopwatch to time events; for instance:
  - How long does a stoplight stay red?
  - How long can a person hold their breath?

Examples of observation:

- Count occurrences of events of interest; for instance:
    - How many times a person can snap their fingers in 30 seconds
    - How many times someone can skip rope before they trip
    - How many cars go through an intersection from 5:00 to 6:00 p.m.

- Record information from nutrition labels; for instance:
    - Milligrams of sodium per cup
    - Grams of fat per tablespoon
    - Calories per half-cup serving
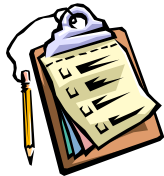
- Record the price of items of interest; for instance:
    - Real estate prices (e.g., per square foot for homes; per acre for land)
    - Automobile prices
    - Gasoline prices (per gallon)
    - Prices of grocery items

Creating and Administering Surveys

If you choose to administer surveys to collect data, you must either use an existing survey or create your own survey. Your survey might contain a collection of questions that you will combine to measure a *construct* (see discussion above, under Constructs). You may create your own survey construct or use one that has already been developed. Appendix C contains several surveys designed and validated by professionals to measure specific constructs. Appendix B also contains a collection of survey variables and constructs previously used by students or suggested by faculty members in a variety of disciplines. When using constructs, you should determine how a score will be calculated for the construct; however, the procedure for scoring should not be included with the survey given to participants.

The process of creating a survey is not arbitrary and should be carefully undertaken. You need a high quality survey to obtain good data, and high quality surveys must be carefully planned. For the purposes of these projects, the survey should be short, less than a page in length. Similar question types should be grouped together.

Data collected from surveys should always be kept confidential so that no individual may be linked to any individual responses. If the survey asks for any sensitive information, consider making the survey anonymous—instruct participants not to put their name on the survey. It also helps to have participants fold their responses up and insert them into a box. This helps participants feel more secure about answering the questions, which in turn will give you better data.

Writing Good Survey Questions

The questions on your survey should be easy to answer, and there are often advantages to closed response items.

Problematic: What is your favorite color?
Productive: Circle your favorite color from the choices given.
a) red   b) blue
c) green   d) another color

The number of possible answers for the first question (open response) is almost infinite. This may cause a problem for the participant since the choices may be overwhelming; and it may cause a problem for the data collector since decisions would need to be made about how to record and analyze this type of information.   The second question (closed response) eliminates the problem and guides the participant to the type of data needed by the researcher.

When choosing between open response and closed response questions, consider the advantages of each, as well as the range of responses you can expect to collect.

Problematic: How many Facebook friends do you have? Circle one:
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Problematic: How many Facebook friends do you have? Circle one:
0 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400
Problematic: How many Facebook friends do you have? Circle one:
0-49   50-99   100-149 150-199 200-249 250-299 300 or more

Productive: How many Facebook friends do you have? _____

The question has too many possible answers to enumerate them all in a closed response format. Consequently, the first list of possible answers (already too long and cumbersome) does not contain a response in the appropriate range for most respondents. The second list of answers can confuse respondents by skipping values, and it still runs the risk of not having answers in the right range. The third list of answers, although it covers all possibilities, limits the analysis you will be able to perform with the responses you collect. For instance, with answers given in a range (like those in the third list) you cannot compute a mean for a t-test or use the responses as coordinates (x or y values) in a scatter plot. The open response question in this case is more desirable, as it allows the respondent to give a specific and accurate value.

Closed response questions are advantageous when there are a limited number of acceptable answers. The question format allows the respondent to complete the survey efficiently by narrowing the range of responses and allowing the respondent to circle one quickly. However, the response options to such closed response questions should be carefully thought out.

Problematic: How many days in the last week did you exercise for at least 30 minutes?
Circle one: 1   2   3   4   5

Productive: How many days in the last week did you exercise for at least 30 minutes?
Circle one: 0   1   2   3   4   5   6   7

The first question does not list all the possible answers for the period given (one week). The second question does. Also remember that zero is often a valid response.

Where appropriate, questions should ask for a snapshot of recent events, not an estimate.

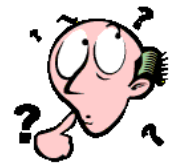> Problematic: On average, how many hours a night do you sleep?
>
> Productive: How many total hours have you slept in the last three days?

