STAT 515

Statistical Methods I

Sections 2.1-2.3

Graphical Displays and Measures of Center

Brian Habing
Department of Statistics
University of South Carolina

Outline

- The main idea of descriptive statistics
- Graphical displays for qualitative variables
- Basic graphical displays for quantitative variables
- Shapes of distributions
- Mean, Median, and Mode

Chapter 2 – Descriptive Statistics

Once we've made sure to collect data that should contain the answers to our research questions, the next step is to find out what the answers are.

This can often be very daunting, consider that many of the large data sets used in research are huge – matrices with thousands or tens-of-thousands of rows (one for each person studied) and dozens or hundred of columns (one for each variable measured). It isn't even feasible to get all of the values of one variable on a single page to look at.

Chapter 2 – Descriptive Statistics

The goal of descriptive statistics is to use either graphical or numerical summaries of the data that are both

- Succinct (the reader can easily take them in and understand them)

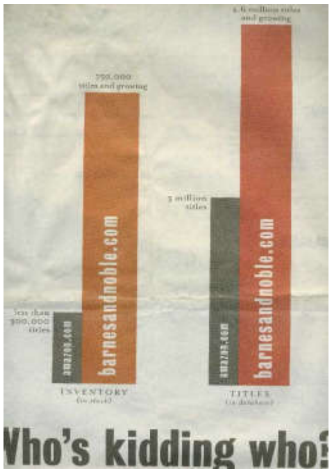
And

- Not Misleading (they don’t leave out important pieces of information or lead the reader to make false conclusions)

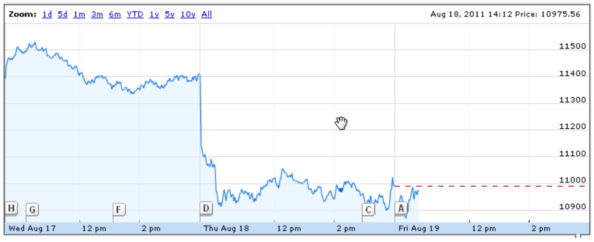
Chapter 2 – Graphical Displays

While being succinct is often easy, avoiding misleading the reader can be much more difficult. In the case of graphical displays, it involves:

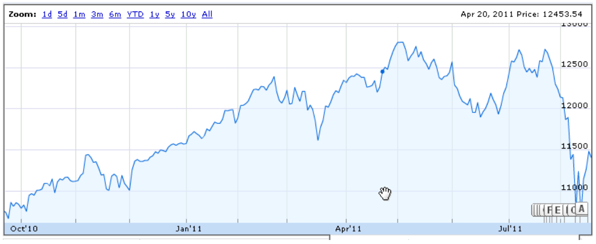
- Choosing appropriate titles, captions, and labels
- Not abusing human spatial perception
- Not using the methodology to hide statistical differences.



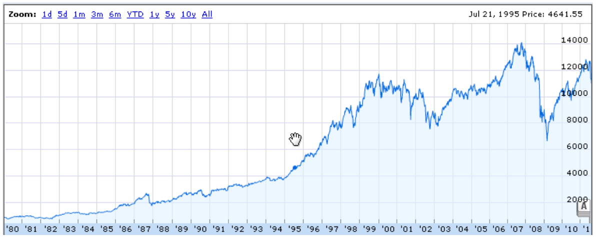
Dow Jones Industrial Average

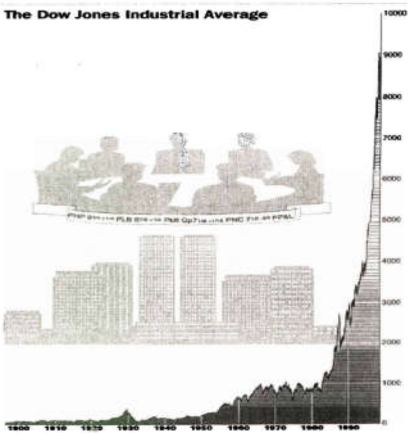


Dow Jones Industrial Average



Dow Jones Industrial Average





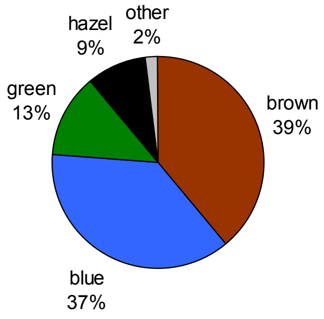
Eye Color of 46 Students

		Relative
Eye Color	Frequency	Frequency
Brown	18	39%
Blue	17	37%
Green	6	13%
Hazel	4	9%
Other	1	2%
Total	46	100%

