Q SCI 381: Introduction to Probability and Statistics
S. Scherba, Jr. Assignment #4: Review Exercises

Shorter answers (*23 points*).

1. (*3 points*) In your own words please summarize the meaning of:
    a. (*1 point*) The Law of Large Numbers.
    b. (*1 point*) The Central Limit Theorem.
    c. (*1 point*) The theorem underlying the distribution of the sample mean and standard deviation.
2. (*1 point*) Why are estimates of population parameters (statistics) useless without a measure of their inherent error?
3. (*3 points*) What does the sampling distribution of a statistic, e.g., the sample mean, represent?
    a. (*1 point*)
    b. (*1 point*)
    c. (*1 point*)
4. (6 *points*) With regard to experimental design:
    a. (*2 points*) Define a treatment and give an example.
    b. (*2 points*) Distinguish between systematic and random errors.
    c. (*2 points*) Distinguish between an experimental and an observational study.
5. (*1 point*) Complete the following statement by John Tukey: "Numerical quantities focus on expected values, graphical summaries on …."
6. (*1 point*) In the phrase random sample, what does the word random mean?
7. (*4 points*) Suppose we wish to investigate the effect that different foods have on a species of fish. We place the food in the tanks containing the fish. We record the weight increase of each fish.
    a. (*1 point*) What is the response?
    b. (*1 point*) What is the experimental unit?
    c. (*1point*) How could we redesign the experiment to make the fish the experimental unit?
    d. (*1 point*) Was an actual treatment applied?

8. (*4 points*) Suppose we wish to conduct a gene expression experiment. We take the RNA from two groups of male subjects (those with a disease and those without it). We apply their RNA to the microarray chip (slide) that contains genetic material (genes). The RNA is hybridized to the microarray, and fluorescence is used to measure the gene expression. There are several types of arrays available and different manufacturers for each.

    a. (*2 points*) What might be two possible sources of error from confounding if we are not careful?

    b. (*1 point*) What is the experimental unit?

    c. (*1 point*) What is the treatment?

Problems (*17 points*).

9. (5 points) In 2010, the Physicians Foundation conducted a survey of physician's attitudes about health care reform, calling the report "a survey of 100,000 physicians." The survey was sent to 100,000 randomly selected physicians practicing in the United States, 40,000 via post office mail and 60,000 via email. A total of 2,379 completed surveys were received.

    a. (*3 points*) State carefully what population is sampled in this survey and what is the sample size. Can you draw conclusions from this study about all physicians practicing in the United States?

    b. (*1 point*) Given that the nonresponse rate is defined to be (1 – the percentage of responses), what is the nonresponse rate?

    c. (*1 point*) Why is it misleading to call the report "a survey of 100,000 physicians"?

10. (*12 points*) You measure the lengths (in cm) of 14 specimens of a certain insect and find that sample mean to be 22 cm and the sample standard deviation to be 1.2 cm. You may assume that the population is approximately normally distributed.

    a. (*10 points*) Find a 95% confidence interval for the population mean. **Show all of your work including clearly stating the distribution you assume and the numerical values of the corresponding critical values. Start by stating the confidence interval in general terms.**

    b. (*2 points*) Provide a one sentence description of what this confidence interval means in words. **Be explicit in your description.**

Problems from the text. You should not fail to do these, but they are <u>NOT TO BE TURNED IN FOR GRADING</u>. My solutions will be posted.

From F & P, please do, in good mathematical form the following problems: 10.4, 10.33, 10.35, 10.39, 10.45; 11.1, 11.6, 11.7, 11.9, 11.11, 11.15, 11.19, 11.22, 11.29, and 11.31.

Draw a picture to help you visualize the problem wherever possible, e.g., draw probability distributions to help yourself determine the area under the curve the problem is requiring you to determine. **Use the tables in the back of the text; not R**. We will be using R in the analysis section of this assignment. Most of these problems are very short.

In addition, the following will NOT GRADED, BUT PLEASE READ FOR YOUR PERSONAL ENRICHMENT!

An important central concept and two open-ended questions (with some answers).

Everything we have done and everything we will do in this course explicitly or implicitly assumes that our samples from the population are random. The subjects we are currently discussing, sampling, parameter estimation, and confidence intervals are the appropriate place to bring up a very important topic that is beyond the scope of our class by asking two questions.

[**1**] When might we have samples that are not random?

One area is the use of sampling along line transect in biology. Other areas are genetics and cladistics. For example:

W. F. Fagan, M. J. Fortin, and C. Soykan. (2003). Integrating Edge Detection and Dynamic Modelling in Quantitative Analyses of Ecological Boundaries. *Bioscience*. 53: 730 – 8.

D. P. Faith. (1991). Cladistic Permutation Tests for Monophyly and Nonmonophyly. *Systematic Zoology*. 40: 366 – 75.

J. Felsenstein. (1986). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*. 39: 783 – 91.

J. Felsenstein. (1992). Estimating Effective Population Size from Samples of Sequences: A Bootstrap Monte Carlo Integration Method. *Genetical Research, Cambridge* 60: 209 – 20.

There are many other areas.

[**2**] What do you do when your sample is not random?

To explore an answer to the last question, Google and / or do a Google Scholar search on Bradley Efron, resampling, jackknife, and the bootstrap. The implementation of Efron's groundbreaking work required the development of high-speed

microprocessors. The theorem that Efron stated and proved in the mid to late 1970's arguably represents the great revolution in 20th century statistics (the work of R. A. Fisher representing the great revolution in statistics in the first half of the 20th century). This material is now a central part of statistical research and analysis (including AI).

You may also wish to look at the book:

Bryan F. J. Manly. 2007 (3rd edition). **Randomization, Bootstrap and Monte Carlo Methods in Biology.** Chapman & Hall / CRC.

Two relevant UW courses are:

Q SCI / STAT 403 and BIOST 558.