Math 113 - Semester Project - 100 points

Your task is to perform some real-world inferential statistics. You will take a claim that someone has made, form a hypothesis from that, collect the data necessary to test the hypothesis, perform a hypothesis test, and interpret the results. If you use pre-existing data, rather than collecting it yourself, then you will need to do more analysis to get the full points.

You should try to come up with something of interest to you instead of some contrived situation. Several groups have tested to see if their company met their sales goals. One person tested to see if 60% of patients show up for their doctor's appointment in the clinic where she worked. One waitress tested to see if the average tip was 15% and another tested to see if gender plays a role in the amount of the tip.

You may work in groups of up to three (3) persons. Pick people you can work with; part of the grade will be assigned by the people in the group as to the work you contributed. Do not necessarily pick your friends, pick people who will do a good job.

You need to submit a proposal defining what it is that you wish to test and how you wish to go about testing it (think back to the types of sampling). The instructor will peruse these proposals, make suggestions and give it back to you. The proposal will be included as part of the final project. If your group can't decide on a project or needs help defining it, see the instructor.

Be sure to define clearly what the population being studied is (Richland Students, People listed in the Decatur phone book, People driving a car, etc).

Some of your projects will require information from chapters in the book not yet covered. See the instructor if you have trouble identifying these areas. If you read the section(s) and don't understand the material, see the instructor for an explanation. Don't wait for the class to cover the material, it may be too late.

Make sure you get the project cleared with the instructor before you go collect the data. One person wanted to telephone survey some people and was talking in the range of $100 phone bill if she called everyone she said she was going to. The project should not cost you very much money to implement. It will take some time, however, and you should not wait until it's due to get started on it.

While you are not precluded from doing any of the given examples, it is certainly better if you can come up with something original which has an interest to you. Things like the Pepsi vs Coke, while acceptable, are really poor choices.

The instructor will keep a copy of your final project.

The project will be comprised of several parts, due at different times during the remainder of the semester.

Project Components

Proposal (5 points)
This is to make sure you're on the correct track before wasting lots of time collecting useless information. Your proposal should also include a time line of when you will have the different components of your project completed. Include when you plan to have your data collected by, when you'll run the analysis, when you'll have the rough draft completed, and when you'll have the final draft completed. Your proposal should include your population, parameter of interest, and sampling frame as well as the who, what, where, when, why, and how for your project. Must be typed.

An *excellent* proposal (5 points) will have the following components.
- The title for the project
- A list of the group members with correct spellings of first and last names
- An explanation of why you find the topic interesting
- The claim being tested
- An identification of the type of test: Is there one, two, or several samples? Are you testing a claim about proportions or means or are you testing linear correlation?
- The context of the data to be collected to test the hypothesis
- A timeline for completion of the project

Status Report (5 points)
The main purpose of the status report is to remind you that there is a project and that you need to be working towards that because it's coming due. What I would like to see in the report is what you've accomplished towards your project so far, even if you haven't accomplished anything. Remember that in your proposal, you put down dates for completing certain tasks. One way to do the status report is to keep a journal of your project. This journal would log the activities of the group, when you worked, who participated, who was assigned what tasks, what was accomplished, etc. The status report must be typed.

An *excellent* status report (5 points) will have the following components.
- The title for the project
- A list of the group members with correct spellings of first and last names
- The tasks accomplished so far
- Difficulties encountered
- Revised timeline (if necessary)

Rough Draft (10 points)
This is a rough draft of the final report so the instructor can suggest corrections. The rough draft is the complete report except that you get a chance to be corrected before the final grade is assigned. Include everything you plan on including in the final report. This includes any graphs, tables, and text. I strongly urge you to make an appointment with the Student Learning Center to have someone proof your draft. The rough draft must be typed and the narrative portions should be double spaced.

The grade here will be based on having the components of the final report present, not on their statistical correctness. This is your chance to make mistakes before it really affects your grade.

Final report (60 points)
The final report will include a description of the problem, and why you think it is important, or what you hope to gain from testing the hypothesis. It should also include the context of the data, all data collected, and the values generated by Minitab or the calculator. A decision and conclusion should be stated. An analysis should follow with what the conclusion means in terms of the original problem. The final report should be in narrative format, must be typed, and should be double spaced.

An *excellent* final report (60 points) will have the following components.

