Descriptive Statistics
Classes 1 - 3 (8/28 - 9/4)

Overview
You should read chapters 1 and 2 in the Rosner text and complete selected problems (or more if you like) at the end of chapter 2. The problems require you to: compute descriptive statistics; create graphical displays of data; comment on various aspects of the data based on the computed statistics and graphics. You will notice that there is a lot of emphasis in chapter 2 on how to compute statistics and how to draw graphics. You should know the formulas that allow computation of descriptive statistics. However, given the availability of statistical software, it is very rare to compute such statistics by hand or with a calculator. The same is true for the production of statistical graphics. You should know when a specific type of graphical display (histogram, box plot, stem-and-leaf plot, x-y scatterplot) is appropriate, but again, it is very rare to draw such graphics by hand.

StatCrunch will be used to compute descriptive statistics and to create graphics. The emphasis will be not on computation, but on interpretation of numeric and graphical results

Chapters 1 and 2 in the Cartoon Guide to Statistics are similar to those in Rosner. No matter how basic the material, the presentation in the Cartoon Guide really helps to understand the text in Rosner (or for that matter, any conventional statistics or biostatistics text). You are NOT REQUIRED to read the material in the Cartoon Guide. However, it will help to understand that material you are assigned in the Rosner text.

NOTE: You are welcome to try more problems in Rosner than those that are assigned. If you do extra problems and have questions, you can ask questions about them in class or by e-mail. We can also go over them in class.

Learning Objectives
1. Understand measures of the central tendency of data distributions: mean, median, mode.
2 Understand measures of variation of data distributions: range, quantiles, standard deviation (variance), coefficient of variation.
3. Learn appropriate methods for displaying the distribution of data.

Readings
Chapters 1 and 2: 8/30

Problems
2.1- 2.3, 2.12- 2.18, 2.31- 2.32, 2.38- 2.40: 9/4

What You Should Know
Chapter 2 begins by discussing measures of location. There are several statistics that are described that define the center or middle of a sample of data. The most important of these is the arithmetic mean, often referred to in common language as the average. You need to know how to calculate the arithmetic mean (Definition 2.1). While there are several other measures of location which define the middle or center of a sample, they are less often utilized. They are, however, important. The median (Definition 2.2) and the mode (Definition 2.3) are two such measures. The relationship between the arithmetic mean and the median should be understood. How these two measures compare when a distribution is symmetric, positively skewed or negatively skewed is an important concept to understand.

When we have a sample of data and add a constant to each data point we have translated the sample. Equation 2.1 tells us what happens to the mean of this new translated sample. Essentially, if we add a constant to each data point of our original sample, the mean of the translated sample is the sum of the mean of the original sample and the constant. Likewise, if we multiply each data point of a sample by a constant, the mean of the new sample is the product of the original mean and the constant (Equation 2.2). When we do this multiplication of each data point by a constant we have rescaled the original sample. Make sure you completely understand the material covered in Section 2.3 (Some Properties of the Arithmetic Mean).

In addition to summarizing a sample by calculating measures of location, we may also be interested in the variability or spread of the data set. We can usually pretty well describe a sample of data by calculating both a measure of location and a measure of spread. Several measures of spread are defined in this Chapter which are important. You should be familiar with how to calculate the range of a sample (Definition 2.5) and percentiles (Definition 2.6). However, the most important measures of spread or variability are the variance and standard deviation. You must understand how these two statistics are calculated (Definitions 2.7 and 2.8) even though they will almost always be produced using a software package. Example 2.19 is a good practice exercise that you should completely understand before moving on.

Just as we did with the arithmetic mean, we can see what happens to the variance and standard deviation of a sample when we add a constant to each data point or multiply each data point by a constant. Adding a constant to each data point has no impact on the spread of the data (this should be intuitive) and thus the variance or standard deviation is unchanged. This relationship is given is Equation 2.5. Multiplying each data point by a constant, does, however, impact the variance and standard deviation. This is shown in Equation 2.6. You should also understand

how to compute the Coefficient of Variation as defined in Definition 2.9. The Coefficient of Variation is useful in comparing the variability of several samples which have different arithmetic means because, as a general rule, when a sample has a higher arithmetic mean it tends to have higher variability.

Finally, you should make sure you understand how to construct a bar graph, a histogram, a stem-and-leaf plot and a box plot. These graphical presentations of data are almost always available by using software such as SAS and you will probably never have to draw one of these graphs by hand. Nonetheless, please know what they show and how they are generally constructed.

Problems

Problems 2.1- 2.3

These are good problems in which to practice manual calculation of descriptive statistics. The data set is small and it would be easy to compute the answers to problems 2.1 and 2.2 with a calculator. You could then compare the results of your calculations with those computed by StatCrunch, Excel, or any other software you have available. Though we do not expect you to compute statistics manually, manual calculation of a few statistics can assist in understanding their meaning and use.

HOSPITAL data are also available on the course web site as a SAS data set and an Excel spreadsheet.

Problems 2.12- 2.18

Try this problem MANUALLY.

The data are available on the course web site in both a SAS data set and an Excel spreadsheet (named CARDIO) and it will be used in class to show you how do the problem using SAS.

Problem 2.31-2.32

This set of problems requires the use of a SAS data set or Excel spread sheet (named LEAD).

In both, the values for the variable GROUP are: 1=control; 2=exposed.
In both, the values for the variable GENDER are: 1=male; 2=female.

Problems 2.38-2.40

This set of problems requires the use of a SAS data set or Excel spread sheet (named BONEDEN). These data are explained in section 2.10 of the text. Variables have been added to the original data found on the disk based on the formulas shown in problem 2.38:

LS_C : value of C for LUMBAR SPINE bone density - the answer for problem 2.38
PDGROUP: (difference in pack-years of smoking between the twin pairs in five groups, heavier smoking twin 2 - lighter smoking twin 1) - groups are based on the information in problem 2.38