Chapter 2: Describing, Exploring & Comparing Data

Section	Title	Notes Pages
1	Overview	2
2	Frequency Distributions	3 – 4
3	Histograms	9 – 12
4	Statistical Graphs	19 – 20
5	Critical Thinking: Bad Graphs	21

§2.1 Overview

There are 2 types of statistics: **descriptive** and **inferential**
It is the descriptive statistic that we will be discussing in this chapter. A *descriptive statistic* describes the characteristics of data.

Following is a list of the 5 important characteristics:

1) The **center** of the data which a representative value that indicates where the middle of the data lies. The most common descriptive statistic for measuring center is the average.
2) The **variation** or scatter of the data is also important.
3) The **distribution** of the data tells the shape of the data. We are very familiar and will become even more so with a distribution call the Normal – this is a bell shaped curve, the type of distribution that class grades follow.

4) **Outliers** which are data points that lie outside the "normal" variation of the data.
5) The change in characteristics of data over **time**.

Just like memorizing order of operations (Please Excuse My Dear Aunt Sally, PEMDAS) or the order of the planets in our solar system (Matilda Visits Every Monday Just Stays Until Noon Period) or the names of the Great Lakes (HOMES) , there is a mnemonic device to help you remember these important characteristics of data:

Computer **V**iruses **D**estroy **O**r **T**erminate (CVDOT)

It is my belief as a statistics instructor that you can not understand a statistic until you have learned how to calculate that statistic by hand, so even though we will be using technology to calculate many statistics ***I will still require you to be able to find basic descriptive statistics by hand*** for small data sets.

§2.2 Frequency Distributions

To help us in our study of the shape of data (**distribution**) we will be learning how to compile a **frequency distribution** (table), a **relative frequency distribution** and a **cumulative frequency distribution**. A frequency table will be created which will give categories (not like the categories for ordinal data) and the number of experimental units (data points) within each category. From these we will learn how to draw pictures that show us distributions. Let's go through an example to learn the process. Here is a summary of the process to start:

Creating a Frequency Distribution (Table)

1. Decide upon the # of **classes** (the categories). There should be between 5 and 20 classes, but if many have 1 or 0 data points then you have chosen too many. I will normally tell you how many classes to use.
2. Find the **class width** (the range of data points in each class). This is the
   (Maximum – Minimum)/ # of classes *

* You may need to round, usually up

3. Select a lower limit for the 1st class. This point does not have to be the lowest data point, but it should make sense in terms of the data (no negatives if negatives don't exist, etc.). **To this lower limit add the class width.** This is the lower limit of the next class, **not** the upper limit of the first class!!!! Continue until you've gotten all your classes.
4. Since the last class has an upper limit that you have not found, be sure to fill it in, as well as the other upper limits.
5. **Make your counts** as to the number of data points in each class.

Example: The ages of people arrested for purse snatching are:
16, 41,25,21,30, 17, 29, 50, 30, & 39

Using 4 classes create a frequency distribution (table) for the data.

Classes:
Class Width:

Lower Limits:

Place Upper Limits:
Make counts:

Some places where students go wrong and that you should be cautious of are:

1) Are your classes mutually exclusive?
2) Did you include all classes even if one was zero?
3) Does the sum of frequencies add to the number of data points? **This is really important!**

Next, we need to discuss the **relative frequency distribution** which is simply our frequency distribution listed in terms of percent of the whole. To create a *relative frequency distribution* you first create a frequency distribution and then divide each class frequency by the total number of data points (sum of all frequencies).

Rel. Freq. = Class Freq. ÷ Sum of Freq.

Example: For the above data of purse-snatchers create the relative frequency distribution.

Finally, we need to know how to create a **cumulative frequency distribution**. This is just a running total that includes all classes below, but not any above. In a cumulative frequency distribution the cumulative frequency in the last class must be the total number of data points. We write the classes in a little different manner for a cumulative frequency dist. – we use less than and use the lower limit of the next class.

Example: For the purse-snatchers create a cumulative frequency dist.

Now for the real work, what does this all mean? Well, this is just a preview of how the data is behaving. We are learning where the data stacks up, what percentage lies in certain areas and at what point have we seen most of the data. A frequency distribution is really just the first step in seeing the shape of the data, and we will be using it to draw an actual pictorial representation in the next section. But, at this early stage we can begin by **comparing frequency distributions** to other similar data or data over time and **looking for patterns** – and really *looking for patterns and finding out how mathematically significant* those patterns are is what statistics is all about!

§2.3 Histograms

This section concerns itself with looking at pictures that show the shape of the data. We will talk about histograms, which are types of graphs created from frequency & relative frequency dist. We will also discuss line graphs of frequency (freq. polygon) and cumulative frequency dist. (ogives). We then move to another visual called a stem and leaf plot and its relative the dot plot. Finally we will discuss a scatter diagram, but will leave the other visuals such as the Pareto Chart and Pie Chart and Time Series to the book.

Let's begin with the **histogram**. A *histogram* is constructed from a freq. or rel. freq. dist. The horizontal axis consists of the scale of the data with either the midpoint of the class (called class marks or midpoints) labeled or the class boundaries (the midpoint between upper and lower limit) marked. The vertical scale is the frequency. Each class is shown with a vertical bar. There are no gaps between the bars on a histogram! Everything must be clearly marked!!

Creating a Histogram

1. Create a frequency or relative frequency table
2. Find the class boundaries (or the class marks)
3. Create an axis system with the class boundaries on the horizontal axis & frequencies on the vertical axis (label clearly)
4. Create a bar the width of the class and the height of the frequency to represent each class in the table.