- The title for the project
- A list of the group members with correct spellings of first and last names
- An introduction to the problem including the claim(s) being tested.
- How and when the data was collected including possible problems
- Descriptive statistics
- Appropriate graphs
- Inferential statistics including ...
    - the null and alternative hypotheses written symbolically
    - any assumption violations with the data or sampling method
    - statistical output including a test statistic and p-value
    - the decision and a conclusion written in terms of the original claim
- Conclusion
- Suggestions for the next time this project is done
- No statistical usage errors

Presentation (10 points)
Classroom presentation of 3-5 minutes on why you picked the project you did, and what your results were. There is a projector with Powerpoint on it if you would like to make a slideshow. You can also make transparencies or write on the board if needed. The class and/or instructor may ask questions on why you did something the way you did. These points will be assigned by the other class members, not by the instructor. You will be assigning point totals to the group as a whole, not each individual member of the group. The grade you receive will be the average of the grades the class gives you. If you are not here for your presentation or the presentations of any of the other groups, you will receive a zero for this portion of the project.

Each group presentation will be rated as excellent, average, or poor in the areas of teamwork, effort in preparation of presentation, clarity of presentation, knowledge of project, and correct statistical usage.

Individual evaluations (10 points)
This is the only part of the project that is not a group grade.
Evaluate yourself and the other people in your group as to their ability to perform within a group and their ability to do assigned work. On a single sheet of paper, write a separate paragraph about each person in the group, including yourself, and what they contributed or did not contribute to the group effort. Assign a grade between 0 and 10 to each member. The grade you receive will be the average of the scores from each of the group members. The individual evaluations must be typed.

If you work alone on the project, you must still complete an evaluation to receive these points.

What can we test?

Some things are easier to test than other things. The purpose of this project is not to do a full-scale PhD level research project, it is to expose you to the process of hypothesis testing in a real-world application. You may test means, proportions, or linear correlation. It is also possible (but not covered in your textbook) to test a standard deviation. You may have one or more samples. You may categorize your variables in one or two ways.

If you are dealing with one sample, then you will need some numerical value to test against. The claim "more people prefer Pepsi than Coke" becomes a claim that the proportion of Pepsi drinkers is greater than 0.5. There are not two independent samples (Pepsi drinkers / Coke drinkers), just one sample categorized in two ways. A problem with the Pepsi / Coke thing is that it omits other soft drinks because that is more difficult to do. A chi-square goodness of fit test would be more appropriate in this case. Realize that some of their topics are really lame and you should, if at all possible, come up with a claim that you have heard or that interests you rather than one out of the book.

Categorical Data

If your data consists solely of categories and not measured quantities, then you should be looking at proportions or counts. Chapters 19-21 tell you how to conduct a test about a single proportion, chapter 22 tells about testing two proportions, and chapter 26 talks about dealing with 3 or more categories and tests for independence (when there are two ways of breaking down the results).

Things to look for that let you know you're dealing with categorical data or proportions include: proportions, percents, counts, frequencies, fractions, or ratios.

This list is a guideline, but counts can also be used as quantitative data as well. You really need to think about the response that was recorded for each case (a row in Minitab terms). Did you record a yes/no response for each case or did you record a number that means something? If it was a yes/no or other categorical data, then this is the place to be.

Example Claims:

- 93.1% of Americans feel there should not be nudity on television during children's viewing time. http://www.parentstv.org/PTC/publications/lbbcolumns/2003/0528.asp
  - This is a claim about a single proportion. We know this because the value includes a percentage and the data is categorical (yes or no), not numerical. The original claim here could be written as p=0.931.
- Blacks are more likely to die from a stroke than whites. http://www.msnbc.com/news/875288.asp
  - Depending on how this is analyzed, it could either be a comparison of two independent proportions or one dependent sample. If you go ahead and read the article, it talks about there actually being four race groups, which would suggest independent samples and a test for independence (see next problem). If you compare the percent of blacks that die from stroke to the percent of whites that die from stroke, then you have a test of two independent proportions from chapter 22 and your original claim could be written as $p_b > p_w$. If you look at just those who died from strokes, then you have only one sample and each person can be classified as either black or white (success or failure). In that case, you're testing that the proportion of blacks is greater than 50% and your original claim could be written as p>0.50. If you look at just those that died from strokes but you include the other two races mentioned in the article, then you have a chi-square goodness of fit test and you can look at the number of blacks, whites, Hispanics, and

other races that died to see if they're all 25% or if one is different. If one is different, then you could look at the blacks to see if it was higher than the others.

- Sexual orientation is related to how strongly someone feels about a written nondiscrimination policy that includes sexual orientation. http://www.louisharris.com/news/allnewsbydate.asp?NewsID=678
  - This is actually a test for independence out of chapter 26. There is the categorical variable for sexual orientation (heterosexual or gay/lesbian/bisexual/transgendered) and a categorical variable for agreement with the statement (strongly agree, somewhat agree, neutral, somewhat disagree, strongly disagree). It is possible that grouping variables can be used for measurement data as well, but what makes this fall in the categorical data is that each response is counted, not measured.

