Solutions To Homework Assignment 1

General Comments:

- The solutions given below are (quite a bit) more extensive than would have been necessary to get full credit. I use the answer key as an opportunity to make important points, or mention commonly made mistakes. Nonetheless, the answer key should give you an idea of the type of solutions I would like to receive. Solutions to the turn in problems will be added on October 3rd after the assignments are handed in.
- To keep the solutions from getting too long and tedious I have given the complete STATA output and then just the accompanying commands for the SAS output. If you are doing this in SAS and are not sure your output is correct feel free to check with me.

Warmup Problems

(1) Bad News Burgers:

(a) This is one of those confusing situations where you could argue the hypotheses either way. I have asked you to give a reason why it might make sense for Hamburger Heaven's null hypothesis to be that the meat is contaminated. Suppose I am the owner of the chain and it has just been announced in the news that people are getting sick at my restaurant. What will happen? I will lose lots of business UNLESS I CAN PROVE TO PROSPECTIVE CUSTOMERS THAT THE REMAINING MEAT IS SAFE. The thing I want to prove is my alternative–namely that the meat is not contaminated. My null hypothesis must be that the meat is contaminated. Another way of saying this is that I prefer to play it safe and not risk serving the meat until I am sure it is OK–I assume the worst and try to prove the best. (Of course, if I were a consumer advocate, trying to shut the restaurant down I might need to prove there was the contamination, thereby reversing the hypotheses.)

(b) A type I error occurs if you reject the null hypothesis when the null hypothesis is true. In this problem the null hypothesis is that the meat is contaminated. Therefore a type I error would consist of deciding that the hamburger is not contaminated when in fact it is. A type II error occurs if you fail to reject the null hypothesis when the null hypothesis is false. In this example you would make a type II error if you decided that the hamburger was contaminated when in fact it was not. The probability of a type I error is α and the probability of a type II error is β. The power is the probability of rejecting the null hypothesis when the null hypothesis is false. Here that means the probability of deciding the meat is safe when it is in fact safe–that is the likelihood that once the owner has cleaned up the problem he will be able to prove it to the satisfaction of his customers. Mathematically this number is $1 - \beta$. In this problem it is much more serious to make a type I error than a type II error. If we make a type I error, contaminated burgers will be sold, many more people will get sick, and Hamburger Heaven will be in big trouble. If we make a type II error we will only lose a bit of money by failing to use some hamburger that was actually OK. Therefore we want to make α as small as possible.

(2) Nuclear Test Scare:

(a) Often the hardest thing in an hypothesis testing problem is correctly identifying the null and alternative. Here, the public health services manager is concerned that some citizens are in danger–and if they

are, he needs to prove it so that the nuclear testing will be stopped! The important hypothesis is that the people downwind are getting too much radiation. Thus, this is the alternative hypothesis. The alternative is the thing that is important or requires action, or that you want to prove. The null hypothesis is that the testing is safe and the downwind residents are not getting any more exposure than their counterparts. (Note that if you were the agency performing the nuclear experiment, you might want to prove that it was perfectly safe. In that case, the hypotheses would be reversed.) In symbols, we have $H_0 : \mu = 1$ and $H_A : \mu > 1$. (Note that I wrote the null hypothesis as equality rather than as less than or equal to. This is because I didnt feel there was any reason to believe the people downwind could be better off in terms of radiation exposure. However, it makes no difference to the calculations which way you write the null hypothesis.)

(b) Since the sample is small it is not too hard to do the calculations by hand. Recall that the best estimate of the population mean is the sample mean which is given by

$$\bar{X} = \frac{10 + 2 + 2 + 2 + 2 + 5 + 5 + 7 + 1}{9} = \frac{36}{9} = 4$$

Our best guess is that the mean level of radiation in the bone marrow is 4 picocuries for people living downwind from the plant–four times the level that was supposed to be safe. This sounds bad! However our sample is very small so we need to be sure this wasn't just due to bad luck. The variability in the sample will help us to get a better handle on this. To estimate the standard deviation we first compute the sample variance which is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(10-4)^2 + (2-4)^2 + \cdots + (1-4)^2}{8} = \frac{72}{8} = 9$$