The first question asks for an estimate of the average. An answer for this question may be influenced by many factors, including the number of hours per night a participant thinks he should get or even the number of hours per night he has been told he gets. The second question eliminates most subjectivity for the participant, requiring the participant to remember a short period of time in recent history and to count total hours.

Questions should be precise, yet not cumbersome or confusing.

> Problematic: During the last month, how many total servings of alcohol have you consumed? A serving of alcohol is defined as a 12 oz. beer, an 8 oz. glass of wine, a shot of hard alcohol, or a 10 oz. mixed drink.
>
> Productive: In the last week, how many total servings of alcohol have you consumed (glass of beer/wine, shot, or mixed drink)?

The first question asks about a time period too far away from the present for the participant to answer with any certainty. The question also attempts to be as specific as possible by defining a serving of alcohol. However, the precision becomes too cumbersome to answer without frustration.

Surveys often contain Likert scale questions. These are typically statements that provide a range of numbers corresponding to the participant's opinion or level of agreement.

> Problematic: Rate your college on a scale of 1 to 5.
>
> Productive: Circle the answer that best describes your agreement with this statement: "I am glad that I am a student at this college."

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

The first question is too broad and vague; respondents may rate different aspects of the college (academic quality, social life, sports, etc.) The question also does not specify what 1 and 5 represent. The second question narrows the focus to a particular sentiment about the college, provides a specific statement, and provides sufficient guidance for selecting an answer.

Questions should avoid bias.

Problematic: Do you think the recent cuts to the unemployment program are fair to the poor, helpless citizens with no job who really need money to survive?

Very Unfair	Unfair	Neutral	Fair	Very Fair
1	2	3	4	5

Productive: Circle the answer that best describes your reaction to the recent budget cuts to the unemployment program:

Strongly Oppose	Oppose	Neutral	Favor	Strongly Favor
1	2	3	4	5

The first question "leads" the respondent and contains emotionally charged words and phrases. Those respondents who did not have a strong opinion may be persuaded by the wording of the question to answer a certain way. The second question states the issue and the possible answers neutrally, without value judgments or persuasive language.

Writing a Project Proposal

Your instructor may provide a project plan form for you to complete, or you may be required to write a more formal proposal. In either case, you will need to decide on the following aspects of your project, and you will need to describe them clearly in your plan:

1) State the research question you have chosen.

2) Specify the population(s) that will be represented in your study.

3) Identify the variable(s) that you will collect, including how they will be measured.

4) Describe your plan for collecting the necessary data. This should include:
   a. A description of where you will obtain your data (website, physical location, etc.)
   b. A sampling strategy
   c. A copy of your survey if you intend to use one

5) Describe your plan for analyzing the data you collect.

6) Your instructor may also ask you to explain what you *expect* to find when you analyze your data. (If you have no reason to expect a particular result, consider selecting a different topic.)

Your instructor may suggest revisions to your project plan before (s)he will grant approval. It is important that you not begin any additional steps in your research until your instructor has approved your proposal.

Collecting and Organizing Data

Once your proposal is approved, it's time to conduct your study. The first step is to collect your data. You should also organize the data so that you will be able to perform the desired analysis.

Sampling Strategies

Remember that your project requires you to collect information from a sample of individuals to represent a particular population. Therefore, before you begin sampling, two things should be very clear in your mind:
1) What population are you studying?
2) What is the best way to select a sample of individuals that will represent that population?

You may recall that the validity of various statistical procedures depends on obtaining a ***random sample***. This is easier to accomplish in some studies than in others. You should determine whether a truly random sample is possible for your study. If it is not, you should at the very least devise a plan to select individuals that fairly represent the population as much as possible.

Collecting Online Data

Because data are available in abundance on the Internet, it may be tempting to use every individual in a population for your project. For instance, you can easily find data on the Internet for every National League pitcher for the last 10 years. However, remember that part of your project assignment is to use a sampling technique and draw conclusions about the population based on statistics from your sample. Using data for the entire population can undermine this aspect of the project: If you can compute the exact mean salary of all National League pitchers, there is no need to hypothesize about it.

