Chapter 3

Graphical Methods for Describing Data

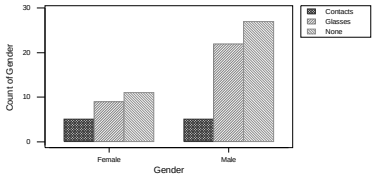
3.1 Displaying Graphical Data

Frequency Distribution Example

The data in the column labeled vision for the student data set introduced in the slides for chapter 1 is the answer to the question, “What is your principle means of correcting your vision?” The results are tabulated below

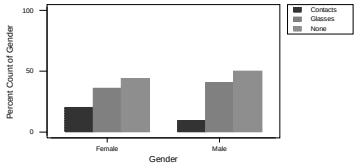
Vision Correction	Frequency	Relative Frequency
None	38	38/79 = 0.481
Glasses	31	31/79 = 0.392
Contacts	10	10/79 = 0.127
Total	79	1.000

Histogram Chart Examples



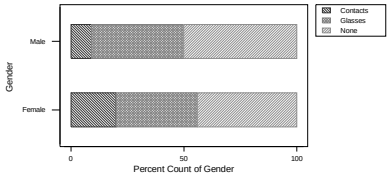
This comparative bar chart is based on frequencies and it can be difficult to interpret and misleading.
Would you mistakenly interpret this to mean that the females and males use contacts equally often?
You shouldn't. The picture is distorted because the frequencies of males and females are not equal.

Histogram Chart Examples



When the comparative bar chart is based on percents (or relative frequencies) (each group adds up to 100%) we can clearly see a difference in pattern for the eye correction proportions for each of the genders. Clearly for this sample of students, the proportion of female students with contacts is larger then the proportion of males with contacts.

Bar Chart Examples



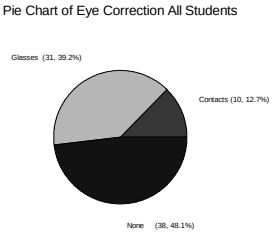
Stacking the bar chart can also show the difference in distribution of eye correction method. This graph clearly shows that the females have a higher proportion using contacts and both the no correction and glasses group have smaller proportions then for the males.

Pie Charts - Procedure

1. Draw a circle to represent the entire data set.
2. For each category, calculate the “slice” size.
   Slice size = 360(category relative frequency)
3. Draw a slice of appropriate size for each category.

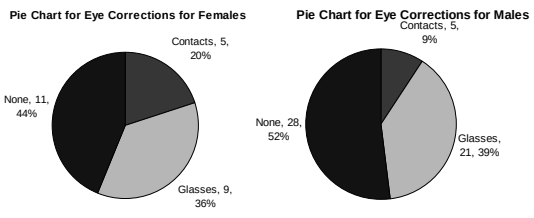
Pie Chart - Example

- Using the vision correction data we have:



Pie Chart - Example

- Using side-by-side pie charts we can compare the vision correction for males and females.



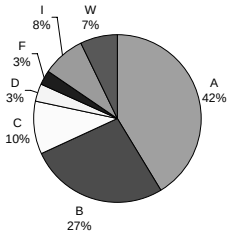
Another Example

This data constitutes the grades earned by the distance learning students during one term in the Winter of 2002.

Grade	Students	Student Proportion
A	454	0.414
B	293	0.267
C	113	0.103
D	35	0.032
F	32	0.029
I	92	0.084
W	78	0.071

Pie Chart – Another Example

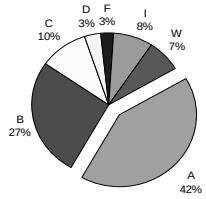
- Using the grade data from the previous slide we have:



Grade Distribution

Pie Chart – Another Example

- Using the grade data we have:



Grade Distribution

By pulling a slice (exploding) we can accentuate and make it clearing how A was the predominate grade for this course.

3.2: Numerical Data:

Stem and Leaf

A quick technique for picturing the distributional pattern associated with numerical data is to create a picture called a stem-and-leaf diagram (Commonly called a stem plot).

1. We want to break up the data into a reasonable number of groups.
2. Looking at the range of the data, we choose the stems (one or more of the leading digits) to get the desired number of groups.
3. The next digits (or digit) after the stem become(s) the leaf.
4. Typically, we truncate (leave off) the remaining digits.