Example: Create a Histogram for the purse-snatcher data.

What does this do for us? Well, it shows you the shape of the data! We can see that as age increases the number of people snatching purses goes down and that in-between age we have the highest frequency. We see that it is a skewed data set; it is not normally distributed (that bell-shaped curve that we see in grades).

If you create a **relative frequency histogram** you will not see any difference in shape. You are seeing the exact same trends; it is just represented in terms of percentages instead of raw numbers.

When we talk about a **frequency polygon** we are talking about the points that correspond to the class marks (mid-point of the class) and the frequency of each class. From these points we create a line graph. This line graph can be shown super-imposed on our histogram. The line graph is nice for showing trends, because our eye follows the slope of the line. As a result of seeing the slope we see increases and decreases according to class.

Example: Draw the frequency polygon over the histogram above.

§2.4 Statistical Graphs

An **ogive** is a line plot for a cumulative freq. These are constructed starting with the class boundaries and the cumulative frequencies. The major use would be to see the number of values below a particular point in the data. It can also help us to see an overall trend in data – where we are getting the most increase and where things level off. With some data this is important (see the example in the book) and with others it is not as useful.

A **stem-and-leaf plot** gives us a nice visualization of the data shape, but unlike the histogram, allows us to keep the original data. This is what makes it superior to a histogram – there is no loss of the original data.

This type of plot works with a **stem**, which is based upon the tens, hundreds, etc. and the **leaf**, which completes the number. You can think of this process like writing a number in expanded form! For large data sets it is sometimes necessary to give each stem two representations to spread the data out a little more (this does add some level of error to the process and thus misrepresentation of the data is possible), called an **expanded** stem-and-leaf. We can also bring the data together a little more for smaller data sets creating a **condensed** stem-and-leaf (the same problem with misrepresentation can occur; see p. 70 for an example).

Creating a Stem-and-Leaf Plot

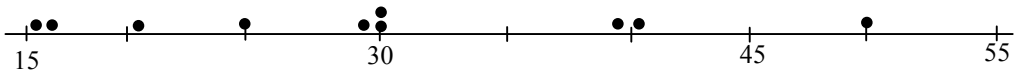
1. Look over the data and decide upon stem and leaf (all below 100, use a stem of 10's)
2. Write a table with stems on left and leaves on right (label the stem & leaf unit value)
3. Fill in the leaves to represent all the data points

Example: Create a stem and leaf plot for the purse-snatchers data
16, 41, 25, 21, 30, 17, 29, 50, 30 & 39

Now, if you look at this on its side, you will see the nearly the same shape as the histogram, but with the added benefit of still being able to see the original data. Another advantage of the stem-and-leaf is that it *sorts the data*, which prepares us for finding some other important descriptive statistics that require ordered or ranked data (such as medians, quartiles & percentiles).

Another type of visual representation is the **dot plot**. The dot plot also keeps all the original data, sorts it, but is worse at showing shape than the histogram and dot plot. The dot plot is good at showing outliers, however.

Use a number line representative of the range of data values and place a dot for each point just above its representative number on the number line to construct a dot plot. For our purse-snatching data, a dot plot is not too informative! This is what it looks like.



The final type of visual that I want to discuss in class is called a scatter diagram or scatter plot. This is very easy to construct and shows us trends that exist in the data. We will use this type of plot for our excursion into regression in chapter 9. This type of visual can only be used for a data set that has ordered pairs (two characteristics that are somehow related, collected from the same people or at the same time& place, etc.) You'll be happy to know that I'm going to need a new data set to show this! All we have to do is construct ordered pairs based upon paired data and plot those points on a coordinate system using appropriately labeled axes for the independent and dependent variables.

Example: The following data is the score of a math reasoning test and the starting salary of the person scoring thusly on the test. Create a scatter diagram based upon it.

X = score	78	85	92	100	85
Y = salary	89	93	99	100	84

Scatter plots can be used for noticing a relationship between two sets of data but we do not want to make the mistake of causality. Just because you receive a high score on the above test does not guarantee a high salary. There is a relationship (called a correlation), but this does not mean that a high salary is caused by a great score, it could be that a person had a lot of prior experience in the job and received a high salary for this reason. This will be discussed at length in Chapter 9.

This chapter also includes discussions of the Pareto chart, pie charts and time-series graphs. The Pareto chart is a bar graph for qualitative data where the bars are ordered in descending height to help tell the story of what class (category) is most important. A pie chart is the same type that you are familiar with from your everyday life and uses the Relative Frequency Distribution to show which classes (categories) are the largest or most influential. The time-series graphs are useful for seeing trends over time.

This section and the preceding are very helpful in showing how to visualize data (CVDOT) without showing time progression. We can easily see the **center** of the data, the **variation**, the **distribution** via a histogram and with the dot plot or stem and leaf we can see **outliers** very easily. We can not see **time** trends with these techniques however!

§2.5 Critical Thinking: Bad Graphs

I've already said my piece about the contents of this section. I want you to read the section and do the exercises for review. In summary, the following are some of the topics that I have discussed that you will find elaboration upon.

Graphs & Pictographs can be used in misleading ways.

Graph Ex: Showing only a portion of a graph exaggerating the difference between the things being shown. See p. 71

Pictograph Ex: Pictures that don't represent the intended difference. Such as using a cube with double side length to represent 2 times the amount. See p. 72.