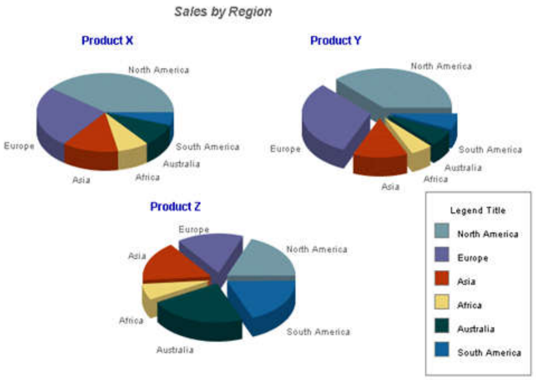
Frequency = Area

The most common graphical displays are based on calculating the relative frequency (percentage) that different groups occur, and then giving those groups areas based on those relative frequencies.

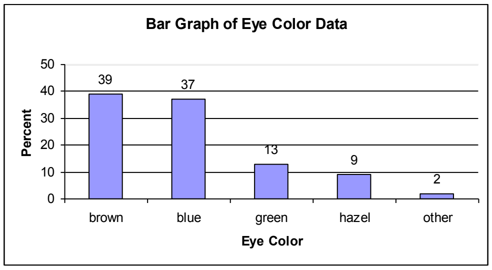
A pie chart is one common method of presenting such data… it's generally not the best choice…



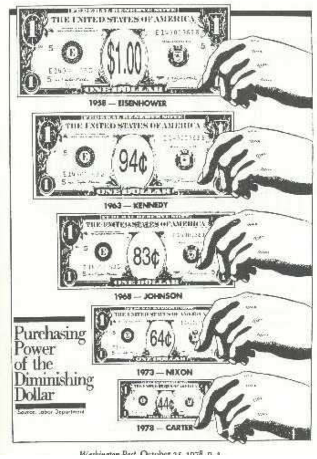
… and is often made even worse.



A <u>bar graph</u> (or bar chart) is another common, and better, choice…



Pictograms use images besides bars… and are often made incorrectly.



Washington Post, October 25, 1978, p. A

Categorical versus Quantitative

Eye color above was an example of categorical data, there was no need to have the eye colors in a certain order.

When the data is quantitative (e.g. numeric) like age or income then a bar chart with the bars in order is commonly used, with the bars touching… and is called a histogram.

Classes

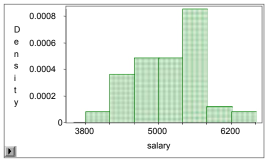
In many cases, quantitative data can take a large number of different values (consider annual salary), and the possible values are then broken into classes before calculating the frequency table.

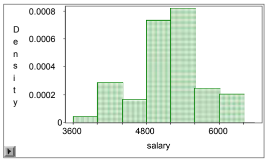
For income, perhaps $0-$19,999.99, $20,000-$39,999.99, etc…

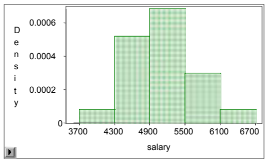
Which Classes

Section 2.2 gives some examples of constructing histograms and some general guidance on choosing the intervals. Most computer programs will have some default way of doing this.

Unfortunately histograms are very easy to manipulate…







Coming Later

Because of this, the box-plot (which is covered in section 2.8 in the fourth introductory video) and the quantile-quantile-plot (in section 5.4) are often much more reliable ways of getting a feeling about what a data set really looks like.

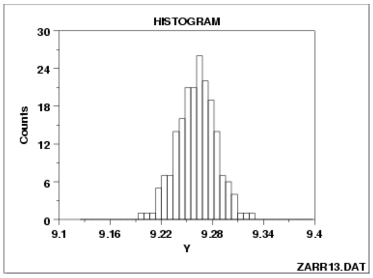
Shape, Center, and Spread

A data set is often described in terms of its shape, its center, and its spread, with shape being determined by looking at a graphical display.

