Announcements

- Please make sure you register your clicker at the clicker website. Link at course webpage.
- When you turn in homework, please do problems in the order assigned. (All of Friday, then Monday, etc.) Put your name, Discussion Section number, and ID (or last 6 digits) in upper right. Tear off ragged edges. Turn in here or in slot on wall opposite 2202 Bren Hall by 6pm Friday.
- Homework: Chapter 2, #42ac, 51, 61, 75

Sections 2.4 to 2.6 Summarizing quantitative variables

…including one quantitative and one categorical variable

Data used for some examples today:

Dataset "UCDavis1" from CD – measured many variables on 173 students in an intro stats class. Four of the variables were:

Sex (Male or Female)

Height (in inches)

Exercise (hours per week, on average)

Alcohol (drinks consumed per week, on average)

Data for the first 6 students:

Sex (Category)	Height (inches)	Exercise (hours/week)	Alcohol (drinks/week)
Female	66	10	12
Female	64	5	0
Male	72	2	0
Male	68	3	0
Male	68	6	0
Female	64	6.5	5

Summary Features of Quantitative Data

1. Location (Center, Average)
2. Spread (Variability)
3. Shape
4. Outliers (Unusual values)

We use pictures *and* numerical information to examine these.

Questions about quantitative variables:

One Quantitative Variable

Question 1: What interesting summary measures, like the average or the range of values, can help us understand the collection of individuals who were measured?

Example: What is the average exercise per week, and how much variability is there in exercise amounts?

Question 2: Are there individual data values that provide interesting information because they are unique or stand out in some way?

Example: What is the oldest verified age of death for a human? Are there many people who have lived nearly that long, or is the oldest recorded age a unique case?

(Note: So far, oldest was 122 years, 164 days; died 1997.)

One Categorical and One Quantitative Variable (Comparing across categories)

Question 1: Are the quantitative measurements similar across categories of the categorical variable?

Example: Do men and women exercise the same amounts, on average? Do they drink the same amounts?

Question 2: When the categories have a natural ordering (an ordinal variable), does the quantitative variable increase or decrease, on average, in that same order?

Example: Do high school dropouts, high school graduates, college dropouts, and college graduates have increasingly higher average incomes?

2.4 Pictures for Quantitative Data

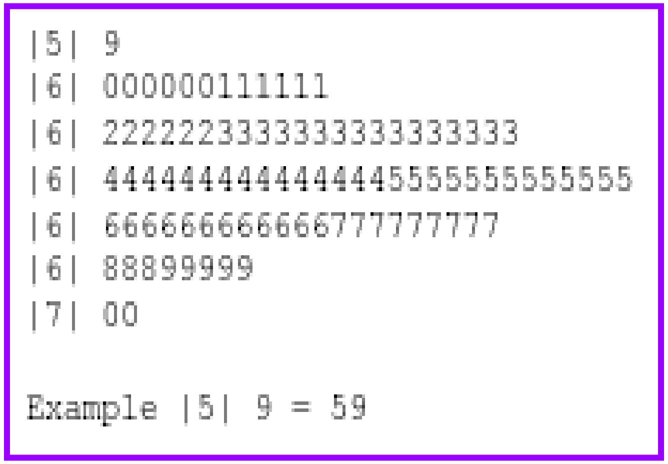
- **Look at** shape, outliers, center (location), spread, gaps, any other interesting features.

Four common types of pictures:

- **Histograms**: similar to bar graphs, used for *any number* of data values.
- **Stem-and-leaf plots** and **dotplots**: present *all individual values*, useful for *small to moderate* sized data sets.
- **Boxplot** or **box-and-whisker plot**: useful *summary* for *comparing* two or more groups.