The good news is that when you find a comprehensive set of data for an entire population, you can obtain a truly random sample from it. For instance, in the case of the National League pitchers, you could place all of their names in a hat and draw 20 names at random; or you could assign each pitcher a number and generate random numbers to select the sample.

There are cases where the data set appears to represent a very small population, and you may feel you need all the individuals listed just to have enough data. In these cases, consider re-defining your population more broadly so that the data you have could be thought of as a sample from a greater population. For instance, suppose you have found a list of all NFL quarterbacks for this year. It's a small enough data set that you would like to be able to use the whole list. But then, how is your population any different from your sample? One way to deal with this is to define your population more broadly—do a little more research and find quarterbacks for the last several years; then sample randomly as described above.

Another example of this issue could occur if you are using the Internet to research menu items from a particular restaurant; you may find that there are only 12 entrees listed. These could be considered a sample of all the possible entrees that the restaurant ***tends*** to serve—especially if their menu changes often. You could also extend your population by combining the menus from several similar restaurants, so that you could sample from the larger set of entrees.

Collecting Data from People or Places

Obtaining a truly random sample is more difficult when you are administering a survey or recording data about items in a store, in a neighborhood, in a natural setting, etc. In these cases, be aware of the *purpose* of sampling randomly-- to ensure that every individual in the population has the same chance to be included in the sample.

To follow this principle as much as possible, you should carry out your data collection carefully, keeping examples such as these in mind:

- If you only administer surveys to college students in dormitories, you are excluding students who live off campus.

- If you only administer surveys to students in your classes, you may be excluding students with different majors or students who are in a higher or lower year than you.

- If you only administer surveys from 8:00 to 9:00 a.m., you are excluding students who sleep later, students who work at that time, students who are in class, etc.

- If you select arbitrary items on grocery store shelves, take care not to pick more items from shelves at eye level than from shelves that are very high or very low.

- If you select arbitrary items in any setting, take care not to pick items simply because they are closer, larger, more convenient, available in greater supply, more brightly colored, or more appealing.

Organizing Your Data

The most effective way to organize your data will be to use a spreadsheet. Typically, you would use one row for each individual in your sample, and one column for each variable.

If you conducted a survey with a multi-question construct, you may choose to score each survey first, and only enter the totals. Another option is to enter each survey response into a separate column and use a formula within the spreadsheet to compute the scores for you.

For tips on using Microsoft Excel to organize your data, consult the Statistics Project Technology Guide provided.

Analyzing Data and Writing a Report

Below is a general outline of the topics that should be included in your report for either project. The type of project you are conducting will govern the specific steps you should follow to analyze and report your data. For project-specific outlines, see the Project Tasks (LR) section for linear regression projects, and the Project Tasks (CP) section for comparison projects. These sections will give you more detail about what should be covered with each topic.

1. **Introduction.**
   State the topic of your study.

2. **Define Population(s).**
   Define the population(s) that you intend for your study to represent.

3. **Define Variable(s).**
   Define clearly the variable(s) that you obtained during your data collection; this should include information on how the variable is measured and what possible values this variable has.

4. **Data Collection.**
   Describe your data collection process, including your data source, your sampling strategy, and what steps you took to avoid bias.

5. **Study Design.**
   Describe the procedures you followed to analyze your data.

6. **Results: Descriptive Statistics.**
   Give the relevant descriptive statistics for the sample(s) you collected.

7. **Results: Statistical Analysis.**
   Describe the results of your statistical analysis.

8. **Findings.**
   Interpret the results of your analysis in the context of your original research question. Was your hypothesis supported by your statistical analyses? Explain.

9. **Discussion.**
   What conclusions, if any, do you believe you can draw as a result of your study? If the results were not what you expected, what factors might explain your results? What did you learn from the project about the population(s) you studied? What did you learn about the research variable? What did you learn about the specific statistical test you conducted?

Presenting Your Results

Your instructor may require you to present your project formally to the class. This presentation should describe your variables, your research question, the steps you followed during data collection and analysis, and your results. These are all topics that you should be able to pull from your project report and summarize briefly for an audience.

Below are some tips for making your presentation successful.

- Find out from your instructor what the allotted time will be for your presentation. Use this to decide how much detail you will include. Typically, this type of presentation is very brief (e.g., 5 minutes).