To get the sample standard deviation we take the square root and get $s = \sqrt{9} = 3$. Of course it would be much easier to do this in STATA, using the **summarize** command, or in SAS using **proc univariate**. The commands and the corresponding printouts are given below:

```
In STATA:

summarize radiation

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   radiation |         9           4           3          1         10

In SAS the command would be:

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

(c) We are asked for a 98% confidence interval. Since all we have is sample data we must use the t distribution, not the Z distribution. (For those of you who like to be technical, this really assumes the original population distribution of radiation levels is normal since $n = 9$ is NOT large enough for the Central Limit Theorem to imply the sample mean is normal unless the original population is normal, but we will not worry about this here.) The basic confidence interval formula for a single mean is

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

For a 98% CI $\alpha = .02$, so $\alpha/2 = .01$. Since $n = 9$, our t distribution has $n - 1 = 8$ degrees of freedom. (When you are working with a one-sample situation, the degrees of freedom for the t-distribution is always $n - 1$. This is the degrees of freedom corresponding to the sample standard deviation. It is $n - 1$ because one degree of freedom is used up calculating the sample mean which must be used to compute the SD. If you would like a refresher on degrees of freedom I'll be happy to discuss it in office hours.) Looking in the t table we see that $t_{.01,8} = 2.896$. Therefore our confidence interval is $4 \pm (2.896)(3/\sqrt{9})$ or $[1.104, 6.896]$. Alternatively you could do this in STATA using the **ci** or **cii** commands or in SAS as part of proc ttest. The STATA commands and output are shown below. In SAS you would get the CI as part of the printout for the hypothesis test. The command is shown in part (d).

```
In STATA either

(i)  ci radiation, level(98)

    Variable |        Obs        Mean    Std. Err.       [98% Conf. Interval]
-------------+--------------------------------------------------------------
   radiation |          9           4           1        1.103541    6.896459

or

(ii)  cii 9 4 3

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
             |          9           4           1        1.693996    6.306004
```

Note that using the "immediate" version of the command we first write the sample size, then the sample mean and finally the sample standard deviation. Naturally the results of the two commands are the same. The interpretation of this interval is that we are 98% sure the range 1.69 to 6.30 picocuries contains the true mean bone marrow radiation level for people living downwind from the plant. Since this whole interval is above the population mean of 1 picocurie things sound bad!

(d) We are asked to find the chance of seeing a more extreme value than what we observed assuming the null hypothesis from part (a) is true. This is just the p-value for the hypothesis test in disguise! In this case, large positive values of $\bar{X}$, that is high radiation levels in the downwind sample, favor the alternative hypothesis. We therefore need the probability of getting an average radiation reading of 4 or **higher** if the population mean level of radiation is really $\mu = 1$ for people living downwind of the plant. We get

$$P(\bar{X} \geq 4|\mu - 1) = P(\frac{\bar{X} - \mu}{s/\sqrt{n}} \geq \frac{4-1}{3/\sqrt{9}}) = P(t_8 \geq 3) = .0085$$

Note that we can not get the exact p-value from the t-table though we can deduce from the table that it is less than .01 since our test statistical value $t_{obs} = 3$ is greater than $t_{.01,8} = 2.896$. I got the exact value from the STATA/SAS printouts of the hypothesis test shown below:

```
IN STATA

. ttest radiation==1

One-sample t test
```

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
radiat~n |       9           4           1           3    1.693996    6.306004
------------------------------------------------------------------------------
    mean = mean(radiation)                                        t =   3.0000
Ho: mean = 1                                     degrees of freedom =        8

    Ha: mean < 1                 Ha: mean != 1                 Ha: mean > 1
 Pr(T < t) = 0.9915         Pr(|T| > |t|) = 0.0171          Pr(T > t) = 0.0085

. ttesti 9 4 3 1

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9           4           1           3    1.693996    6.306004
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   3.0000
Ho: mean = 1                                     degrees of freedom =        8

    Ha: mean < 1                 Ha: mean != 1                 Ha: mean > 1
 Pr(T < t) = 0.9915         Pr(|T| > |t|) = 0.0171          Pr(T > t) = 0.0085

IN SAS the command is

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

(e) Based on my confidence interval from part (c) I am inclined to reject the null hypothesis. The 98% CI I computed lies entirely above the null hypothesis mean of $\mu = 1$. Since there is only a 2% chance over all that the interval does not contain the true μ, and in fact only a 1% chance that the interval lies entirely above the true μ, I am more than 99% sure that I will not make a mistake by rejecting the null hypothesis. The data are not at all consistent with the idea that the people living downwind have normal radiation levels! I draw the same conclusion from my calculation in (d). This probability of getting radiation readings this high by chance is VERY small-certainly less than 1%–if the people downwind are unaffected. But I did get radiation readings this high. Therefore I REJECT the null hypothesis that $\mu = 1$ and conclude the alternative that $\mu > 1$. It appears at least "99% certain" that the people downwind from the site are being exposed to additional radiation, and something should be done to address the problem. (Note: The probability in (d) means that IF the downwind people were just like everyone else, there is virtually no chance they would have shown such a high level of radiation in their bone marrow. It is not actually the probability that the null hypothesis is true! Since we did observe an average radiation level of 4, we suspect that it is our null hypothesis assumption was wrong.)

(f) For a test of a single mean, the estimated effect size is just the difference between the observed mean and the null hypothesis mean, divided by the standard deviation:

$$d = \frac{\bar{X} - \mu_0}{s} = \frac{4-1}{3} = 1$$

In this case, our observed mean is a full standard deviation above the null hypothesis mean which is by the conventions of Cohen a very large effect size.

(3) Teaching Aids:

(a) and (b) The population of interest to Dr. Goss is all statistics students, or at least all statistics students at this campus who might ever be in a class that could use the tapes. We can deduce this because Dr. Goss stated objective is to determine "whether all statistics students would benefit by using the tapes." He is not just interested in knowing whether his own students benefitted but whether ANY student has or will in the future benefit. Therefore, Dr. Goss class cannot be the population. Instead, it constitutes a sample: the students in the data set are a subset of all statistics students at the school, present and future. With the given information it is hard to tell whether the class is a random sample or not. If this class is an honors class, that might influence the exam scores and whether the students checked out the tapes, in which case the sample would not be random. However if there is nothing special (with regard to testing ability) about the students in this class then the class would be a random sample for our purposes. Either answer would be acceptable provided you explained what assumptions you were making and gave your reasoning.

(c) The central problem of this case is one of statistical inference because Dr. Goss wants to decide (or infer) based on some sample data, whether the tapes are useful for all students (the population). When you draw a conclusion or make a decision about a population based on sample data you are performing statistical inference. Almost every problem we do in this class will involve statistical inference.

(d) There are two variables measured for each student–their final exam score and whether or not they used the tapes. The final exam score is a quantitative variable because it is numerical. The variable concerning checking out the tapes is qualitative because it is not a number. It simply says whether the students fell into one of two categories "checked out tapes" or "didnt check out tapes." This variable is in fact what is called *nominal* because there is no particular ordering of the categories.

(e) This study is observational because Dr. Goss has not controlled who checked out the tapes. He has simply observed which students checked out the tapes and then looked to see whether those who did had higher exam scores. The fact that this was an an observational study may make it difficult for Dr. Goss to draw conclusions about the effectiveness of the tapes. For instance, it is possible that the students who checked out the tapes were the more serious students and would have done better on the exams anyway. To be sure the tapes were helping Dr. Goss would need to do an experimental study in which he randomly chose which students would get to see the tapes and which students wouldnt. However, such a study would be hard to perform. It would not be fair to prevent students who wanted to see the tapes from doing so because it might put them at a disadvantage on the exam. Similarly, it would be hard to force students who werent interested in the tapes to watch them carefully. One way to get around this problem would be to consider two equivalent classes and make the tapes available to one class and not the other. You can probably think of a variety of other snags....

(f) The general form for a confidence interval for the difference in two means is

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, n_1+n_2-2} s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If we let Group 1 be the students who watched the tapes and Group 2 be the students who didn't we see that $n_1 = 11$ and $n_2 = 12$. To get the means and standard deviations we can use the **summarize** command in STATA or **proc univariate** in SAS. The command lines and output are shown below:

