Chapter 6 Overview

Introduction:
This chapter begins our study of statistics. As we stated at the beginning of the course, probability and statistics are two different sides of the same coin. They are, respectively, the pre-game and post-game analysis of randomly occurring events.

When we discussed continuous random variables, we said that certain types of distributions can model different types of random variables. For example, we said that the exponential distribution could be used to model the lifespan of some machine. But how do we know what parameter to use, and how do we even know that this is an accurate model? If we wanted to think about this in the real world, what we would do is to gather information on lifespans of many of these machines to look for trends. This process of gathering data, analyzing it, and making inferences based on it is the core of the subject of statistics.

In this chapter, we look at the basics of gathering data, presenting data usefully, and computing some simple statistics.

6.1 Experimentation

It's important to keep the picture in Figure 6.1 on page 268 in your mind throughout the rest of the course:

{unknown probability distribution $f(x)$}

↓ (experimentation)

{sample data set}

↓ (statistical inference)

{estimate properties of $f(x)$}

To help us, we make some definitions.

Population and Samples

A **population** consists of all possible observations available from a particular probability distribution. A **sample** is a particular subset of the population that an experimenter measures. Samples may be used to investigate the unknown probability distribution. A **random sample** is one in which the elements of the sample are chosen at random from the population, and should be representative of the population.

The data observations mentioned above can be of several different types:

1) Categorical or nominal data is generally non-numerical in nature and groups the observations into non-overlapping categories. If a Categorical variable has only two possible observations it is called a **binary** variable.

   Examples:

   a. Observation - color of a car {white, gray, red, blue, . . .}. The statistician must decide beforehand whether silver and gray are the same color or different, etc.

   b. Observation – Zip code. This is still categorical because even though a zip code is a number, the number is meaningless.

   c. Observation – On or off. This categorical variable is a binary variable.

2) Numerical data may be either real numbers or integers. The book calls these variables **continuous**. (The use of this word here is a bit non-standard, and may mean something slightly different in other textbooks).

   Examples:

   a. In boxes of computer chips, count the number of chips that are defective

   b. Measure the weights of newborn baby hippos

   c. Measure the lifespan of several copies of a machine

In all these cases, we should remember the importance of the sample we get being random, which means we are getting enough elements of the population in the sample, and that the elements we get are representative of the whole population, not just a certain part. Often, this means we have to think about what the population we want is. For example, to get a sample of boxes of computer chips, do we want the population to be all the boxes produced in all factories? Only one factory? Only a single assembly line in one factory? If we want the population to be all the boxes in all factories, but we only choose boxes from a single assembly line in a single factory, then our sample will not be representative.

End of 6.1. For the exercises, note that they require you to access some datasets. These datasets can be found at http://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111827045&token= by clicking on “Datasets” on the left. You can download them as a zip file, which contains all the datasets for the rest of the course, in a variety of formats.

6.2 Data Presentation

We study a few ways to present data to make it easier to analyze. For categorical data, we use bar charts, Pareto charts, and pie charts. For numerical data, we focus on histograms. In 6.3, we'll also consider boxplots.

Categorical Data:

Bar chart: Each category has a bar whose height is proportional to the frequency associated with that category. Never truncate the vertical axis! Always start at 0!

Pareto chart: A bar chart in which the categories are arranged in order of decreasing frequency. As an option a line is drawn showing the cumulative frequency over the bars.

Pie chart: A pie chart uses the proportion of the total data set in each category to divide the circle into slices that cover the same proportion of the circle. If there are n observations of which r are in a specific category then the slice of pie for that category should have a central angle of $\frac{r}{n} \bullet 360°$.

Numerical Data:

Histograms look like bar graphs, but since they are used to represent numerical data, the x-axis will have numerical values instead of categories.

A histogram is a graph that displays quantitative data by using contiguous vertical bars of various heights to represent the frequencies of the classes (if the frequency of the class is zero the "height" is zero). The width represents a quantitative not qualitative variable (piston rod lengths rather than quality of a television picture).

Example: Consider the following data, which gives corinne levels in tobacco users.