Stem and Leaf

For our first example, we use the weights of the 25 female students.

Choosing the 1st two digits as the stem and the 3rd digit as the leaf we have the following

150	140	155	195	139
200	157	130	113	130
121	140	140	150	125
135	124	130	150	125
120	103	170	124	160

```
10 | 3
11 | 3
12 | 154504
13 | 90050
14 | 000
15 | 05700
16 | 0
17 | 0
18 |
19 | 5
20 | 0
```

Stem and Leaf

Typically we sort the order the stems in increasing order.
We also note on the diagram the units for stems and leaves

```
10 | 3
11 | 3
12 | 014455
13 | 00059
14 | 000
15 | 00057
16 | 0
17 | 0
18 |
19 | 5
20 | 0
```

Probable outliers

Stem: Tens and hundreds digits
Leaf: Ones digit

Stem-and-leaf – GPA example

The following are the GPAs for the 20 advisees of a faculty member.

GPA					
3.09	2.04	2.27	3.98	3.70	2.99
3.72	3.23	3.13	3.50	2.26	3.15
2.80	1.75	3.89	3.38	2.74	1.65
2.22	2.66				

If the ones digit is used as the stem, you only get three groups. You can expand this a little by breaking up the stems by using each stem twice letting the 2nd digits 0-4 go with the first and the 2nd digits 5-9 with the second. Call it Low & High

The next slide gives two versions of the stem-and-leaf diagram.

Stem-and-leaf – GPA example

```
1L |
1H | 67
2L | 0222
2H | 6978
3L | 01123
3H | 57789
```

Stem: Ones digit

Leaf: Tenths digits

```
1L |
1H | 65,75
2L | 04,22,26,27
2H | 66,99,74,80
3L | 09,13,15,23,38
3H | 50,70,72,89,98
```

Stem: Ones digit

Leaf: Tenths and hundredths digits

Note: The characters in a stem-and-leaf diagram must all have the same width, so if typing a fixed character width font such as courier.

Comparative Stem & Leaf Diagram

Student Wt (Comparing 2 groups)

When it is desirable to compare two groups, back-to-back stem and leaf diagrams are useful. Here is the result from the student weights.

From this comparative stem and leaf diagram, it is clear that the males weigh more (as a group not necessarily as individuals) than the females.

```
     3 | 10 |
     3 | 11 | 7
554410 | 12 | 145
 95000 | 13 | 0004558
   000 | 14 | 000000555
 75000 | 15 | 0005556
     0 | 16 | 00005558
     0 | 17 | 000005555
       | 18 | 0358
     5 | 19 |
     0 | 20 | 0
       | 21 | 0
       | 22 | 55
       | 23 | 79
```

Comparative Stem & Leaf Diagram Student Age

```
female |  | male
     7 | 1 |
  9999 | 1 | 88888999999999999999
 1111000 | 2 | 00000001111111111
 3322222 | 2 | 2222223333
     4 | 2 | 445
       | 2 | 6
       | 2 | 88
     0 | 3 |
       | 3 |
       | 3 |
     7 | 3 |
     8 | 3 |
       | 4 |
       | 4 |
     4 | 4 |
     7 | 4 |
```

From this comparative stem and leaf diagram, it is clear that the male ages are all more closely grouped then the females. Also the females had a number of outliers.

3.3: Frequency Distributions & Histograms

- When working with discrete data, the frequency tables are similar to those produced for qualitative data.
- For example, a survey of local law firms in a medium sized town gave

Number of Lawyers	Frequency	Relative Frequency
1	11	0.44
2	7	0.28
3	4	0.16
4	2	0.08
5	1	0.04

Frequency Distributions & Histograms

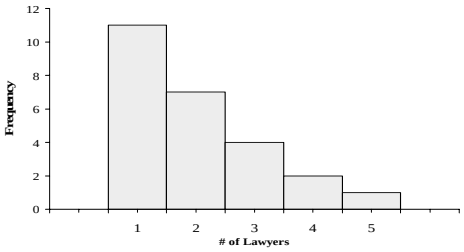
- When working with discrete data, the steps to construct a histogram are
  1. Draw a horizontal scale, and mark the possible values.
  2. Draw a vertical scale and mark it with either frequencies or relative frequencies (usually start at 0).
  3. Above each possible value, draw a rectangle whose height is the frequency (or relative frequency) centered at the data value with a width chosen appropriately. Typically if the data values are integers then the widths will be one.

