MATH 3070 Introduction to Probability and Statistics
Lecture notes
Graphical Representation

Objectives:

1. Identify bar graphs, pie graphs, and histograms
2. Interpret a histogram

Graphical Display of Data

It is easiest to interpret and absorb data when it is presented in a graphical format. There are several different ways to display data depending upon the purpose.

A **bar graph** or **bar chart** is a simple graphical technique for illustrating a categorical data set. (Hayter) The displayed value is either the count or percent of items/individuals who fall in a particular category. Each category has a bar whose length is proportional to the frequency associated with that category. The bars may be either vertical or horizontal.

A **pie chart** divides a complete circle into slices, one corresponding to each category with the central angle of the slice proportional to the category relative frequency. A pie chart emphasizes the proportion of the total data set that is taken up by each of the categories. If a data set of n observations has r observations in a specific category, then that category receives a slice of the pie with an angle of $\frac{r}{n} \times 360$.

A **dot plot** consists of a horizontal scale on which dots are placed to show the numerical values of the dat point. If a data value repeats, the dots are piled up at that location, one dot for each repetition. (Cryer and Miller)

For small datasets, a **stem and leaf plot** is another way to display data. In a stem and leaf plot, the "stem" and "leaf" are parts of the data points. The leaf is usually the rightmost digit, in the one's place. The stem is the remaining digits. This isn't a hard and fast rule, you can make the stem the leftmost digit, if you prefer. The rightmost is more common.

List the stems in a column, smallest to largest. Then list each leaf value, to the right of the stem, in increasing order. If there are two values with the same value, list them together.

If you rotate the stem and leaf plot on its side, you will get the same type of bar graph as with a histogram.

Histograms

A **histogram** is similar to a bar chart, but the values represented are numerical rather than categorical. Whereas in a bar chart the values on the horizontal, or x-axis, list the various categories in a histogram the x-axis is numerical. A histogram consists of a number of bands whose length is proportional to the number of data observations that take a value within that band. It is important to carefully construct the width of the bands in a histogram. (Hayter)

Frequency histograms are the most basic histograms. They are graphical displays of frequency distributions. All they require is grouping the data into classes, counting the number of observations in each class, and making a plot.

The width of the bands is related to the number of bands. A suggested method for choosing how many bands is

Less than 25 observations	5 or 6 bands
25 to 50 observations	7 to 14
More than 50	15 to 20

There are about 5 steps to constructing a histogram (Mendenhall).

What are we looking for in a histogram? The shape of the data, basically, but also a few other characteristics.

- Overall pattern
- Striking, noticeable deviations from that pattern
- Its shape, center, and spread of the data
- Any observation that is significantly apart, an outlier

An important question about the shape is, "Is it symmetric?" If the left and right halves of the distribution are approximately mirror images, then the data are said to be symmetric. If the data aren't symmetric, then we talk about the skewness of the data.

- Right, positively, skewed right-hand tail is longer and flatter
- Left, negatively, skewed left-hand tail is longer and flatter

It might appear that there isn't a single, central point to the data. There might be two (or more). If there are two separate peaks, or humps, in the data then we say it is bimodal. What does this mean? It may be that the data are actually from two different groups or populations. For example, a data set measuring some attribute of people may more usefully be separated into one data set for men and one for women.

An outlier is a data point that appears to be separate from the rest of the data set. What do we do with an outlier? It may be a misrecorded value, or a typo. It could be a valid data point that is just that extreme. (Bill Gates is an outlier in a sample of CEO salaries, but he's not an error.)