MATH 124 Spring 2005

Lecture: 3
Date: Feb 10, 2005

Skills you should acquire from this lecture:
- Knowledge of graphical ways to examine distributions
- Ability to draw and interpret a Stem and Leaf plot
- Ability to draw and interpret a Histogram
- Knowledge that more than one variable may be used to measure the same thing, each variable having its own advantages and disadvantages

Related readings in the textbook:
- Section 1.1
- Problems 1.5-1.9, 1.28a

What is a distribution?

In statistics we use the word *distribution* to refer to information which tells us which values a variable takes and how often it takes these values. When we spend time exploring data we are often trying to get a feeling for how the data is distributed.

Graphing categorical variables

You may often see *bar graphs* and *pie charts* in newspapers and magazines as these are nice ways to show the distribution of a categorical variable, in a few lectures we will talk about how sometimes these tools are used to mislead or distort the data. However, they are not particularly useful for data analysis purposes because we can usually see the distribution by just looking at the data.

Stem and Leaf plot

The stem and leaf plot is useful for getting a quick picture of the shape of the distribution while also including the actual numerical values into the graph. It is typically most useful when you have a small number of data values. Each observation is separated into two parts, a *stem* and a *leaf*. Stems consist of all but the least significant (most right) digit. The leafs contain only a single digit. The stem and leaf plot is created by first writing the stems in a vertical column and then the leafs in increasing order to the

right of the stems. Usually a vertical line is drawn to separate the stems from the leafs.

An example

Suppose we have the following data:

```
-0.9 -2.0  0.1  0.1  0.6  0.4  0.4  1.1  0.4 -0.5  0.6
 1.0 -2.6  0.7 -0.6  0.7  1.6  0.7  1.4  0.9  2.0  1.6
 1.1  2.3  1.8
```

then stem and leaf plot will be

```
-2 | 60
-1 |
-0 | 965
 0 | 11444667779
 1 | 0114668
 2 | 03
```

The stems can also be split, This is particularly useful with large datasets. We split the stems so that the leaves 0, 1, 2, 3, 4 fall on one stem and 5, 6, 7, 8, 9 fall on the other stem. With observed variables with many digits you might find it easier to round the numbers before making the stem and leaf plot.

Histograms

Histograms are more versatile than stem and leaf plots. A histogram breaks the range of values of a variable into intervals and displays only counts of observations that fall into each interval. This makes them much easier to use when there are a large number of observations. The intervals for a histogram should be of equal length.

Steps to drawing a histogram

1. Divide the range of the data into equal length classes

2. Count the number of individuals/units/observations in each class. A table of these values is called a frequency table.

3. Draw the histogram. The height of each bar represents the number of observations in the class. The width represents the length of the class.
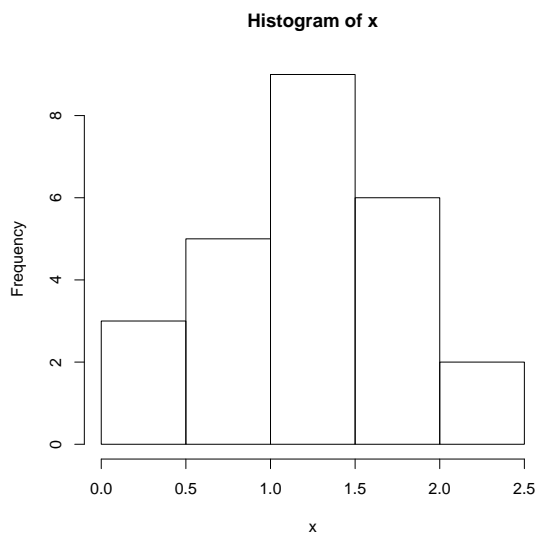
An Example

The data

```
0.75 2.04 1.99 1.59 0.84 1.38 0.68 1.22 1.31 0.43
1.55 1.97 1.76 1.05 1.19 0.78 0.21 1.25 1.03 1.23
0.68 2.39 1.81 0.33 1.05
```

A frequency table using class intervals of length 0.5

```
Class                Freq
0.00 < x <= 0.5      3
0.50 < x <= 1.00     5
1.00 < x <= 1.50     9
1.50 < x <= 2.00     6
2.00 < x <= 2.50     2
```

The resultant histogram

Histogram of x

Results of Exercise

Draw the stem and leaf plot for the same data, but this time split the stems. Draw the same histogram but this time use class intervals of length 0.25.

Drawing the Stem and Leaf plot with Stems split

```
-2 | 6
-2 | 0
-1 |
-1 |
-0 | 965
-0 |
 0 | 11444
 0 | 667779
 1 | 0114
 1 | 668
 2 | 03
```

Notice how the plot is more spread out and -2.6 and -2.0 stand out much clearer as being outliers.

Histogram with different length class intervals.

```
Class                Freq
0.00 < x <= 0.25     1
0.25 < x <= 0.50     2
0.50 < x <= 0.75     3
0.75 < x <= 1.00     2
1.00 < x <= 1.25     6
1.25 < x <= 1.50     3
1.50 < x <= 1.75     2
1.75 < x <= 2.00     4
2.00 < x <= 2.25     1
2.25 < x <= 2.50     1
```

The resultant histogram

Histogram of x