1	2	131	173	265	210	44	277	173	208
32	3	35	112	477	289	227	103	1	284
222	149	313	491	130	234	164	198	86	48
17	253	87	121	266	290	123	167	245	250

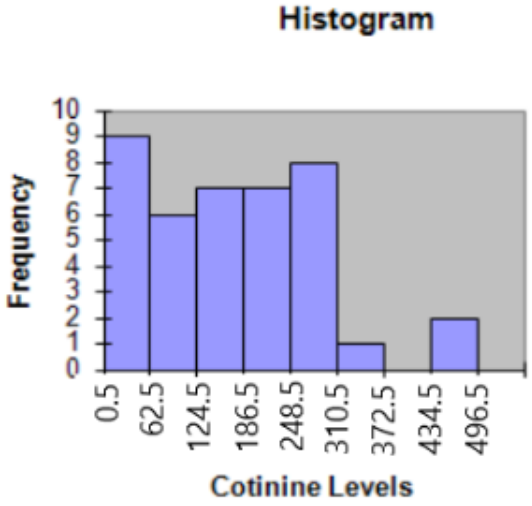
From this data, we can make a table separating the numbers into bands. In determining the width of the bands (bandwidth), we have some considerations. Making the bands too wide will result in fewer "classes" of data, which will make the histogram rougher, in that it will not distinguish

as cleanly between different data points. Making the bands too narrow will make the histogram look more like the raw data, which makes it more difficult to detect trends.

In this case, we'll split into 8 bands of equal width. Since the range of the data is from 1 to 491, our bandwidth should be about (491-1)/8 = 61.25. We always round up, to make sure all the data falls in one of the 8 bands. Thus, we'll use bandwidth of 62. As a rule of thumb, the boundaries between bands should occur between data points, though this is not always necessary. We can make the following table:

Bands (classes)	Frequency
(0.5 , 62.5]	9
(62.5 , 124.5]	6
(124.5 , 186.5]	7
(186.5 , 248.5]	7
(248.5 , 310.5]	8
(310.5 , 372.5]	1
(372.5 , 434.5]	0
(434.5 , 496.5]	2

Then we can form the following histogram:



Ideally, the histogram gives a rough idea of what the probability density function of the underlying distribution should look like. In later chapters, we'll try to answer the question of how close in shape the histogram is to the actual probability density function.

We can try to gain some knowledge about the underlying distribution based on the shape of the histogram. Is the histogram symmetric? If not, is the tail on one side longer and flatter than the other? Is there a single peak where the frequency of data is highest? Are there two peaks? Are there data points on the tails that seem out of place and don't fit with the rest? We have definitions for these situations.
Unimodal: One peak.
Bimodal: two peaks.
Symmetric: The book says symmetric when they really mean symmetric and unimodal (for example, bell-shaped)
Right-skewed (positively skewed): right-hand tail longer and flatter than left-hand tail
Left-skewed (negatively skewed): left-hand tail longer and flatter than right-hand tail
Outliers: Unusually large or small pieces of data. One must consider whether or not they can be ignored.

6.3 Sample Statistics

From a data set, there are some basic statistics that we define.

Sample Mean
The sample mean of a data set is the sum of all the data values divided by the number of values in the sample. This is just the arithmetic average.

The sample mean is denoted by $\bar{x}$ and is given by the formula
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ where n represents the total number of observations.

The mean is affected by outliers (unusual pieces of data that don't fit the trend). One extremely small or large data point can have a disproportionately large effect on the sample mean. However, when we want to estimate the expectation *E(X)* of the unknown underlying probability distribution, one place to start is the sample mean.

Sample Median
The median is the midpoint of the data array.

To find the median
1) Arrange the data in order.
2) Select the middle point. If the middle point falls between two data values add them and divide by two.

The sample median may be used to estimate the population median. If data is significantly skewed, then the median may be a better estimate of the "average" than the mean.

Sample Trimmed Mean

A trimmed mean is obtained from a data set by eliminating and equal number of highest and lowest values. Usually this is 10% off the top and bottom, so that if you had 60 observations you would remove the lowest 6 and the highest 6. This does not change the sample median, but it gets rid of any outliers and reduces the affect of any skewing of the data on the sample mean. That skewing may be important, so some thought should be given before employing this method. The value of a Sample Trimmed Mean is generally somewhere between the values of the sample mean and the sample median.