```
IN STATA

. bysort usedtapes: summarize testscore
```

```
-------------------------------------------------------------------------------
-> usedtapes = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   testscore |        12          67    21.99173         25         99

-------------------------------------------------------------------------------
-> usedtapes = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   testscore |        11    81.18182    9.568889         66         95

-------------------------------------------------------------------------------
```

```
IN SAS the command would be

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

The difference in means is $\bar{X}_1 - \bar{X}_2 = 81.2 - 67 = 14.2$, the pooled estimate of the standard deviation is

$$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-2)s_2^2}{n_1+n_2-2}} = \sqrt{frac10 * 9.57^2 + 11 * 21.99^2 21} = 17.23$$

The corresponding standard error is

$$s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 17.23\sqrt{\frac{1}{11}+\frac{1}{12}} = 7.1926$$

Since we are asked for a 90% confidence interval and we have $n_1 + n_2 - 2 = 21$ degrees of freedom our t-value is $t_{.05,21} = 1.721$. The resulting confidence interval is $14.2 \pm (1.721)(7.192) = [1.62, 26.38]$. Thus we are 90% sure that students who use the tapes score between 1.62 and 26.38% higher on Dr. Goss' exams. This makes the tapes look pretty good except for the problems noted with the study design above.

Of course we could also have computed the confidence interval in STATA or SAS. This is most easily obtained as part of the hypothesis tests and is therefore shown in part (g) below.

(g) Dr. Goss wants to see if there is a difference in performance between students who do and do not use the tapes. Since it is simply stated that he wants to check for a difference rather than trying to prove the tapes improve performance, we use a two-sided test. Our hypotheses are

$H_0 : \mu_1 = \mu_2$–the average test score is the same whether or not you watched the tapes.
$H_A : \mu_1 \neq \mu_2$–there is a difference in average test score between the two groups.

In a two-sided test the alternative is ALWAYS inequality–we will never be able to prove that the means are exactly equal since any data that supported that would also support values of the means that were extremely close to equal. In this case that there is a difference is what we want to show in any case so it makes sense as H_A. Using the values computed above, the test statistic is

$$t_{obs} = \frac{\bar{(}X_1 - \bar{X}_2) - 0}{s_{pooled}\sqrt{1/n_1 + 1/n_2}} = \frac{14.2}{7.192} = 1.97$$

Since this is a two-sided test, our p-value is $2P(t_{n_1+n_2-2} \geq t_{obs}) = 2P(t_{21} \geq 1.97) = .0620$. This p-value is less than $\alpha = .10$ so at this significance level we reject the null hypothesis and conclude that there is a difference ine xam score between students who do and do not use the tapes. (Looking at the data tells us the students who used the tapes did better but that is not the formal conclusion of the test.) The exact p-value can not be obtained from the t-table. I got it off the STATA and SAS printouts which are shown below.