Frequency Distributions & Histograms

- Look for a central or typical value, extent of spread or variation, general shape, location and number of peaks, and presence of gaps and outliers.

Frequency Distributions & Histograms

- The number of lawyers in the firm will have the following histogram.



Clearly, the largest group are single member law firms and the frequency decreases as the number of lawyers in the firm increases.

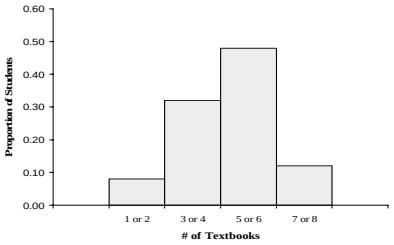
Frequency Distributions & Histograms

- 50 students were asked the question, “How many textbooks did you purchase last term?” The result is summarized below and the histogram is on the next slide.

# of Textbooks	Frequency	Relative Frequency
1 or 2	4	0.08
3 or 4	16	0.32
5 or 6	24	0.48
7 or 8	6	0.12

Frequency Distributions & Histograms

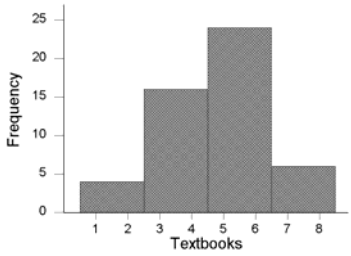
- "How many textbooks did you purchase last term?"



The largest group of students bought 5 or 6 textbooks with 3 or 4 being the next largest frequency.

Frequency Distributions & Histograms

- Another version with the scales produced differently.



Frequency Distributions & Histograms

- When working with continuous data, the steps to construct a histogram are
  1. Decide into how many groups or "classes" you want to break up the data. Typically somewhere between 5 and 20. *A good rule of thumb is to think having an average of more than 5 per group.**
  2. Use your answer to help decide the "width" of each group.
  3. Determine the "starting point" for the lowest group.

*A quick estimate for a reasonable number of intervals is $\sqrt{\text{number of observations}}$

Example of Frequency Distribution

- Consider the student weights in the student data set. The data values fall between 103 (lowest) and 239 (highest). The range of the dataset is 239-103=136.
- There are 79 data values, so to have an average of at least 5 per group, we need 16 or fewer groups. We need to choose a width that breaks the data into 16 or fewer groups. Any width 10 or large would be reasonable.

Example of Frequency Distribution

- Choosing a width of 15 we have the following frequency distribution.

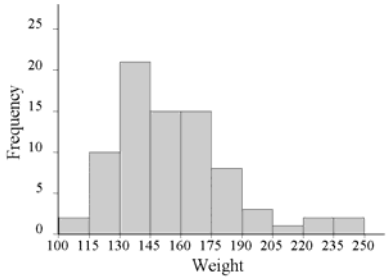
Class Interval	Frequency	Relative Frequency
100 to <115	2	0.025
115 to <130	10	0.127
130 to <145	21	0.266
145 to <160	15	0.190
160 to <175	15	0.190
175 to <190	8	0.101
190 to <205	3	0.038
205 to <220	1	0.013
220 to <235	2	0.025
235 to <250	2	0.025
	79	1.000

Histogram for Continuous Data

- Mark the boundaries of the class intervals on a horizontal axis
- Use frequency or relative frequency on the vertical scale.

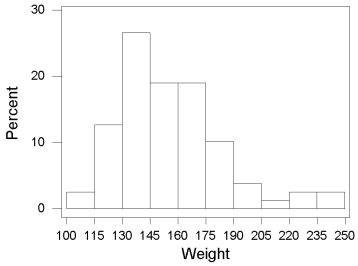
Histogram for Continuous Data

- The following histogram is for the frequency table of the weight data.



