Math 161

Glen Pugh

September 27, 2007

Our Work so Far

Our Work so Far

- Focus to date: Collect information on individuals of a population through either observational studies or experiments.
- The information collected is in the form of values (numerical or descriptive) assigned to variables representing characteristics of an individual.
- In the case of experiments, response variables record the changes in the characteristics of subjects due to the experiment. The explanatory variables are thought to explain the outcomes recorded by the response variables.

The Next Step: Organizing Data

- Graphing
- Numerical Descriptions

Chapter 10: Graphs

Graphs in a Nutshell

- Use graphs to visualize data, i.e. use pictures to represent the values a variable takes and the frequencies at which the values occur.
- More precisely, we wish to graphically represent the distribution of a variable: what values the variable takes and how often it takes these values.

Graphs and Data Types

- The type of graph used depends on the data type of the variable.
- Quantitative variables take on numerical values.
- Categorical variables take on descriptive or qualitative values.
- Example:

$$\begin{aligned} \text{individual} &: \text{person} \\ \text{a numerical variable} &: \text{age, values} = 18, 19, 20, \ldots \\ \text{a categorical variable} &: \text{sex, values} = \text{male, female} \end{aligned}$$

Bar Graphs

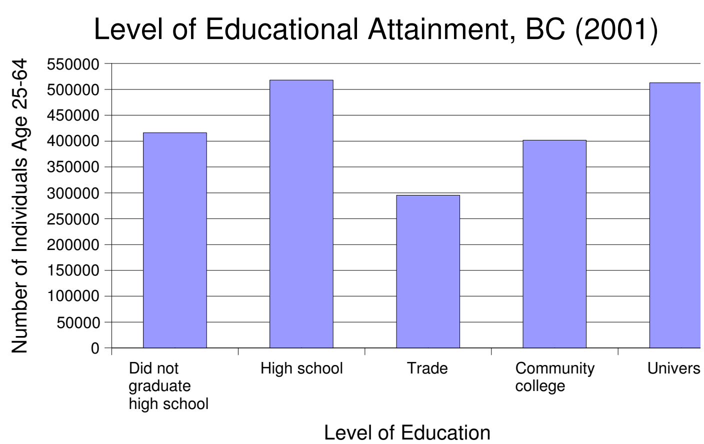
- Useful for data associated with categorical variables.
- Categories listed on the horizontal axis, heights of bars represent the size of each category.
- Example: Level of Educational Attainment for the Age Group 25 to 64, 2001 Counts for Both Sexes, British Columbia:

Level	Count	Percentage
did not graduate high school	416,245	19.4%
high school	518,150	24.2%
trade certificate	295,180	13.8%
community college	401,755	18.7%
university	512,715	23.9%
total	2,144,050	100.0%

Source: Statistics Canada

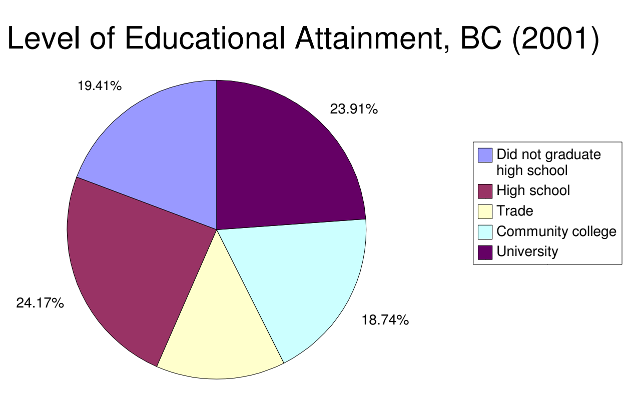
Bar Graph Example Continued

- Most spreadsheet programs can produce decent bar graphs.
- Using ‘OpenOffice’ (free):



Pie Charts

- Useful for categorical data in which the size of each category is given as a percentage of the total.
- Our Level of Education example again:

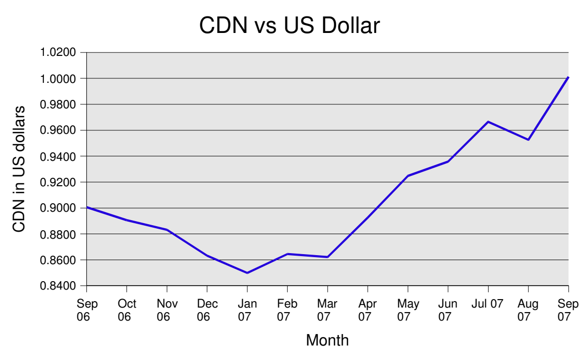


Line Graphs

- Typically used to graph the values of a variable varying in time.
- Time is plotted on the horizontal axis, and values of the variable on the vertical axis.
- Example: the value of the Canadian dollar vs the US dollar over the past year...

