Estimation

Overview
Up until now, we have been dealing with describing data (Chapter 2), assuming that we already know how data are distributed and that we sometimes know population parameters such as the mean and variance (Chapter 3 through 5). Examples 6.1 through 6.4 are typical of the situation where we assume that the number of patients with a particular characteristic or the measurements taken from a screening test follow a certain distribution.

A much more common situation is where data are collected in some systematic fashion for the purpose making an inference as to how the data are distributed and what underlying characteristics it exhibits. This is called statistical inference and is the subject of much of the remainder of this course. Statistical inference is broken down into two main components: estimation where we estimate the value of specific parameters of a distribution; hypothesis testing where we test whether the value of some specific parameter equals a specific value. Chapter 6 covers the topic of estimation.

Please note that while you should read all of Chapter 6 and understand the material in the section titled *Interval Estimation---Exact Methods* on pages 207-210 of Rosner, you will not be asked to calculate intervals using the methods described in that section. Calculation of exact confidence intervals is best done with software (e.g. in SAS using PROC FREQ and the BINOMIAL option), not with tables Also, you need not concern yourself with the details in section on *Random-Number Tables*, pages169-173. It is important to know about both random samples and random assignment of individuals to groups in a study. However, in your own work, you will most likely use software to accomplish these tasks.

Most of the material covered in this Chapter of Rosner is also covered in parts of Chapter 6 and most of Chapter 7 of the *Cartoon Guide*. Looking at the Cartoon Guide might help you understand the material in Rosner a bit better.

Learning Objectives
1. Understand what a simple random sample is.
2. Understand the relationship between a sample and a population.
3. Understand the design features of a randomized clinical trial.
4. Know how to calculate a point estimate of the mean of a distribution and the standard error of the mean.
5. Understand the implications of the Central Limit Theorem and why it enables us to use the normal distribution to perform statistical inference despite the non-normality of individual observations in a sample.
6. Know how to calculate confidence intervals for the mean using both the normal distribution and the t distribution.
7. Know how to calculate a point estimate of the variance of a distribution.
8. Know how to calculate confidence intervals for the variance using the chi-square distribution.
9. Know how to calculate a point estimate for the parameter p in a binomial distribution.
10. Understand how to use normal theory to obtain an interval estimate of the parameter p.
11. Know how to calculate a point estimate for the parameter lambda of a Poisson distribution and an exact confidence interval around this parameter.
12. Understand the difference between one-sided and two-sided confidence intervals.

Readings
Chapter 6:

Problems
6.7- 6.17 (skip 6.9), 6.31-6.34, 6.38- 6.39, 6.44- 6.46, 6.58-6.62, 6.88-6.91

What You Should Know
We now move onto to the topic of statistical inference. It involves looking at a sample of data and inferring something about the underlying distribution of the population from which this sample is drawn. Chapter 6 deals with the problem of estimating the value of specific population parameters from a sample of data. You need to understand what a simple random sample is (Definitions 6.1 and 6.2) because for the remainder of the course we will be assuming that all samples are simple random samples. Carefully read Examples 6.9, 6.10, and 6.11. Thy provide excellent examples of health related situations where one would take a sample to estimate a population parameter.

Section 6.4 of Rosner contains essential background information on randomized clinical trials. This is one of those topics that differentiates biostatistics from the more general topic of statistics. Please make sure you understand the topics contained in this section. It is important for you to be familiar with the terminology associated with randomized clinical trials. You will encounter many of theses terms (block randomization, stratification, double blind, single blind, etc) when reading journal articles and you should know what they mean.

We now start studying actual estimation procedures for some of the more common parameters: the mean and variance. Common sense seems to indicate that if you want to estimate the mean of a population, you might want to take a random sample from the population, calculate the mean of that sample and then use that sample mean as an estimate of the population mean. You need to know why this works. You need to know what is meant by an unbiased estimator (Definition 6.1) and what is meant by the minimum variance unbiased estimator of the population mean (discussed on page 171 of Rosner). Basically an estimator is unbiased if the average value of the estimator over a

large number of repeated values equals the population value you are trying to estimate. If you want to estimate the mean value of birth weight in a population and take a large number of samples from that population and the average value of the sample means actually equals the population mean, then that estimator (the sample birth weight) is unbiased. This seems to make sense. Now you need to understand what is meant by the standard error of the mean. As you read this material in Rosner, please note what the standard error of the mean is and how it differs from the standard deviation. Read Definition 6.12 carefully and remember that the standard error is NOT the standard deviation of an individual observation but rather of the sample mean.

One of the most important concepts in statistics is the Central Limit Theorem (Equation 6.3). This theorem allows the use of normal probability theory to solve many problems by confirming that when the sample size is large enough, the distribution of sample means will be approximately normal, even when the underlying distribution in the population is not normal.

Chapter 6 also covers using the estimates of a population mean and its variability to construct a confidence interval for the mean. Equation 6.4 shows how the standardized normal distribution can be used to calculate intervals when the value of the population variance is known. This is rarely the case and Equation 6.5 gives us the relationship that can be used to calculate interval estimates for the population mean using the t-distribution and the standard deviation of a sample. Because for large samples (greater than 200), there is virtually no difference between a normal distribution and a t-distribution, confidence intervals for the mean can be calculated using the normal distribution for large sample sizes. You should be very familiar with Equations 6.6 and 6.7 and the material in the Section entitled *t Distribution*. You will definitely be tested on this material.

Estimation of the variance of a distribution is also a topic in Chapter 6. If repeated random samples of size n are selected from a population, then the average of the variances from these samples over a large number of samples will be the population variance, i.e. the sample variance is an unbiased estimator of the population variance. Putting confidence intervals around the variance is not as easy as for the mean. The Central Limit Theorem is of no help. Equation 6.15 gives a method for calculating an interval estimate for the variance using the Chi-Square Distribution which works well when the samples are normally distributed. If they are not normally distributed, the use of Equation 6.15 is not recommended. You need to familiarize yourself with the Chi-Square Distribution and the methodology to construct interval estimates for the variance when the samples are normally distributed. Example 6.41 demonstrates the technique. As noted in this example, the calculated interval is not symmetric as was the case for intervals around the mean because the chi-square distribution is not symmetric.

Chapter 6 continues by discussing methods for estimating the parameter p from a binomial distribution. Equations 6.16. 6.17, 6.18 and 6.19 are very important concepts that need to be understood. In particular, you need to be able to compute confidence intervals for the binomial parameter p using the formula in Equation 6.19. Note the requirement for using Equation 6.19, the product of the sample size, the estimate of p and the estimate of q must be greater than or equal to 5. You also need to understand how one goes about calculating confidence intervals for the parameters of the Poisson distribution. Make sure you know how to use Table 8 in the Appendix to calculate the confidence intervals for the expected number of events over time period t. Finally, familiarize yourself with the material contained in Section 6.10 so you understand the difference between two-sided and one-sided confidence intervals.

Problems

6.7- 6.17 (skip 6.9)
These problems all relate to the data presented in Table 6.9 on page 218. They should provide you with practice in computing confidence intervals. Since these problems are for a rather limited number of data points, you can solve them using just a calculator.

6.31-6.34, 6.38- 6.39, 6.44- 6.46, 6.88-6.91
You should be able to answer all of these questions using just a calculator. They provide you with good practice in creating confidence intervals and interpreting results (note: 6.91 is OPTIONAL).

6.58-6.62
Though the title of this problem set might be scary (SIMULATION), the problems give you practice taking random samples.