IN STATA

```
ttest testscore, by(usedtapes) level(90)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      12          67    6.348467    21.99173    55.59888    78.40112
       1 |      11    81.18182    2.885129    9.568889    75.95263      86.411
---------+--------------------------------------------------------------------
combined |      23    73.78261    3.821593    18.32772    67.22038    80.34484
---------+--------------------------------------------------------------------
    diff |          -14.18182    7.192961               -26.55905   -1.804582
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.9716
Ho: diff = 0                                     degrees of freedom =       21

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0310         Pr(|T| > |t|) = 0.0620          Pr(T > t) = 0.9690

OR using just the means and standard deviations:

 ttesti 11 81.18 9.57  12 67 21.99, level(90)

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      11       81.18    2.885464        9.57    75.95021    86.40979
       y |      12          67    6.347966       21.99    55.59978    78.40022
---------+--------------------------------------------------------------------
combined |      23    73.78174    3.821354    18.32657    67.21992    80.34356
---------+--------------------------------------------------------------------
    diff |               14.18      7.1926                1.803385    26.55661
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   1.9715
Ho: diff = 0                                     degrees of freedom =       21

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
```

```
 Pr(T < t) = 0.9690         Pr(|T| > |t|) = 0.0620          Pr(T > t) = 0.0310
```

```
IN SAS the command would be

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

Note that in STATA I used the level option to get 90% CIs. Also note that in the first STATA printout the order of the groups was reversed which is why the confidence interval is negative. This is not a problem–it simply means people who did not use the tapes had LOWER scores than people who did use them which is completely equivalent.

(h) As noted above when calculating the confidence interval the difference in means between students who did and did not use the tapes was 14.2. The pooled estimate of the standard deviation was 17.23. Therefore the effect size for the difference between using and not using the tapes is estimated as

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s} = \frac{14.2}{17.23} = .82$$

This is a large effect size by the conventions of Cohen which we learned in class. Our best estimate is that use of the tapes is associated with close to a full standard deviation improvement in the test score but of course since this is an observational study we have no idea whether there is a causal effect of the tapes or it is just that the better students chose to watch them.

(4) Practicing Power: The commands and output for the power calculations are shown below with brief explanations.

(a) Here we are given power and want to calculate the necessary sample size. The STATA output is given below. We need about $n = 32$ subjects to achieve the desired power.

```
sampsi 0 5, sd(10) power(.80) alpha(.05) onesample

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
 alternative m =        5
            sd =       10

Estimated required sample size:

             n =       32
```

(b) To calculate sample size or power for a standardized effect size, you simply set the standard deviation to 1, the reference mean to 0 and the alternate mean to the effect size. The STATA command and output are

below. The necessary sample size is $n = 8$ which is much smaller than in part (a) where the effect size we were trying to detect was much smaller, $d = 5/10 = .5$. In general it is much easier to detect large deviations from the null hypothesis and so one can use a smaller sample size.

```
sampsi 0 1, sd(1) power(.80) alpha(.05) onesample

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
 alternative m =        1
            sd =        1

Estimated required sample size:

             n =        8
```

(c) Now we are given the sample size and asked to calculate the power. The STATA set up is essentially the same. For the specified parameters, our power is 78% which is very close to the usual standard of 80%.

```
sampsi 0 .5, sd(1) n(30) alpha(.05) onesample

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =       .5
            sd =        1
 sample size n =       30

Estimated power:

         power =   0.7819
```

(d) Now we want a sample size calculation for a two-sample t-test. The output is shown below. We need just over 60 subjects per group.

```
sampsi 0 .5, sd1(1) sd2(1) power(.80) alpha(.05)

Estimated sample size for two-sample comparison of means
```

```
Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =        0
            m2 =       .5
           sd1 =        1
           sd2 =        1
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =       63
            n2 =       63
```

(e) We use the same command as in part(d) except we have add the ratio option which tells us the relative size of group 1 and group 2. We need a total of approximately 144 subjects.

```
sampsi 0 .5, sd1(1) sd2(1) power(.80) alpha(.05) r(2)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =        0
            m2 =       .5
           sd1 =        1
           sd2 =        1
         n2/n1 =     2.00

Estimated required sample sizes:

            n1 =       48
            n2 =       96
```

(f) and (g) This is the same situation as parts (d) and (e) but now we are given sample sizes. The output is shown below. These sample sizes do not give adequate power (49% and 61% respectively) for the desired tests.

```
sampsi 0 .5, sd1(1) sd2(1) n1(30) n2(30) alpha(.05)

Estimated power for two-sample comparison of means
```

```
Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
            m1 =        0
            m2 =       .5
           sd1 =        1
           sd2 =        1
sample size n1 =       30
            n2 =       30
         n2/n1 =     1.00

Estimated power:

         power =   0.4907

sampsi 0 .5, sd1(1) sd2(1) n1(60) n2(30) alpha(.05)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
            m1 =        0
            m2 =       .5
           sd1 =        1
           sd2 =        1
sample size n1 =       60
            n2 =       30
         n2/n1 =     0.50

Estimated power:

         power =   0.6088
```