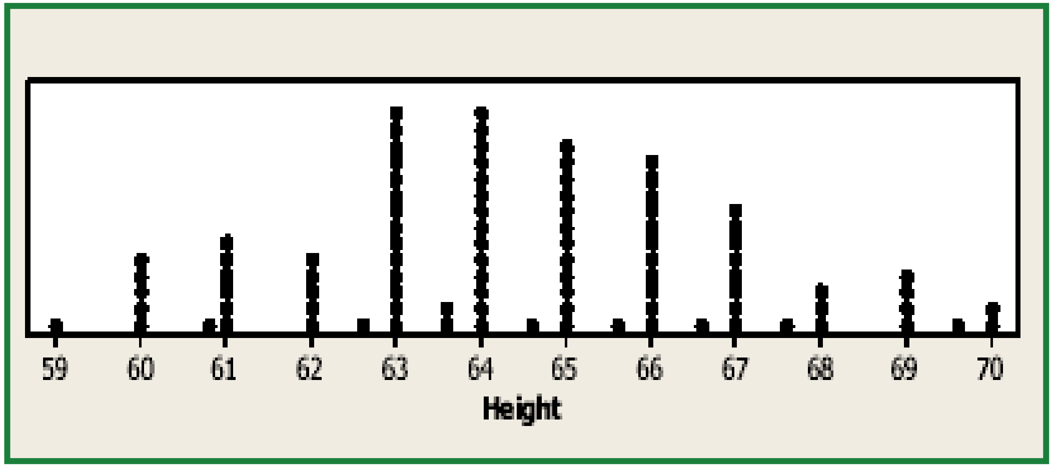
Stemplots, Dotplots and Histograms

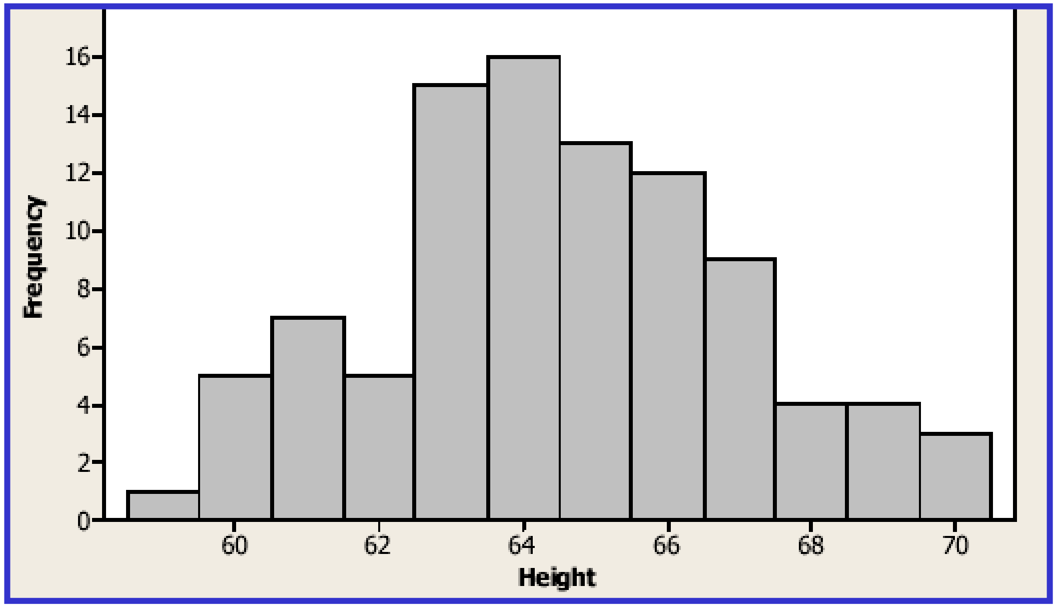
EX: Heights for 94 Females



```
|5| 9
|6| 000000111111
|6| 2222223333333333333333
|6| 4444444444444445555555555555
|6| 666666666666777777777
|6| 88899999
|7| 00

Example |5| 9 = 59
```





- Values are centered around 64 or 65 inches.
- “Bell-shaped,” no outliers
- Spread is 59 to 70 in.

