Homework Assignment 1

Due Date: Monday, October 3rd

Note: There are 7 problems on this assignment. The first 4 provide examples and extra practice. They are relevant for the exams but do not need to be turned in. Solutions for them are available on the class web site. You must turn in Problems 5-7 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Monday, October 3rd. Officially it is due in class but you may turn it in to the TA Adam King's mail folder (in CHS 51-254) any time before 3:00 with no penalty.

Note: Output from any calculations done in STATA or SAS MUST be included with your assignment for full credit. If I do not specify which way to do a problem you may chose whether to do it by hand or on the computer. All the STATA/SAS commands needed to complete this homework are given at the end of the assignment and will be reviewed in the lab. You do not need to hand in a separate lab report–simply turn in the relevant output as part of your homework.

Note: You are encouraged to work with fellow students, as necessary, on these problems. However, each of you MUST write up your solution ON YOUR OWN and IN YOUR OWN WORDS. The style of your write-up is as important as getting the correct answer. Your solutions should be easy to follow, and contain English explanations of what you are doing and why. You do not have to write an essay for each problem, but you should give enough comments so that someone who has not seen the problem statement can understand your work. You do not have to type your assignments. However, if they are too sloppy to read, too hard to understand, or give just numbers with no comments, you WILL lose points. Problems labeled (GS) are adapted from our optional text, *Primer of Applied Regression & Analysis of Variance* by Stanton Glantz and Bryan Slinker.

Warm-up Problems

(1) Bad News Burgers (WBCH): Hamburger Heaven has just found out that one of it's customers has become ill after eating a burger infected with E. Coli. They are concerned that a whole shipment of meat may have been spoiled and want to be able to assure their customers that the food is safe so they decide to test several randomly selected packages of patties.

(i) Hamburger Heaven's null hypothesis is that the meat is contaminated. Explain briefly why this makes sense.

(ii) Explain in English what the type I and type II errors and power for this problem would be. Would the company want a small α or a small β if it had to choose one or the other? Explain.

(2) Nuclear Test Scare: A manager of public health services in an area downwind of a nuclear site is worried that citizens in his area are in danger. To get the plant operations changed he needs strong evidence. He decides to test whether or not the mean amount of radiation (measured in picocuries) in the bone marrow of residents who live downwind of the plant exceeds the average level of $\mu = 1$ found in citizens not living near the plant. Measurements of bone marrow radiation are taken from nine citizens who live downwind. The data values are

$$10, 2, 2, 2, 2, 5, 5, 7, 1$$

(a) State the null and alternative hypotheses for this test, both mathematically and in English, with a justification of your choice.

(b) Find estimates of the mean and standard deviation of the amount of bone marrow radiation in people living downwind from the power plant. You may do the calculations either by hand or in STATA or SAS.

(c) Find a 98% confidence interval for the mean amount of radiation in the bone marrow of people living downwind from the plant.

(d) Find the probability that the average radiation level in a sample of size $n = 9$ would exceed the esimtated value you found in part (b) if the nuclear plant had NO effect on radiation levels. (In other words, if x is the value you got for the sample mean in part (b), find $P(\bar{X} \geq x)$.) Hint: You may not be able to get this value directly from your probability table (which one should you be using?) but you can work out approximately what it is, or you can use STATA or SAS to perform the test and get the exact value.

(e) Suppose that you need evidence at the 99% confidence level that citizens living downwind have increased levels of radiation before you will be able to get the plant administrators to take action. Based on the your results from parts (c) and (d), what should happen? Explain briefly.

(f) What is the estimate of effect size for comparing the amount of radiation found in the bone marrow of people downwind from the plant compared to the general population? Using the conventions we learned in class, is this a small, medium or large effect?

(3) Teaching Aids: This problem is modified from a textbook I used to use in my previous life across town called *Statistics for Management and Economics* by Watson, Billingsley, Crofts, and Huntsberger.

Midcontinent Polytechnical Institute requires all its students to take Statistics 101 in their freshman or sophomore years. In the past year alumni donations were used to set up a statistics lab with reference books, computers, tutors, and videotaped lectures. Dr. Goss, the chairman of the statistics department, is interested in knowing whether all statistics students would benefit from using the tapes. To study this issue, he surveyed his previous semesters class and compared the test scores of students who had used the tapes to those who had not. Dr. Goss data is shown below.

Used Tapes: 79, 89, 66, 72, 70, 81, 95, 87, 83, 77, 94
Didn't Use Tapes: 77, 45, 99, 66, 80, 90, 86, 55, 41, 75, 25, 65

(a) What is the population of interest to Dr. Goss? Explain.