Histogram for Continuous Data

- The following histogram is the Minitab output of the relative frequency histogram. *Notice that the relative frequency scale is in percent.*



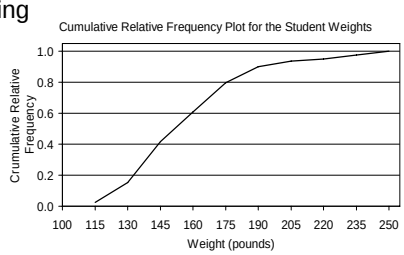
Cumulative Relative Frequency Table

- If we keep track of the proportion of that data that falls below the upper boundaries of the classes, we have a **cumulative relative frequency table**.

Class Interval	Relative Frequency	Cumulative Relative Frequency
100 to < 115	0.025	0.025
115 to < 130	0.127	0.152
130 to < 145	0.266	0.418
145 to < 160	0.190	0.608
160 to < 175	0.190	0.797
175 to < 190	0.101	0.899
190 to < 205	0.038	0.937
205 to < 220	0.013	0.949
220 to < 235	0.025	0.975
235 to < 250	0.025	1.000

Cumulative Relative Frequency Plot

- If we graph the cumulative relative frequencies against the upper endpoint of the corresponding interval, we have a **cumulative relative freq plot**.



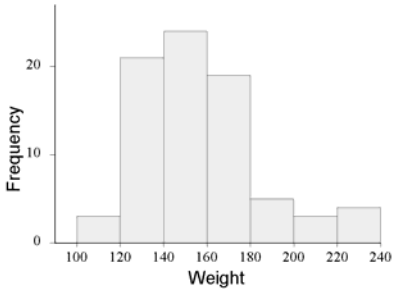
Histogram for Continuous Data

- Another version of a frequency table and histogram for the weight data with a class width of 20.

Class Interval	Frequency	Relative Frequency
100 to <120	3	0.038
120 to <140	21	0.266
140 to <160	24	0.304
160 to <180	19	0.241
180 to <200	5	0.063
200 to <220	3	0.038
220 to <240	4	0.051
	79	1.001

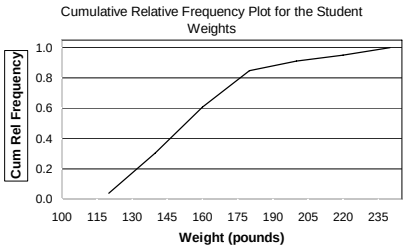
Histogram for Continuous Data

- The resulting histogram.



Histogram for Continuous Data

- The resulting cumulative relative frequency plot.



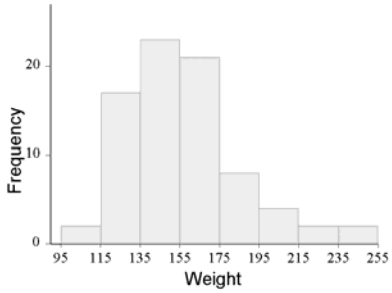
Histogram for Continuous Data

- Yet, another version of a frequency table and histogram for the weight data with a class width of 20.

Class Interval	Frequency	Relative Frequency
95 to <115	2	0.025
115 to <135	17	0.215
135 to <155	23	0.291
155 to <175	21	0.266
175 to <195	8	0.101
195 to <215	4	0.051
215 to <235	2	0.025
235 to <255	2	0.025
	79	0.999

Histogram for Continuous Data

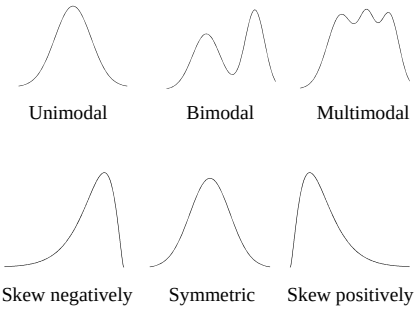
- The corresponding histogram.



Histogram for Continuous Data

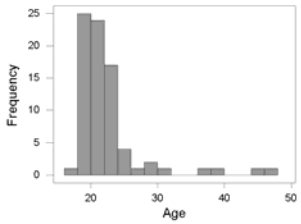
- A class width of 15 or 20 seems to work well because all of the pictures tell the same story.
- The bulk of the weights appear to be centered around 150 lbs with a few values substantially large. The distribution of the weights is **unimodal** and is **positively skewed**.