Quantitative Data

If your data consists of measured quantities, then you will probably be testing a mean or perhaps correlation between two variables. It is possible to test a claim about a standard deviation, but that is rare, and not covered in your text.

There are four main ways to analyze means. A single mean (requiring a numerical value) is discussed in chapter 23. Comparison of two independent means is discussed in chapter 24 and the comparison of two dependent means is discussed in chapter 25. Finally, comparison of three or more independent means is discussed in chapter 28, which is on the CD, but not in the textbook.

You can also perform correlation and regression with two quantitative variables. Simple regression, with just one predictor variable, is covered in chapter 27 and multiple regression, with several predictor variables, is covered in chapter 29 on the CD.

Example Claims:

- Americans have sex an average of 138 times a year. http://www.durex.com/GSSresultsetup.htm
  - This is a claim about a mean (average). This one might seem a little confusing, given that it's a count of something and "count" was one of the keywords for categorical data, but that's why I included it here as an example. Think about what you record for each person (case). Did the data have a yes/no (categorical value) or a number? In this case, each case would have a number and we found the average of those numbers. In the categorical situations previously mentioned, each case would have been yes or no or some other category. The original claim could be written as $\mu=138$.
- Women live five years longer than men. http://stacks.msnbc.com/news/743069.asp
  - This is a claim about two averages, the average lifespan of women and that of men. We don't know the average of either gender (they're given in the article), we just know that women are supposed to live five years longer than men. When you're working with one sample, it's important to have a value to compare against, but with two samples, you don't need a value for each, just the difference between the two (in this case 5 years). The original claim here could be written as $\mu_w-\mu_m=5$ (the difference in the mean ages of women and men is 5 years).
- Gasoline costs more on the West Coast than other regions. http://tonto.eia.doe.gov/oog/info/gdu/gasdiesel.asp
  - This information comes from the US Department of Energy and includes a sampling frame of 115,000 gas stations from across the country. The US is broken down into regions of the East Coast, Midwest, Gulf Coast, Rocky Mountain, and West Coast. Since we are looking at the average of more than two independent samples, we'll use the Analysis of Variance from chapter 26. Notice that there is only one measurement variable (gasoline prices) but there is also a

categorical variable (region). The categorical variable is used only for grouping purposes. The ANOVA tests that all the means are equal, written as $\mu_E = \mu_M = \mu_G = \mu_R = \mu_W$.

- There is a direct relationship between the amount of ozone and the emergency room visits. http://www.meha.net/PDF/Air-Pollution-and-Asthma_MLH.pdf
  - This was on page 8 of the document linked above and involved a five summer study in New Jersey. Here we have two quantitative variables ozone level and number of emergency room visits. The ozone level is measured in parts per million (ppm) and the emergency room visits would be counts. Remember that you can not perform correlation and regression with categorical variables. The original claim that there is a relationship would be written as $\rho \neq 0$.

Some previous projects:

These are some of the many projects that students have worked on before. You should not limit yourself to these topics, but they may give you guidance for picking your topic. Topics that are related to people's work usually turn out to be the best projects.

You can also get ideas from reading newspapers or online news sites. I went to MSNBC.com and typed in keywords like "average", "more likely", or "correlation" to get some of the claims I used in the examples.

- Are the rates paid by the insurance company for dental cleaning in line with the rates charged by the dentists? A student called 30 dentists to find out the rates.
- Does the blood type ratios in McClean county agree with the national percentages as published by the American Red Cross? Students went through Red Cross records using stratified sampling until there were over two hundred people in the sample.
- Do people prefer Pepsi over Coke? People's preference was asked and then they were given a taste test.
- Are the men and women's shoe prices at Foot Locker, MC Sports, and Finish Line the same?
- Does Firestone/Bridgestone produce splices with the mean size claimed?
- Is the GPA of smokers lower than the GPA of non-smokers?
- Do higher priced bullets have a smaller shot pattern?
- Do Chex potato chips have 60% less fat than their competitors?
- Can Wonder Bread claim they have 200% more calcium than other regular white breads?
- Is there a difference in GPA between students coming from public and private high schools?
- Do patrons at Cheddar's tip 15%?
- Are there more absences on Fridays than on Mondays?

Sample Final Report

Available online are some sample projects prepared by the instructor. There are sample student projects available in the classroom. These are older projects from a different textbook that didn't emphasize the context of the data or the sampling techniques, so don't follow them too closely for your projects.

An Analysis of College Algebra Exam Scores

Are College Algebra scores different depending on the chapter? Are there differences between male and female students? Grades from the Fall 2000 section of Math 116 were compared to look for statistically significant differences.

A Comparison of Textbook Prices between Richland's Bookstore and Online Textbook Stores

Textbook prices between were compared in the Summer 2000 term to see if Richland's bookstore was more expensive as many students felt.