- Select the most relevant facts from your project and organize them into a coherent sequence.

- Use a presentation tool to display the information visually to the audience. These are popular options:
  - PowerPoint
  - Adobe
  - Prezi
    Prezi is an online presentation software tool that is free to students. See http://prezi.com/.

- Plan to spend 30-60 seconds on a screen. For a 5-minute presentation, you might aim for 5-8 screens. Use each screen to summarize one component of your project.

- Be selective and avoid cramming too much information onto a screen. Do not write information in paragraph form; summarize with brief bullet items.

- Set off the information on the screen with visual aids.
  - On some screens, you will want to show certain graphs (e.g., histograms for descriptive statistics, scatter plots for a linear regression project).
  - On other screens, you may choose to include an image that complements the subject matter; take care not to clutter the screen by overdoing it.

- It's very important to rehearse your presentation several times before giving it.
  - Get very familiar with each screen and with the order of topics.
 - Ensure that the presentation takes the right amount of time.

Evaluation: How Will My Project Be Graded?

Your instructor will evaluate your project using a scoring guide, also known as a scoring rubric. Below is an *example* of a rubric for scoring a student's work on a jigsaw puzzle.

Item	Poor(0-1)	Minimal (2)	Fair (3)	Good (4)	Excellent (5)
Completion *Weight: x3* 15 points	[0] No parts complete [1] 1-3 sides of outer edge complete	All of outer edge complete OR 1 significant region (25% of puzzle) complete	All of outer edge AND 1 significant region (25%) complete OR 2 significant regions (50%) complete	All of outer edge AND 3 significant regions (75%) complete	Entire puzzle complete
Strategies *Weight: x4* 20 points	[0] No strategy evident [1] Applied ineffective strategies	Applied one strategy listed at right	Applied two strategies listed at right	Applied three strategies listed at right	Applied all 4 strategies below: • Separated edge pieces first • Used image on box as a guide • Grouped similar pieces together • Paid attention to compatible shapes of puzzle pieces
Time 5 points	[0] Did not complete [1] Took over 3 hours	Completed in 3 hours or less	Completed in 2 hours or less	Completed in 1 hour or less	Completed in 30 minutes or less

Because the project has many components, a scoring rubric can help the instructor identify and evaluate each of the project requirements, as well as weight them appropriately to determine your final score. Using a rubric also helps your instructor to grade each project submission the same way, so that all students' scores are based on a consistent set of criteria.
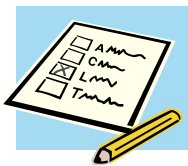
Ideally, your instructor will provide you with a copy of the scoring rubric (s)he intends to use to evaluate your project. If you have this information in advance, you can use it as a guide when preparing your project for submission. Try scoring your own project using the rubric. If you encounter something that is missing or insufficient, you can refine your project further. This is a good strategy to ensure you have not omitted a critical component of the project.

The Linear Regression Project

Project Design (LR)

In this project, you will select 2 quantitative variables that you believe may be related in a specific population. After collecting your data, you will investigate the relationship by constructing a scatter plot, computing correlation, and conducting a linear regression analysis.

Here are some example variables and corresponding research questions and hypotheses:

1) Batting averages and salaries among professional baseball players
    Population: Professional baseball players
    Explanatory Variable: Batting average
    Response Variable: Annual salary
    Research Question:
      Is there a correlation between professional baseball players' batting averages and their salaries?
    Hypothesis: Players with higher batting averages have higher salaries.
              [**positive correlation is hypothesized**]

2) Calories and fat content in grocery store snack foods
    Population: Grocery store snack foods
    Explanatory Variable: Fat content in grams per serving
    Response Variable: Calories per serving
    Research Question:
      Is there a correlation between snack food calories and fat content?
    Hypothesis: Snack foods with higher fat content also have higher calories.
              [**positive correlation is hypothesized**]

3) Engine horsepower and gas mileage in automobiles
    Population: Automobiles
    Explanatory Variable: Engine horsepower (HP)
    Response Variable: Gas mileage (miles per gallon)
    Research Question:
    Is there a correlation between cars' horsepower and their gas mileage?
    Hypothesis: Cars with higher horsepower have lower gas mileage.
    [**negative correlation is hypothesized**]

4) Perfectionism score and Anxiety score among students at your school.
    (*see surveys in Appendix C*)
    Population: NGCSU students
    Explanatory Variable: Perfectionism score
    Response Variable: Anxiety score
    Research Question:
      Is there a correlation between students' perfectionism and anxiety scores?
    Hypothesis: Students with higher perfectionism scores have higher anxiety scores.
              [**positive correlation is hypothesized**]

Project Tasks (LR)

Recall the overall sequence of tasks you will follow to complete the linear regression project.

