

Guidelines for Statistics Class Project

Math 140: Introductory Statistics

April 13, 2009

The purpose of your class project is to carry out, in a real-life setting, all the different techniques of statistical analysis we have learned this semester. The most important concepts we have learned are statistical inference (i.e., test of confidence and test of significance) and regression. Recall that inference is based on the Central Limit Theorem which may be applied to either the Sample Means random variable, \bar{X} , or the Sample Proportions random variable \hat{P} . In your project you will perform inference for each of these two random variables. Also recall that we are unable to have any reliable statistics unless we have a random and unbiased sample that is a well-representative of the underlying population. Hence, we must pay extra attention to our design for data production.

For your project you may study any question that interests you. Use your imagination and creativity to explore and come up with questions that intrigue you. As a general recommendation, in order to see statistics that resemble your homework problems, use quantitative variables. Some typical questions that have been studied by Math 140 students in the past are provided at the bottom of these notes.

CONTENTS: Your class project is a report that should be submitted according to the following **ordered** list. You will be graded based on the mathematical content and accuracy of your report, and the presentation of your results.

- 1) **Table of Contents:** Include a table of contents for your report.
- 2) **Sample Survey Form:** Include a blank copy of your sample survey form.
- 3) **Data Production:** Explain how you obtained your sample. Namely, explain how the locations, time slots, days, and combinations with your group members were done. Provide all detail including the lines in Table B that were used.
- 4) **Central Limit Theorem:** This section has two parts.
 - a. **Sample Means R.V., \bar{X} :** (i) Explain how your 100 sample means, \bar{x} , were found; (ii) Give the probability distribution table for \bar{X} ; (iii) Give the histogram for the distribution of \bar{X} ; (iv) Describe the distribution of \bar{X} ; (v) Give a five-number-summary for the distribution of \bar{X} ; (vi) Give the mean, \bar{x} and the IQR for the distribution of \bar{X} .
 - b. **Sample Proportions R.V., \hat{P} :** (i) Explain how your 100 sample proportions, \hat{p} , were found; (ii) Give the probability distribution table for \hat{P} ; (iii) Give the histogram for the distribution of \hat{P} ; (iv) Describe the distribution of \hat{P} ; (v) Give a five-number-summary for the distribution of \hat{P} ; (vi) Give the mean, p and the IQR for the distribution of \hat{P} .

5) Data Analysis:

- a. Your report should contain **at least one of each** of the following diagrams.
 - Histogram
 - Stemplot
 - Bar Chart
 - Pie Chart
 - Boxplot
- b. Each distribution in (a) should be accurately described.
- c. When applicable, give a five-number-summary for each distribution in (a).
- d. When applicable, give the mean and the IQR for each distribution in (a).

6) **Linear Regression:** For this section you will need to make a hypothesis that involves two variables: an explanatory variable and a response variable [Example: The more a CSUN student studies during a week (X), the higher his/her GPA (Y)]. Your report for this section consists of the following.

- a. Specify your hypothesis, the explanatory variable X , and the response variables Y .
- b. Make a scatterplot of your data. Mark (with a different color or symbol) any outlier that you observe in your data.
- c. Describe the scatter plot. Does the scatter plot support your hypothesis?
- d. Calculate r and r^2 once for the entire data, and once without the outliers. What can we conclude about the nature of the linear correlation between X and Y using these two numbers. If your data contains any outliers, what is their effect on the values of r and r^2 ? Does that make sense? Why?
- e. Find the equation of lines of regression once using all of your data, and once excluding the outliers. Then, graph both lines (using different colors) on your scatter plot. Does the difference between the two lines make sense? Which line do you think we should use for making predictions for values of X and Y ? How is your answer to this question related to the value of r^2 ?
- f. Pick an arbitrary value for X . Calculate the corresponding predicted \hat{y} -value using the two lines of regression in (e). Now plot these two points (using different colors and/or symbols) on your scatter plot. Which predicted \hat{y} -value is more accurate? Why?
- g. Is X the only variable that influences the outcomes of Y ? Suggest a few lurking variables for this study. How important do you think the influence of lurking variables are for the outcome of your hypothesis?
- h. Based on the statistical analysis you performed above, evaluate the validity of your hypothesis and make your final conclusion(s).

7) **Inference for a Population Mean, μ :** This section has two parts.

- a. **Test of Confidence:** Find a 98% confidence interval for a population mean, μ . Provide the conclusion of this test in a complete sentence.

- b. **Test of Significance:** Make reasonable null and alternate hypotheses for a population mean μ . Use a test of significance to make a conclusion about your hypotheses. The conclusion of this test should be given in a complete sentence and referring to the underlying variable X .

8) **Inference for a Population Proportion, p :** This section has two parts.

- a. **Test of Confidence:** Find a 98% confidence interval for a population proportion, p . Provide the conclusion of this test in a complete sentence.
- b. **Test of Significance:** Make reasonable null and alternate hypotheses for a population proportion p . Use a test of significance to make a conclusion about your hypotheses. The conclusion of this test should be given in a complete sentence referring to the underlying variable \hat{P} .

9) **Self and Groupmate Evaluation:**

This report is meant to be a collaborative activity. All group members are required to participate equally in the production of this report. Thus, in order to maintain fairness to all group members you are required to submit an honest evaluation of your participation in this report, as well as that of all other students in your group. For your convenience an evaluation form is posted to my website. Please download a copy of this form, complete it as accurately as possible, place the completed form in an envelop, and submit the sealed envelop directly to me. No one other than me will have access to your confidential evaluations. You will NOT receive a grade for your class project until you have turned-in your confidential group evaluation envelop.

Sample Questions. Below are a few typical questions and hypotheses.

- a. On the average how many hours per week does a CSUN student work (watches TV, studies, commutes, spends online, uses cell phone, texts, etc)?
- b. What proportion of CSUN students own a pet (car, cell phone, computer, smoke, iPod, iPhone, etc)?
- c. On the average how many tattoos (body piercings, sexual partners within the past year) does a CSUN have?
- d. Opinion on abortion (death penalty, premarital sex, US foriegn policy, environmental issues, etc).
- e. Correlation: The more a CSUN student drinks (watches TV, chats online, texts, etc) per week, then lower his/her GPA.