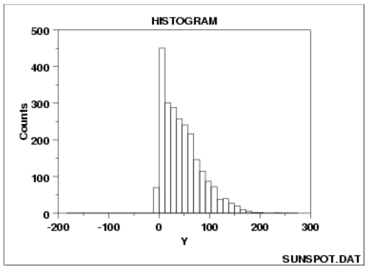
Common descriptions of shape include:

Approximately symmetric – the right and left sides of the histogram are approximately mirror images

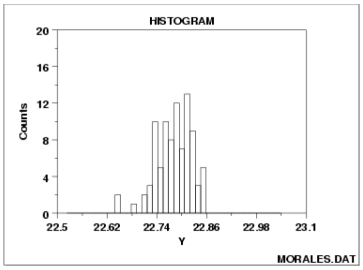
unimodal – the data has one major peak



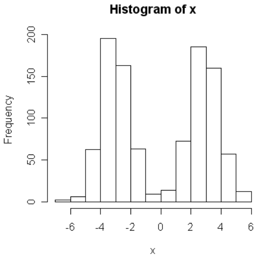
skewed right – the right side of the histogram (the half with the larger values) extends much farther than the left



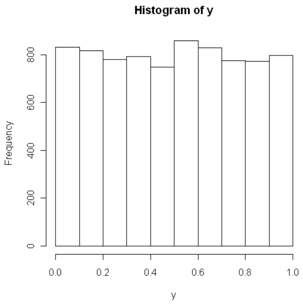
skewed left – the left side of the histogram (the half containing the smaller values) extends much farther than the right



bimodal – the data set contains two major peaks



uniform – all values are roughly equally likely



Measures of Center

A measure of center is often informally called an “average” – but is often taken to mean “what value is typical”.

Three common measures of center are the mean, median, and mode.

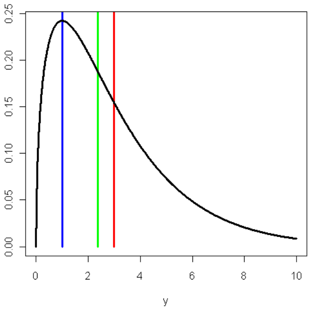
(Section 2.4 gives examples of calculating these measures of center.)

The mean of a sample is found by adding up all of the values and dividing by the sample size.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + \cdots + y_n}{n}$$

Consider 1, 8, 3

The red line in the graph to the right is the <u>mean</u>.



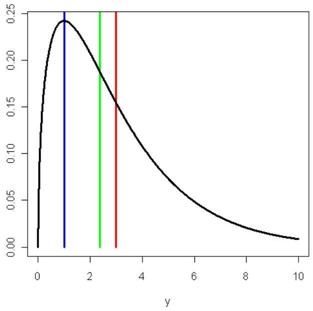
The sample median is the middle of the sorted values in a data set.

Consider 1, 8, 3

Consider 1, 8, 3, 4

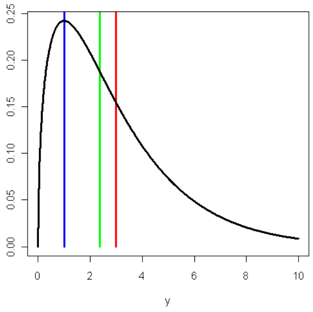
The green line in the graph to the right is the <u>median</u>.

It is the value that has at least 50% of the observations greater than it and at least 50% of the observations less than it.



The blue line in the graph to the right is the <u>mode</u> – the most commonly occurring value.

It is the one that is least used as a statistic because there is often more than one, and it is difficult to develop theory for it.



Which to Use?

Mean and Median both have their own strengths and flaws.

Median – It is not affected by extreme values – if 49% of the observations are doubled, this “measure of center” doesn’t change at all.

Mean – It is very affected by outliers – adding one very large value to a data set can greatly change the mean, even though all of the data points but one are kept the same.

Example Revisited

Consider the data set 1, 3, 8 again

Which To Use?

Data is approximately symmetric and unimodal – use the mean (which in this case should be very close to the median)

Data is skewed – use the median

The mode is typically the odd measure out, but it occurs a great deal in mathematical statistics classes! There are also other (more complicated but possibly better) measures of center in more advanced classes (like STAT 518).