1. Choose a research question:
   a. that can be addressed using linear regression
   b. for which you can collect data to analyze

2. Devise a plan to collect your data

3. Submit a project proposal to your instructor and obtain approval

4. Once your research question and data collection plan are approved, carry out your research:
   a. Collect data
   b. Conduct linear regression analysis, referring to guidelines below
   c. Write your results in a report, using the outline given below

5. Submit your written report in the format specified by your instructor (on paper, on a web page, in a pdf file, etc.)

6. If assigned by your instructor, prepare a presentation of your project design and results.

Project Report Outline (LR)

Below is a detailed outline of the content that should be included in your project report. Your instructor may give you additional guidelines; these may include exclusions, extensions, supplements, or other modifications to the outline as shown below.

These components are listed in outline form so that they can be used as a checklist. However, your project report is expected to be a formal paper (not an outline). Your results should be stated in complete sentences, and your paper should be written in paragraph form. Although you may choose to use headings, you should not number your paragraphs.

1. **Introduction.** State the topic of your study as a research question and/or as a specific hypothesis that you tested. Your hypothesis should indicate what type of correlation you expected to see (positive or negative) and how strong you expected the correlation to be (weak, moderate, or strong). Your hypothesis should describe a specific result that you expected to find AND the *practical reason* that you expected this result (your rationale).

2. **Define Population(s).** Define clearly the population(s) that you intend for your study to represent. (Examples: all NFL football players, all cars manufactured this year, all biology majors at your school, all small towns in the Southeastern U.S., all PetSmart shoppers in your city, etc.)

3. **Define Variable(s).** Define clearly the variable(s) that you obtained during your data collection (e.g., age, salary, price, miles per gallon, score on a particular personality test, miles commuted one-way to school daily, etc.) This must be specific: "time spent watching TV" is too vague; "number of hours spent watching TV in the last 3 days" would be specific enough. If your variable is a measurement (e.g., height) give units (e.g., inches). If your variable is a score (e.g., on a personality quiz), give the range of possible scores (e.g., 0 to 15).

4. **Data Collection.** Describe your data collection process and sampling strategy. If you located data on a website, provide the URL and tell how you selected individuals from that website to include in your sample. If you obtained data from an agency, office, store, or other similar source, explain where you went and how you selected individuals to include in your sample. If you surveyed individuals directly or took measurements, describe how you selected individuals for the sample. If you used a survey, this section must include a copy of your survey. If you collected measurements, describe the device you used (e.g., tape measure, odometer, scale, stopwatch, etc.) If participants were required to do something to be measured, describe the measurement process (e.g., "each participant was asked to hold their breath as long as possible; the number of seconds they held their breath was timed with a stopwatch.") *No matter what data collection process you used, address: a) what steps you took to avoid bias in your sample; and b) whether you believe the sample(s) you obtained were representative of the population. Tell why or why not.* Your instructor may also request that you include a table with all of your raw (not summarized) data as an appendix at the end of the report.

5. **Study Design.** Identify the statistical procedures you used to analyze your data. Give relevant design details (e.g., which variable was selected as the explanatory variable, and which the response variable? Why? What type of correlation did you expect? And so on.) If you are required to conduct a significance test to determine the significance of your correlation, state your null and alternative hypotheses about the correlation, both in words and in appropriate mathematical symbols.

6. **Results: Descriptive Statistics.** Give descriptive statistics *for **each** of your two quantitative variables*. Note that you will be reporting summary statistics for both your explanatory variable and your response variable. Report each set of descriptive statistics using both a table and a chart as described below. All tables and charts should be placed directly in your report.
    a) Table: Give sample size, mean, standard deviation, and 5-number summary.
    b) Chart: Show a histogram that illustrates the distribution of the variable.