(b) Do the students in Dr. Goss' class constitute a sample from that population? If so, do you think they are a random sample? Explain.

(c) Why is the central problem of this study an example of statistical inference?

(d) What are the variables in this study? State whether each of the variables in this study is quantitative or qualitative and explain why.

(e) What problems might be caused by the study design employed by Dr. Goss? Explain.

(f) Find a 90% confidence interval for the mean difference in test score between students who used the tapes and students who did not and carefully interpret your result. You should calculate the interval first by hand and then repeat the calculation in STATA and SAS.

(g) Perform the hypothesis test that Dr. Goss would use to determine whether there is a difference in exam scores between students who do and do not use the tapes. State your null and alternative hypotheses mathematically and in words with a justification of your choice and calculate the test statistic by hand. Perform the test in STATA and SAS to obtain the p-value and explain your real-world conclusion using significance level $\alpha = .10$.

(h) What is the mean difference in exam score between students who do and do not use the tapes? What is the effect size for the difference? Based on these numbers do you feel the tapes have a meaningful effect from a student's point of view? Explain.

(4) Practicing Power: Use STATA and SAS to perform the following power and sample size calculations. Parts (a)-(c) involve a test of a single mean (1-sample t-test). Parts (d)-(g) involve a comparison of two means (two-sample t-test). In all cases assume $\alpha = .05$ and the tests are two-sided.

(a) Find the sample size needed to achieve 80% power if the null hypothesis mean is $\mu_0 = 0$ and the alternative hypothesis mean is $\mu_A = 5$ and the standard deviation is $\sigma = 10$.

(b) Find the sample size needed to achieve 80% power for an effect size of $d = 1$.

(c) What is the power of the test to detect an effect size of $d = .5$ if the same size is n = 30?

(d) Suppose you want to have two samples of equal size and you are testing $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. If $\mu_1 = 0, \mu_2 = .5$ and $\sigma_1 = \sigma_2 = 1$, how many subjects should you have in each group?

(e) Repeat part (d) assuming you will have twice as many subjects in group 1 as group 2. (Do this in STATA.)

(f) Repeat part (d) but assume you have 30 subjects per group and compute the power rather than the sample size.

(g) Repeat part (d) but assume you have 60 subjects in group 1 and 30 subjects in group 2 and compute the power rather than the sample size. (Do this in STATA.)

Problems To Turn In

(5) Exercise Power: A nutrition professor is studying a new diet and exercise program to help children lose weight. She has piloted the study in a small sample of n=9 children and has found that the average weight loss during the study period is $\bar{X} = 5$ pounds and the standard deviation of the weight loss is $s = 10$ pounds. (Note that a negative weight loss would correspond to a weight gain.)

(a) Find a 98% confidence interval for the mean weight loss of children on the diet and exercise program and carefully interpret it. Do this first by hand and then verify your calculations in STATA or SAS.

(b) Suppose that the professor wants to prove that on average chidren using her program lose weight. Perform the appropriate hypothesis test, making sure to state the null and alternative hypotheses mathematically and in words with a justification of your choice, calculate the p-value and give your real-world conclusions. Use $\alpha = .01$. Set the p-value calculation up by hand and then obtain the actual value from STATA or SAS.

(c) Explain in English what the Type I and Type II errors for this problem would be.

(d) What is the observed effect size for this study?

(e) The professor claims that the effect size you computed in part (d) is fairly large and that this indicates the weight loss program really is useful–she just didn't have enough power to prove it. Does her statement have any merit? Discuss.

(f) Suppose that the true mean weight loss for children in this program is $\mu = 5$ and the standard deviation is $\sigma = 10$, just as you observed in the sample. Find the power for the test in part (b) if the sample size is (1) n = 9, (2) n = 25, (3) n = 100. You may use STATA, SAS or an online power calculator. Does your answer have any bearing on part (e)? Discuss.

(g) How would your answer to part (f) change if (1) you made the test 2-sided (2) you used $\alpha = .05$ or (3) the standard deviation were 5 instead of 10. Discuss briefly in each case. You do not need to redo the calculations.

(6) Class Database Part 1: So far, 38 people have provided data on height and gender for the class database. The values may be found in the accompanying HW1 data set in the columns labeled Height1 and Gender1. Heights are given in inches and gender is coded as M/F.

(a) Use STATA or SAS to find summary statistics for the heights of men and women (treated as two separate groups). What is your best estimate of the difference in height between men and women?

(b) Using the values from part (a), find a 95% confidence interval for the difference in height between men and women based on this data and carefully interpret your result. Calculate the interval first by hand and then verify it in SAS or STATA. (Note: you will need to compute the pooled estimate of the standard deviation from the individual standard deviations for the hand calculation.)

