Collaborative Statistics: Projects: Bivariate Data, Linear Regression and Univariate Data*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

In this project, students will collect a sample of bivariate data and analyze the infomration. Students will be asked to describe the center and spread of the data, determine the goodness of fit of a linear regression model, and analyze the relationship between the variables.

1 Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

2 Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

3 Part I: Bivariate Data

Introduction

_ _ _ _ State the bivariate data your group is going to study.

*Version 1.6: Feb 19, 2009 2:52 pm -0600
†http://creativecommons.org/licenses/by/2.0/

EXAMPLES: Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

_____ Describe how your group is going to collect the data (for instance, collect data from the web, survey students on campus).

_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.

_____ Conduct your survey. Your number of pairs must be at least 30.

_____ Print out a copy of your data.

Analysis

_____ On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.

_____ State the least squares line and the correlation coefficient.

_____ On your scatter plot, in a different color, construct the least squares line.

_____ Is the correlation coefficient significant? Explain and show how you determined this.

_____ Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.

_____ Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.

_____ Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

4 Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____ Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.

_____ Answer the following, rounded to 2 decimal places:

1. Sample mean =
2. Sample standard deviation =
3. First quartile =
4. Third quartile =
5. Median =
6. 70th percentile =
7. Value that is 2 standard deviations above the mean =
8. Value that is 1.5 standard deviations below the mean =

_____ Construct a histogram displaying your data. Group your data into $6 - 10$ intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .

_____ In complete sentences, describe the shape of your histogram.

_____ Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in chapter 2 (since you are now using univariate data) to determine which values might be outliers.

_____ Construct a box plot of your data.

_ _ _ _ _ Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_ _ _ _ _ Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

5 Due Dates

- Part I, Intro: _ _ _ _ _ _ _ _ _ _ _ (keep a copy for your records)
- Part I, Analysis: _ _ _ _ _ _ _ _ _ _ _ (keep a copy for your records)
- Entire Project, typed and stapled: _ _ _ _ _ _ _ _ _ _ _

_ _ _ _ Cover sheet: names, class time, and name of your study.
_ _ _ _ Part I: label the sections "Intro" and "Analysis."
_ _ _ _ Part II:
_ _ _ _ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
_ _ _ _ All graphs requested in the project.
_ _ _ _ All calculations requested to support questions in data.
_ _ _ _ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges.

NOTE: **Include answers to ALL questions asked, even if not explicitly repeated in the items above.**