7. **Results: Statistical Analysis.** Report the results of your analysis; include items below.
    a) Scatter plot with a graph of the regression line
    b) Value of the correlation coefficient *r* and interpretation of its meaning
    c) Equation of the regression line
    d) An example of a prediction using the regression equation
    e) Discussion of the slope of the regression line and its meaning
    f) Value of R^2 for the regression model and interpretation of its meaning
    g) Indication of significance of correlation, with significance level (if required)

8. **Findings.** Interpret the results of your statistical analysis in the context of your original research question. Do your analyses support your expected findings? Explain.

9. **Discussion.** What conclusions, if any, do you believe you can draw as a result of your study? If the results were not what you expected, what factors might explain your results? What did you learn from the project about the population(s) you studied? What did you learn about the research variables? What did you learn about the specific statistical analysis you conducted?

The Comparison Project

Project Design (CP)

There are several scenarios for which a t-test can be used to conduct a comparison project. It is possible that your instructor has also covered other statistical tests that may be used to conduct a comparison study; however, those techniques are beyond the scope of this guide. Check with your instructor if you have questions about what statistical tests you might be able to use.

Several t-test option scenarios are named and described briefly below, with an example of each.

1. Elements: *one population, one quantitative variable, a specific value for hypothesis*

   Example scenario:

   Population: All 2-bedroom apartments within 10 miles of your school

   Variable: monthly rent

   Research question: Is the average rent for a 2 bedroom apartment within 10 miles of the school more than $600 per month?

   Hypothesis and rationale: "We predicted that the average monthly rent for a 2 bedroom apartment within 10 miles of the school would be higher than $600. A college planning guide stated that students could probably share a 2-bedroom apartment off-campus 'for about $600 a month.' Based on casual observation, we believe that on average, this estimate is too low."

   Appropriate test: The 1-sample t-test

2. Elements: *one quantitative variable, two populations in which individuals can be matched*

   Example scenario:

   Populations: All canned vegetables sold at Kroger; all canned vegetables sold at Wal-Mart

   Variable: price

          (*specifically, the ***difference*** in price charged for identical items at each store)

   Research question: On average, are canned vegetables cheaper at Wal-Mart than at Kroger?

   Hypothesis and rationale: "We predicted that on average, Wal-Mart would charge less for the same canned vegetable item. This prediction is based on the general observation that things seem to cost less at Wal-Mart."

   Appropriate test: The matched pairs t-test

   Design note: Hypothesis is about the computed <u>difference</u> between the two related values

3. Elements: *one population, two quantitative variable values for each individual*

    Example scenario:

    Population: All students at your school

    Variables: rating (1 – 10) of the taste of Coke; and rating (1-10) of the taste of Pepsi
    (*specifically, taste tests are conducted "blind" and the **difference** in ratings is
    computed for each participant)

    Research question: On average, do students at this school prefer Coke over Pepsi?

    Hypothesis and rationale: "We predicted that on average, students at this
    school would rate Coke better in taste than Pepsi. This prediction is based on
    the fact that in such close proximity to Atlanta, the Coca Cola capital, Coke is
    often the preferred soft drink."

    Appropriate test: The matched pairs t-test

    Design note: Hypothesis is about the computed <u>difference</u> between the two related values

4. Elements: *two populations, one quantitative variable*

    Example scenario:

    Populations: 1) All current NFC football players; 2) All current AFC football players

    Variable: age

    Research question: Is there a significant difference between the average age of NFC players
    and AFC players? (*by "significant" we mean <u>statistically significant</u>*)

    Hypothesis and rationale: "We predicted that there might be a
    difference between the average age of NFC and AFC players. In recent
    years, the AFC has been perceived as the stronger conference; one could
    theorize that on average, they have older players who are wiser and
    more experienced; OR one could theorize that on average, they have
    younger players who are in better shape and have more stamina. The
    purpose of this study is to determine if any such age discrepancy exists."

Project Tasks (CP)

Recall the overall sequence of tasks you will follow to complete the comparison project.