(c) Perform an hypothesis test to determine whether there is a difference in height between men and women. State the null and alternative hypotheses mathematically and in words with a justification of your choice, compute the test statistic and set up the p-value calculation by hand. Then perform the test in SAS or STATA to get the exact p-value and explain your real-world conclusions using $\alpha = .05$. Is your result consistent with part (b)? Discuss briefly.

(d) What is the effect size for the difference in heights between men and women? How large would this effect be considered using the standard conventions we learned in class?

(e) What is the minimum effect size we could detect with these sample sizes assuming we want 80% power? (Note that this will be easier to do in STATA.)

(f) Assuming the true difference in means is 6 inches and the standard deviation of heights is 3 inches (regardless of gender) what is our power for the test in part (c) with the given sample sizes? (Note that this will be easier to do in STATA.) Overall, do you think my "study" was well powered to detect gender differences in height? Discuss briefly.

(g) The columns labeled Height2 and Gender2 contain the heights and gender for 29 randomly selected students from last year's class. Do you think that the heights of students in this year's class differ from those of last year's class either overall or within each of the gender groups? Justify your answer by computing appropriate estimates and tests.

(h) Give one potential problem with comparing the overall average height of the sample from this year's class with that of the sample from last year's class. Do you think that your suggested problem actually had much effect with these samples?

(7) Go to the link below and watch the clip (my life is sometimes like this!)
http://www.genomeweb.com/blog/researcher-takes-biostatistician

(a) To the extent you can determine it from the clip (i) What are the population and sample proposed by the researcher? (ii) What is the inference the researcher is trying to make and what are the corresponding hypotheses? (iii) What information does the statistician want in order to be able to tell whether the researchers' sample size is sufficient?

(b) The minimum detectable effect size is the smallest effect you can detect with a specified power, sample size and significance level. What is the minimum detectable effect size in this context assuming the researcher has 3 subjects per group and wants 80% power using a two-sided significance level of $\alpha = .05$? You may use either STATA or SAS to do the calculation. Do you think it is likely that the funding agency will be happy about the minimum detectable effect size? Explain briefly.

STATA and SAS Commands

To do this assignment you will need to compute summary statistics and confidence intervals, perform t-tests and do power and sample size calculations in STATA or SAS. The necessary commands are given below.

(1) STATA COMMANDS:

(a) To obtain basic descriptive statistics, use the **summarize** command, which can be abbreviated by **su**. If you want to get descriptive statistics for subgroups of a data set you can use the option **by** if the data are sorted by subgroup or **bysort** if they are not. Examples are as follows:

To get the summary statistics for the radiation variable in warm-up problem 2, type

summarize radiation

To get the summary statistics of the exam scores for students who did and did not use the statistics tapes in warm-up problem 3, type

bysort usedtapes: summarize testscore

(b) To get confidence intervals in STATA there are several options. If you already know the mean(s), sd(s) and sample sizes you can use the **cii** command which stands for "confidence interval instant". If you have the data entered in a variable you can use the command **ci**. The hypothesis test commands below also generate confidence intervals. Examples are below:

To get a confidence interval for the mean radiation level in warm-up problem 2 using the variable in the HW1 data file you would type

ci radiation

By default STATA produces a 95% confidence interval. If you want a different α you use the **level** option, e.g. to get a 98% interval type:

ci radiation, level(98)

If you have already calculated the sample mean and standard deviation using the summarize command you would find $\bar{X} = 4, s = 3$ and you could instead type

cii 9 4 3

which uses in order the sample size, the sample mean and the sample standard deviation. To do a confidence interval for a difference in means it is easiest to use the t-test command below. However you can use the **ci** command with the **by** option to get separate confidence intervals for two subgroups. For instance for the exam data you could type

ci testscore, by(usedtapes)

Note that unlike in (a) I did not need to use **bysort**.

(c) To perform an hypothesis test about means in STATA you can use the **ttesti** command if you know the mean(s), sd(s) and sample size(s) or the **test** command if you have the data stored in a column. Note that STATA automatically gives p-values for the two-sided test and both one-sided tests so you do not have to specify which test you are doing or what the significance level is. Examples are as follows:

To perform the test about radiation levels from warm-up problem 2 using the column in the HW1 data file you would type

ttest height==1.

Note that you use a double equals sign and have to specify the null hypothesis value but everything else STATA does for you. To do the test using the mean and variance, $\bar{X} = 4, s = 3$, that we found above and the sample size of $n = 9$ you would type

ttesti 9 4 3 1

which gives first the sample size, then the sample mean, then the sample standard deviation and finally the value to be tested.