Line Graph Example Continued

CDN vs US dollar	
Date	Close (¢)
Sep 2007	100.12
Aug 2007	95.26
Jul 2007	96.65
Jun 2007	93.58
May 2007	92.48
Apr 2007	89.26
Mar 2007	86.22
Feb 2007	86.45
Jan 2007	84.99
Dec 2006	86.32
Nov 2006	88.32
Oct 2006	89.06
Sep 2006	90.06

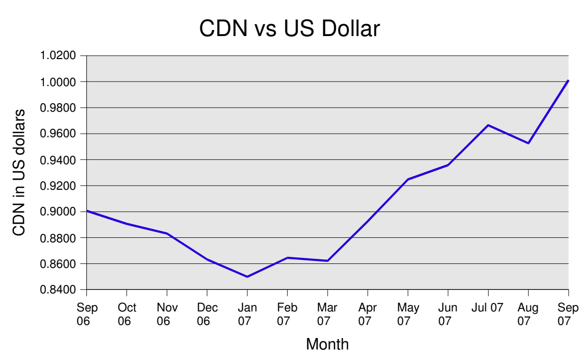


Source: finance.yahoo.ca

A Caution About Graphs

Graphs can emphasize or even misrepresent certain aspects of the data: beware!
Take another look at our Canadian vs US Dollar graph:

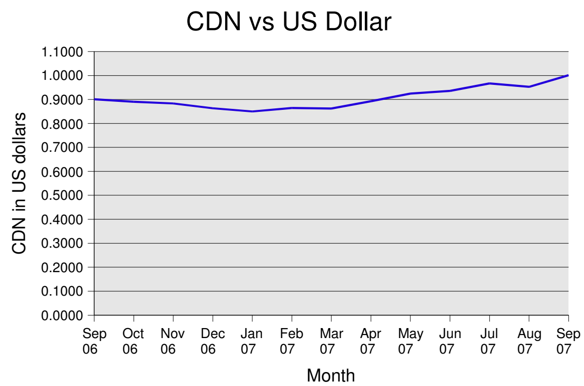
CDN vs US dollar	
Date	Close (¢)
Sep 2007	100.12
Aug 2007	95.26
Jul 2007	96.65
Jun 2007	93.58
May 2007	92.48
Apr 2007	89.26
Mar 2007	86.22
Feb 2007	86.45
Jan 2007	84.99
Dec 2006	86.32
Nov 2006	88.32
Oct 2006	89.06
Sep 2006	90.06



A Caution About Graphs (cont'd)

Notice how the variation is not nearly as dramatic if the vertical scale starts at zero:

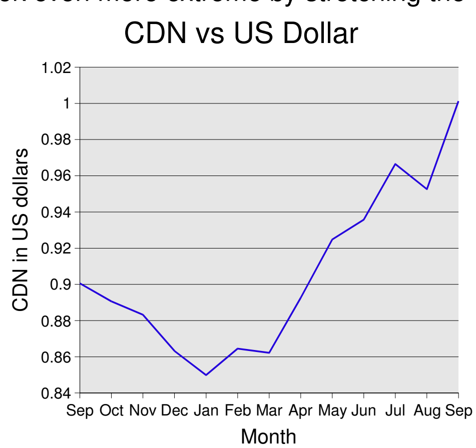
CDN vs US dollar	
Date	Close (¢)
Sep 2007	100.12
Aug 2007	95.26
Jul 2007	96.65
Jun 2007	93.58
May 2007	92.48
Apr 2007	89.26
Mar 2007	86.22
Feb 2007	86.45
Jan 2007	84.99
Dec 2006	86.32
Nov 2006	88.32
Oct 2006	89.06
Sep 2006	90.06



A Caution About Graphs (cont'd)

In the other direction, the variation in monthly values of the Canadian Dollar can be made to look even more extreme by stretching the vertical axis:

CDN vs US dollar	
Date	Close (¢)
Sep 2007	100.12
Aug 2007	95.26
Jul 2007	96.65
Jun 2007	93.58
May 2007	92.48
Apr 2007	89.26
Mar 2007	86.22
Feb 2007	86.45
Jan 2007	84.99
Dec 2006	86.32
Nov 2006	88.32
Oct 2006	89.06
Sep 2006	90.06



Good Graphing Habits

- Label axes stating what is represented, and give units.
- Put titles on your graphs.
- Avoid scaling axes so as to fool the reader (or yourself!)