1. Choose a research question:
   a. that involves comparison using a quantitative variable
   b. for which you can collect data to analyze

2. Devise a plan to collect your data

3. Submit a project proposal to your instructor and obtain approval

4. Once your research question and data collection plan are approved, carry out your research:
   a. Collect data
   b. Conduct appropriate analysis, referring to guidelines below
   c. Write your results in a report, using the outline given below

5. Submit your written report in the format specified by your instructor (on paper, on a web page, in a pdf file, etc.)

6. If assigned by your instructor, prepare a presentation of your project design and results.

Project Report Outline (CP)

Below is a detailed outline of the content that should be included in your project report. Your instructor may give you additional guidelines; these may include exclusions, extensions, supplements, or other modifications to the outline as shown below.

These components are listed in outline form so that they can be used as a checklist. However, your project report is expected to be a formal paper (not an outline). Your results should be stated in complete sentences, and your paper should be written in paragraph form. Although you may choose to use headings, you should not number your paragraphs.

1. **Introduction.** State the topic of your study as a research question and/or as a specific hypothesis that you tested; your specific hypothesis should note a statistically significant result that you expected to find and the ***practical reason*** that you expected this result (your rationale).

2. **Define Population(s).** Define clearly the population(s) that you intend for your study to represent. (Examples: all NFL football players, all cars manufactured this year, all biology majors at your school, all small towns in the Southeastern U.S., all PetSmart shoppers in your city, etc.)

3. **Define Variable(s).** Define clearly the variable(s) that you obtained during your data collection (e.g., age, salary, price, miles per gallon, score on a particular personality test, miles commuted one-way to school daily, etc.) This must be specific: "time spent watching TV" is too vague; "number of hours spent watching TV in the last 3 days" would be specific enough. If your variable is a measurement (e.g., height) give units (e.g., inches). If your variable is a score (e.g., on a personality quiz), give the range of possible scores (e.g., 0 to 15).

4. **Data Collection.** Describe your data collection process and sampling strategy. If you located data on a website, provide the URL and tell how you selected individuals from that website to include in your sample. If you obtained data from an agency, office, store, or other similar source, explain where you went and how you selected individuals to include in your sample. If you surveyed individuals directly or took measurements, describe how you selected individuals for the sample. If you used a survey, this section must include a copy of your survey. If you collected measurements, describe the device you used (e.g., tape measure, odometer, scale, stopwatch, etc.) If participants were required to do something to be measured, describe the measurement process (e.g., "each participant was asked to hold their breath as long as possible; the number of seconds they held their breath was timed with a stopwatch.") *No matter what data collection process you used, address: a) what steps you took to avoid bias in your sample; and b) whether you believe the sample(s) you obtained were representative of the population. Tell why or why not.* Your instructor may also request that you include a table with all of your raw (not summarized) data as an appendix at the end of the report.

5. **Study Design.** Identify the statistical test you conducted to analyze your data. Also tell which *type* of design you used (one-sample, matched pairs, etc.) Give other design details (e.g., was it 1-sided or 2-sided? Left-tailed or right-tailed?) State your null and alternative hypotheses, both in words and in appropriate mathematical symbols. If you used a matched pairs design, explain clearly how the pairs of values were matched and how the difference was computed (e.g., by subtracting before – after, left – right, etc.)

6. **Results: Descriptive Statistics.** Give descriptive statistics *for **each** data set*. Note that t-tests for 2 independent samples will require statistics for 2 sets of data-- one for each of the two separate samples. Matched pairs t-tests will require statistics for 3 sets of data-- one for each of the 2 related values (e.g., before and after) and another one for the difference between each pair of numbers. Report each set of descriptive statistics using both a table and a chart as described below. All tables and charts should be placed directly in your report.
    a) Table: Give sample size, mean, standard deviation, and 5-number summary.
    b) Chart: Show a histogram that illustrates the distribution of the variable.

7. **Results: Statistical Analysis.** Report the results of your test; include the test statistic, degrees of freedom, and the p value of the significance test.

8. **Findings.** Interpret the results in the context of your original research question. Do your analyses support your expected findings? Explain. Interpret p-values and discuss significance levels.

9. **Discussion.** What conclusions, if any, do you believe you can draw as a result of your study? If the results were not what you expected, what factors might explain your results? What did you learn from the project about the population(s) you studied? What did you learn about the research variable? What did you learn about the specific statistical test you conducted?

Appendices