To perform the test comparing the exam scores of students who did and did not use the statistics tapes in warm-up problem 3, the corresponding commands would be

ttest testscore, by(usedtapes)

Note the comma after the main test statement and the use of the by option with parentheses to specify the grouping variable. If instead you had the means and standard deviations you would type

ttesti n1 m1 sd1 n2 m2 sd2

Try doing this using the output of the summarize command above.

(d) Power and sample size calculations in STATA use the command **sampsi**. By default this command produces the sample size needed for the specified means and standard deviations. If you use the option **power** and also specify sample sizes it will give you the power for the indicated test. STATA defaults to a two-sample comparison of means where the null hypothesis is no difference, the desired power is .8 and the significance level is $\alpha = .05$. You can change all of these using appropriate options. To specify a one-sample test of a particular mean value you use the **onesample** option. To compute power for a one-sided test you use the **onesided** option. Examples of the basic syntax are shown below but there are many shortcuts. You can search for help on sampsi to see the variations and some examples.

sampsi m1 m2, sd1(value1) sd2(value2) power(p) alpha(a)

calculates the power for a test of a difference in two means where the true (alternative hypothesis) means for groups 1 and 2 are m1 and m2, their standard deviations are value1 and value2, the desired power is p and the desired significance level is a. For instance, for a situation where the values are $\mu_1 = 0, \mu_2 = .5, \sigma_1 = \sigma_2 = 1$ with 80% power and $\alpha = .01$ we would type

sampsi 0 .5, sd1(1) sd2(1) power(.8) alpha(.01)

Note that this assumes we want the sizes of the two samples to be equal. You can add an option r() which gives the ratio of the sizes of group 2 to group 1 if you wish to get other than equal sized groups. (This is relevant for one of the warm-up problems.)

To specify the sample sizes and get the power we would instead type

sampsi m1 m2, n1(val1) n2(val2) sd1(sdvalue1) sd2(value2) alpha(a)

For a one-sample test where mn is the null hypothesis value and ma is the alternative hypothesis value of the mean to get the sample size you would type

sampsi mn ma, sd(value) power(p) alpha(a) onesample

If instead you wanted to calculate the power given the sample size you would type

sampsi mn ma, sd(value) n(value) alpha(a) onesample

(2) SAS COMMANDS:

Below are the SAS procs and command structure needed to do this assignment. I have typed the data statements as if you loaded the data in from an excel file into the work directory as the file work.hw1. If you loaded directly from a SAS file, just specify the appropriate directory and file name.

(a) Summary statistics are obtaiend with the **proc univariate**. If you just want to summarize a single variable, say the radiation variable for warm-up problem 2, you would type

```
proc univariate data = work.hw1;
var radiation;
run;
```

Note that you need the procedure name, a specification of the data set, the variable you want to summarize (preceded by the **var** option and the run command. Each piece of the procedure is separateed by a semicolon. If you want to get summaries for subgroups based on a second variable, for instance for the exam scores in warm-up problem 3 split by whether students had or hadn't used the tapes, you would add a **class** statement:

```
proc univariate data = work.hw1;
```

```
class usedtapes;
var testscore;
run;
```

(b) To get confidence intervals and tests in SAS you use **proc ttest**. The structure is very similar to that for the descriptive statistics except that for a one-sample test you have to specify the null hypothesis value using the option **ho=**. The default value is **h0 = 0**. (Actually you can do this for a two-sample test as well but there you are usually interested in checking whether the groups are equal so the default is OK.) You can also specify the significance level for the confidence intervals using the option **alpha =** . The default if you don't specify is to give 95% confidence intervals. The corresponding syntax for warmup problems 2 and 3 are as follows:

```
proc ttest data=work.hw1 h0 = 1 alpha = .01;
var radiation;
run;

proc ttest data = work.hw1;
class usedtapes;
var testscore;
run;
```

Note that all this seems very efficient. However SAS gives you lots of extra information in the printouts so be sure you can identify the correct values! Also, SAS only gives p-values for the 2-sided test. Of course, you can work out the 1-sided p-values by dividing by 2 and (if necessary) using complements.

(c) Power and sample size calculations in SAS are done using the Analyst feature. From the main menu select **Solutions/Analysis/Analyst.** Once you are in Analyst, select **Statistics/SampleSize** followed in our case by 1- or 2-sample t test. You will be given a pop-up grid to fill in that lets you specify whether you want power or sample size and asks you to fill in the assorted means, standard deviations, etc. The basic SAS set up assumes equal sized groups. Remember that if you want to calculate power for an effect size you set all the standard deviations to 1 and make the reference mean 0 and the other mean equal to the effect size.