Sample Mode

The sample mode of a set of discrete or categorical data is the data value that occurs the most. It is thus the category or value with the highest probability.

Sample Variance

The sample variance of a data set $x_1, \ldots, x_n$ is given by the formula $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$.

Note that we divide by n – 1 instead of n. This is a correction factor that is only meaningful for small values of n, and corresponds to the fact that samples usually vary less than populations. This is explained more in Chapter 7. Similar to the alternate computational formula for the variance of a random variable, we are given a similar computational formula for the variance of a sample.

$$s^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n \bullet \bar{x}^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n-1}$$

Sample Quantiles

The pth sample quantile (sample percentile) is a value that has a proportion p of the sample taking values smaller than it and a proportion $1 - p$ taking values larger than it. Sample quantiles are usually values between data observations.

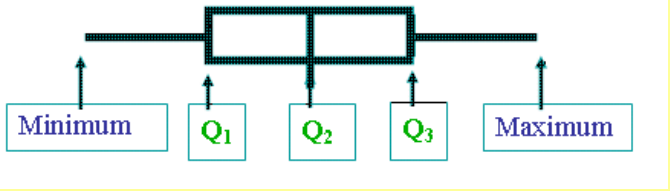
The sample median is the 50th percentile of the sample, the upper quartile is the 75th percentile and the lower quartile is the 25th percentile. As before the interquartile range is the difference between the upper quartile and the lower quartile. The book doesn't really give thorough details on how to do this precisely, so we can use the following methodology for finding quartiles:

1. If there are an even number of data observations: Use the median to divide the ordered data set into two halves. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

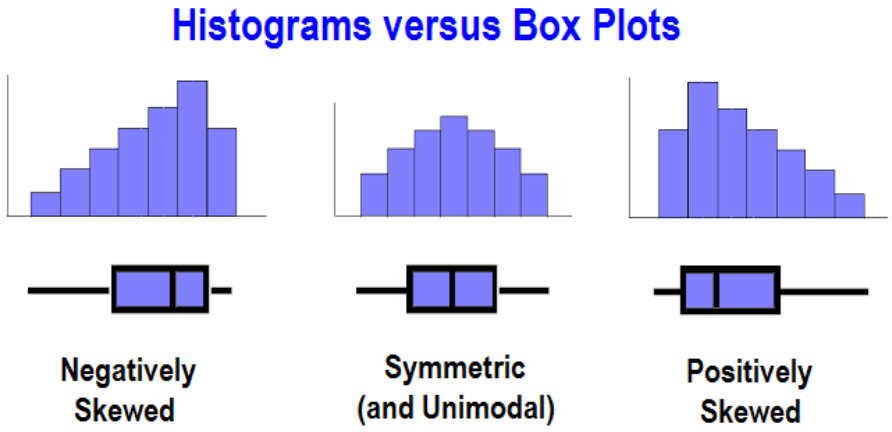
2. If there are an odd number of data points then there are three methods that are sometimes used. The third method is the average of the first two and is generally thought to be best, it is the mean of the first two methods. First use the median to divide the ordered data set into two halves. Now find the median of the lower half two ways, once including the median in the lower half and once excluding it. The lower quartile is the arithmetic mean of these two numbers. Now use the same process on the upper half of the data to find the upper quartile.

Box Plots

A box plot is a graph of a data set obtained by drawing a rectangle with sides at the lower quartile Q_1 and the upper quartile Q_3 and two horizontal lines from the sides of the box to the maximum and minimum data value. Finally, a vertical line is drawn through the middle of the box at the median (Q_2).



Note: The "box" represents half of the data value



Coefficient of Variation

The coefficient of variation (CV) is the standard deviation divided by the mean. $CV = \frac{s}{\bar{x}}$. Note that we can also define this for random variables: $CV = \frac{\sigma}{\mu}$.

The CV measures the variation relative to the average.