Math 113 - Statistics 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 may work alone, but a (very small) part of the final evaluation will be your ability to work as a team.

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. If your group can't decide on a project or needs help defining it, see the instructor.

The proposal is due after we have covered the basics of hypothesis testing, but 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 chapters 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.

The instructor will keep a copy of your final project and rough draft.

The project will be comprised of several parts, due at different times during 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 must be typed, double spaced, and printed.

Wait for approval from the instructor before you beginning your project.

An *excellent* proposal (5 points) will have the following components.

- A list of the group members with correct spellings of first and last names
- The title of the project
- An explanation of why you find the topic interesting
- The claim being tested
- An identification of the type of test: Are there one, two, or several samples? Are you testing a claim about proportions, means, or linear correlation?
- A timeline for completion of the project

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 printed. The narrative portions should be double spaced, but tables and computer output may be single 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.

The rough draft will be returned to you for corrections, but you should turn it back in to the instructor with the final report.

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 like you were writing for a newspaper or magazine, must be typed, printed, and should be double spaced.

An *excellent* final report (60 points) will have the following components.
- The title of 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
- The context (who, what, where, when, why, how) of the data (remember this is in narrative format)
- Descriptive statistics and/or tables depending on your type of data
- Appropriate graphs (every project should have at least one graph or chart of the data in it)
- Inferential statistics including ...
    - the null and alternative hypotheses written symbolically
    - statistical output including a test statistic and p-value
    - a graph showing the critical and non-critical regions, 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 (15 points)
Classroom presentation of 3-5 minutes on why you picked the project you did, and what your results were. It is highly recommended that you create a PowerPoint slideshow to go with your presentation.

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 as well as 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 for 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. Your score will be a combination of the scores given you by each member of the group and the instructor's evaluation of your evaluation.

Turn in a summary paragraph of what each person in the group (including yourself) did and how many points out of ten you would give them for their effort. Be sure to put your name, section number, and what the assignment is at the top of the sheet. These evaluations should be typed up individually and emailed to the instructor. The evaluations may be sent in the body of an email, they do not have to be a separate attachment. The other students in the group will not see what you wrote about them, just the average score they got from all of the students.

You need to evaluate everyone in the group including yourself. If you're the only person in the group and did all of the work, you still need to evaluate yourself or you'll miss out on the participation grade.

When the instructor grades your evaluation, he is looking for things like the quantity and quality of material written about each person, whether the evaluation was submitted on time, whether the instructions were followed, etc.

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 (in your textbook, but not covered in class) 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. You should try to come up with a claim that you have heard or that interests you.

Categorical Data
If your data consists solely of categories and not measured quantities, then you should be looking at proportions or counts.

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

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 about Categorical Data

- 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.medicalnewstoday.com/medicalnews.php?newsid=64812
  Depending on how this is analyzed, it could either be a comparison of two independent proportions or one dependent sample. 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 and your original claim could be written as $p_b > p_w$. However, if the wording was "stroke victims are more likely to be black than white," then your population would just be stroke victims and you would have one sample. 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.
- Sexual orientation is related to how strongly someone feels about a written nondiscrimination policy that includes sexual orientation. http://www.harrisinteractive.com/news/allnewsbydate.asp?newsid=972
  This is actually a test for independence. 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 (Numerical) 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 this course.

There are four main ways to analyze means.

1. A test about a single mean that requires a number as the claimed value.
2. A test about two independent means doesn't need a number because you compare them to each other. This compares the same thing in two different groups.
3. A test for two dependent means, often called paired samples, compares two values for each case in the same group.
4. The Analysis of Variance is an extension of the two independent samples case where there are more than two groups.

You can also perform correlation and regression with two quantitative variables. Simple regression, with just one predictor variable, is covered in the book. Multiple regression, with several predictor variables, is not covered in the textbook but is available online.

Example Claims about Quantitative Data:

- Americans had sex an average of 111 times a year according to a 2004 survey. http://www.durex.com/cm/gss2004Content.asp?intQid=398&intMenuOpen=11
  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 $\mu=111$.

- Women live five years longer than men. http://www.medicalnewstoday.com/medicalnews.php?newsid=18866 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. 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 $\mu_E = \mu_M = \mu_G = \mu_R = \mu_W$.
- Seat belts save lives. http://dot.state.il.us/trafficsafety/seatbelt june 2006.pdf and http://www-fars.nhtsa.dot.gov/FinalReport.cfm?stateid=17&title=states&title2=fatalities_and_fatality_rates&year=2005 Okay, this claim is all over the place, but I wanted to give some links on how it would be tested. You could take the data regarding the percent of people wearing their seat belts and compare it to the fatality rate. These are two numerical values that are paired together for each case (probably based on an annual report). Remember that you can not perform correlation and regression with categorical variables. The original claim that seat belts save lives would be interpreted as a negative correlation (as seat belt use goes up, fatalities go down) and would be written as $\rho<0$.

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 typed in keywords like "average", "more likely", or "correlation" to get some of the claims used as 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 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?

- Is the average drive-through time at Culver's less than 270 seconds?

Sample Final Report

Available online are some sample projects prepared by the instructor. I do not expect your projects to be as long or detailed. There are sample student projects available in the classroom on the filing cabinet.

An Analysis of College Algebra Exam Scores - http://www.richland.edu/james/spring01/m113/algebra.pdf

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 - http://www.richland.edu/james/summer00/m113/textbook/report.pdf

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