Statistical Reasoning in Public Health, 2009
Homework #2

1. Two investigators are interested in studying the health of men 18-24 years old in Baltimore City. Investigator A plans to take a random sample of 100 men age 18-24 from Baltimore City. Investigator B will take a random sample of 1,000 such men. Both investigators will measure the waist size of men in their samples. *(6 points total, 2 points for each of a –c)*

(a) Which investigator will tend to get a bigger standard deviation (SD) for the waist sizes of the men in his sample?
(b) Which investigator will tend to get a bigger estimated standard error of the sample mean waist size?
(c) Which investigator is likely to estimate a narrower confidence interval for the true mean waist size of all men 18-24 in Baltimore City?

2. A study is conducted concerning the blood pressure of 60 year old women with glaucoma. In the study a random sample of 200 60-year old women with glaucoma was taken and the sample mean systolic blood pressure was 140 mm Hg and the sample standard deviation was 25 mm Hg. *(14 points total, 2 points for each of a – g)*

(a) The sample standard deviation of 25 mm Hg is an estimate of what? (ie: this is an estimate of the variation in what values?)
(b) Based on the sample results, what is the estimated standard error of the sample mean?
(c) The quantity in part (b) is an estimate of what? (ie: this is an estimate of the variation in what values?)
(d) Using the sample results given, estimate a 95% confidence interval for the true mean blood pressure of all 60-year old women with glaucoma.
(e) Suppose this study was based on a random sample of only 100 women 60-years old with glaucoma, and the sample mean and standard deviation estimates were the same as those in the study of 200 such women. What is the estimated 95% confidence interval for the true mean blood pressure of all 60-year old women with glaucoma based on this sample of only 100 such women?
(f) Was your interval in (e) wider or narrower than the interval in (d)?
(g) Explain the intuition behind the mathematics that shows that the estimated standard error of a sample mean tends to decrease with increasing sample size.

3. Use the **Sampling Applet** which you can access from our web page or directly on the

internet to learn about sampling distributions and the Central Limit Theorem. The sampling applet is a little computer program (application) on the web that simulates data. See the applet instructions sheet for more information. (URL for applet: http://onlinestatbook.com/stat_sim/sampling_dist)

Choose your own distribution other than the normal for the population. Use the "custom" feature to draw your own distribution. Choose whatever you like but don't make it look normal. This will be your "population distribution" from which you will take multiple random samples to simulate the results promised by the Central Limit Theorem. *(18 points total, 2 points for each of (a) –(i) for the sampling scenarios based on n = 5)*

(a) Describe the distribution you chose. For example, is it left or right skewed? Is it bimodal?
(b) What is the mean and standard deviation of this population distribution?
(c) Simulate a study that consists of 5 observations (n=5) from your population distribution. (Use the "Animated" sample feature). Do this 3 times. How does the mean of these 3 means based on the 3 samples compare, in value, to the population mean?

Repeat the simulation another 1000 times, again each time taking ***5*** observations in each simulation. (now hit the "1,000" feature under "sample" to add another 1,000 sample means to the three sample means you already have). You will now have a histogram consisting of 1,003 sample means each based on random samples of size n = 5 all taken from the same population.

(d) What does the distribution of the 1,003 sample means look like, in particular does it look normally distributed?
(e) What is the mean of the 1,003 sample means?
(f) How does the sample mean of these 1,003 means compare, in value, to the population mean?
(g) What is the sample standard deviation of the 1003 sample means (this is an estimate of the standard error of the means based on samples of size n=5: remember standard error is shorthand for "standard deviation of sample means across all random samples of the same size from a single population")
(h) How does the theoretical value of the standard error for sample means based on samples of this size compare to the estimated standard error based on the 1,003 means in the simulation? (hint: I am talking about standard error here- how could you relate the population standard deviation of individual values and the size of the samples to estimate the standard error, ie variation in sample means?)

Now hit the "10,000" feature under "sample" to add another 10,000 sample means to the 1,003 sample means you already have) You will now have a histogram consisting of 11,003 sample means each based on random samples of size n = 5 all taken from the same population.

(i) What does the distribution of 11,003 sample means look like compared to what you found in part (iv) with 1,003 sample means? Does the "spread" (ie: standard deviation of the 11,003 means, ie: the estimated standard error) of the distribution change much? How about the mean of the 11,003 sample means?

4. Repeat parts a-i from problem 3 using n=25 observations (instead of n=5) in each sample in each simulation. (you may use a new starting population distribution)
(18 points total, 2 points for each of (a) –(i) for the sampling scenarios based on n = 25)

5. **Conduct a survey**

Conduct a survey with a sample size of at least 5 about a "continuous" variable. The survey can be of any topic of your choice. YOU MAY SURVEY PERSONS OTHER THAN SR1 CLASSMATES! Here are some ideas for a survey -Survey students about one of the following:

# hours slept yesterday	# sodas consumed yesterday
# Facebook friends	commuting time to school
age	# cell phones owned in lifetime

Answer the following questions. For questions that require some calculation, you can use a computer (if you know how), a calculator or do it by hand. (20 points total: 2 points for each of parts (a) – (j))

(a) What variable have you chosen for your study?
(b) What was your sample size?
(c) Report your data.
(d) Calculate the sample mean, median, standard deviation and standard error of the sample mean.
(e) What is the population from which your sample was drawn?
(f) Do you think your sample was representative of the population? Is there the possibility of any systematic bias resulting from how you chose your sample?
(g) Do you have suggestions for how your sampling procedure could be improved?

(h) Pretend your variable is approximately normally distributed at the population level - can you estimate an interval that would contain about 95% of the individual values in the population? Does this result make sense?

(i) Calculate a 95% confidence interval for the population mean. Briefly explain what this confidence interval means.

(j) Explain the difference between the intervals computed in parts *h* and *i*.

Sample Quiz Questions (Please note: these questions are a required part of the homework!) Choose the correct answer for the following multiple choice questions and give a sentence or two justifying your answer choice: *(6 points total: 1 point for providing a correct answer, 1 point for a correct justification)*

6. The individual survival times of a random sample of 500 patients with a particular disease following surgery is a right-skewed distribution. The sampling distribution of the sample means based on random samples of 500 observations from this population:

   (a) will be approximately normally distributed.
   (b) will be right-skewed
   (c) No general statement can be made
   (d) It is not possible to comment on the distribution of multiple sample means using the information from a single sample of individual values

7. A survey is performed to estimate the proportion of 18-year old females who have had a recent sexually transmitted infection (STI) defined as an STI in the past year. In a random sample of 300 women, 200 have agreed to participate. Which of the following best characterizes this remaining sub-sample of 200 women?

   (a) The sample of the 200 women will definitely be representative of the population of interest.
   (b) The sample of the 200 women will definitely not be representative of the population of interest.
   (c) No general statement can be made about this sub sample of 200 women without additional information.

8. A random sample of 300 diastolic blood pressure measurements are taken. Suppose a 99% confidence interval for the population mean diastolic blood pressure is 68 to 73 mm Hg. If a 95% confidence interval is also calculated, then

   (a) The 95% confidence interval will be wider than the 99%.
   (b) The 95% confidence interval will be narrower than the 99%.
   (c) 95% and 99% confidence interval will be the same.
   (d) One cannot make a general statement about whether the 95% confidence interval would be narrower, wider or the same as the 99%.

SR1 HW2, 2009.