Illustrated Distribution Shapes



Histograms with uneven class widths

- Consider the following frequency histogram of ages based on A with class widths of 2. Notice it is a bit choppy. Because of the positively skewed data, sometimes frequency distributions are created with unequal class widths.



Histograms with uneven class widths

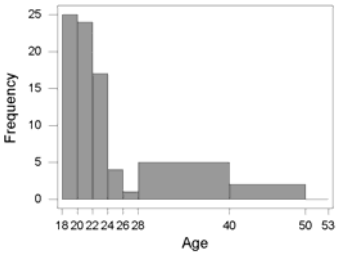
- For many reasons, either for convenience or because that is the way data was obtained, the data may be broken up in groups of uneven width as in the following example referring to the student ages.

Class Interval	Frequency	Relative Frequency
18 to <20	26	0.329
20 to <22	24	0.304
22 to <24	17	0.215
24 to <26	4	0.051
26 to <28	1	0.013
28 to <40	5	0.063
40 to <50	2	0.025

Histograms with uneven class widths

- If a frequency (or relative frequency) histogram is drawn with the heights of the bars being the frequencies (relative frequencies), the result is distorted.

Notice that it appears that there are a lot of people over 28 when there is only a few.



Histograms with uneven class widths

- To correct the distortion, we create a density histogram. The vertical scale is called the density and the density of a class is calculated by

$$\text{density} = \text{rectangle height} = \frac{\text{relative frequency of class}}{\text{class width}}$$

This choice for the density makes the area of the rectangle equal to the relative frequency.

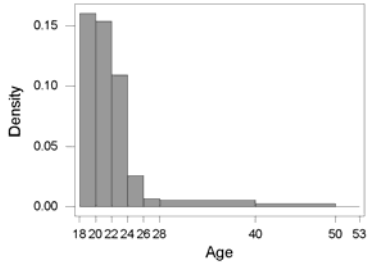
Histograms with uneven class widths

- Continuing this example we have

Class Interval	Frequency	Relative Frequency	Density
18 to <20	26	0.329	0.165
20 to <22	24	0.304	0.152
22 to <24	17	0.215	0.108
24 to <26	4	0.051	0.026
26 to <28	1	0.013	0.007
28 to <40	5	0.063	0.005
40 to <50	2	0.025	0.003

Histograms with uneven class widths

- The resulting histogram is now a reasonable representation of the data.



3.4: Displaying Bivariate Data

Scatterplots

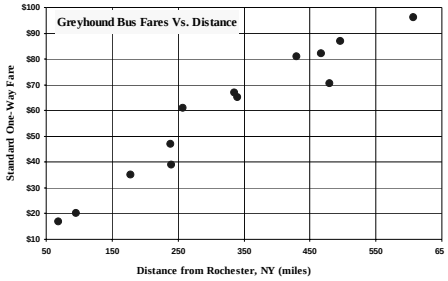
- A scatterplot is a plot of pairs of observed values (both quantitative) of two different variables. It's plotting the (x, y) ordered pair on coordinate plane like you have did in Alg/Geo
- When one of the variables is considered to be a response variable (y) and the other an explanatory variable (x). The explanatory variable is usually plotted on the x axis

Example

A sample of one-way Greyhound bus fares from Rochester, NY to cities less than 750 miles was taken by going to Greyhound's website. The following table gives the destination city, the distance and the one-way fare. Distance should be the x axis and the Fare should be the y axis.

Destination City	Distance	Standard One-Way Fare
Albany, NY	240	39
Baltimore, MD	430	81
Buffalo, NY	69	17
Chicago, IL	607	96
Cleveland, OH	257	61
Montreal, QU	480	70.5
New York City, NY	340	65
Ottawa, ON	467	82
Philadelphia, PA	335	67
Potsdam, NY	239	47
Syracuse, NY	95	20
Toronto, ON	178	35
Washington, DC	496	87

Example Scatterplot



Displaying Bivariate Data

Time series

- Just plot most any univariate data against time (if applicable). As an example you could measure your height every hour during the day (interested in height), then plot with time on the x axis to see the change in ht as the day goes by.