Chapter 11: Displaying Distributions with Graphs: Histograms and Stemplots

Histograms

- A histogram is a type of bar graph used to display the distribution of a variable. The range of data (the variable's values) is first divided into classes, and the horizontal scale represents these classes while the vertical scale represents the frequency (i.e. count or percentage) of data values in each class.
- To construct a histogram:
  1. Divide the data range into non-overlapping classes of equal size
  2. Count the number occurences in each class.
  3. For each class, draw a bar with base covering the class and height equal to the count of occurences in the class.

Histogram Example

List of final grades from a math class:

73	55	76	83	87	93
53	67	97	75	68	61
77	55	85	47	57	
81	38	39	96	78	
53	60	65	64	79	
54	85	58	59	77	

Here the individuals are students, and the variable is 'final grade'. Plot a histogram showing the distribution of this variable. Use a class size of 10, and use a lower limit of 30 for the classes.

Stemplots

- A stemplot (or stem and leaf plot) is a graph which displays both the data values as well as their frequencies in a way which makes it easy to see the distribution without losing information about the exact data values.
- To construct a stemplot:
  1. Separate variable values into a stem consisting of all but the last digit, and a leaf: the last digit.
  2. List stems vertically, smallest to largest, and draw a vertical line beside the list.
  3. List, smallest to largest, each leaf beside its corresponding stem.

Stemplot Example

Contruct a stemplot of the final grade data from the previous example:

Here's the data again:

73	55	76	83	87	93
53	67	97	75	68	61
77	55	85	47	57	
81	38	39	96	78	
53	60	65	64	79	
54	85	58	59	77	

First, sort the data:

38	39	47	53	53	54
55	55	57	58	59	60
61	64	65	67	68	73
75	76	77	77	78	79
81	83	85	85	87	93
96	97				

Stemplot Example (cont'd)

Now list the stems consisting of the first digits:

```
3 |
4 |
5 |
6 |
7 |
8 |
9 |
```

Now add the leaves:

```
3 | 8 9
4 | 7
5 | 3 3 4 5 5 7 8 9
6 | 0 1 4 5 7 8
7 | 3 5 6 7 7 8 9
8 | 1 3 5 5 7
9 | 3 6 7
```

Describing Histograms I

One we have our histogram constructed, we wish to interpret it: what does its shape say about the data?

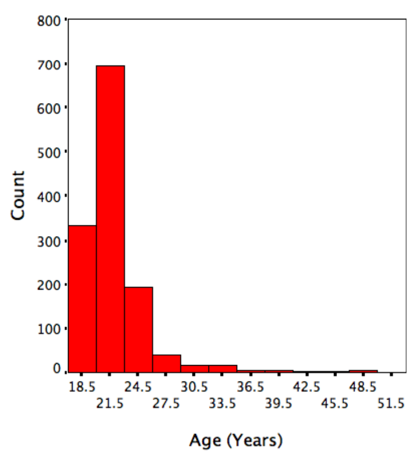
- What is the overall pattern?
  - Is is symmetric: the right and left sides are approximately mirror images of each other.
  - Is it skewed to the right: the right side of the histogram extends much farther out than the left side.
  - Similarly, is it skewed to the left: the left side of the histogram extends much farther out than the right side.

Describing Histograms II

- Are there any outliers: individual observations which lie outside of the overall pattern of the graph.
- What is the center and spread?
  - center: the value with roughly half the observations larger, half smaller.
  - spread: the range of values.
- What value is associated with the peak: the largest class.

Example I

Describe the distribution (symmetric? skewed? outliers? center? spread? peak?)



Example II

Describe the distribution (symmetric? skewed? outliers? center? spread? peak?)