Creating a Histogram by hand

Step 1: Decide **how many** *equally spaced* (same width) **intervals** to use for the horizontal axis. Between 6 and 15 intervals is a good number (more if there are gaps and/or outliers). Decide where to put values that are on the boundary. For instance, does 2 go in the interval from 0 to 2, or from 2 to 4? Just need to be consistent.

Example: Exercise values ranged from 0 to 30 hours a week. Use 15 intervals of width 2 hours each, so intervals are 0 to 2, 2.1 to 4, etc., up to 28.1 to 30.

Creating a Histogram, continued

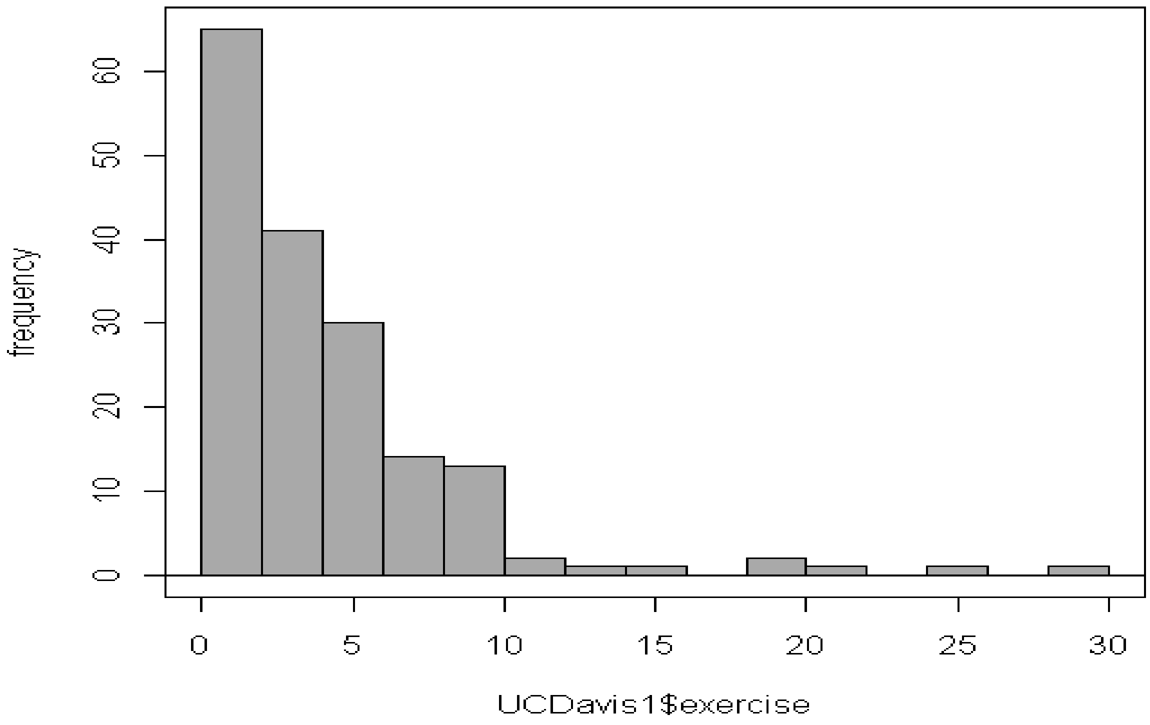
Step 2: Decide to use ***frequencies*** (count) or ***relative frequencies*** (proportion) on the vertical axis.

Example: Use frequencies, i.e *number* of people who exercise each of the amounts, rather than the *proportion* who do so.

Step 3: Draw *equally spaced intervals* on horizontal axis covering entire range of the data. Determine frequency or relative frequency of data values in each interval and draw a bar with corresponding height.

Exercise hours per week, n = 172

Note that 15 intervals are used; some gaps.



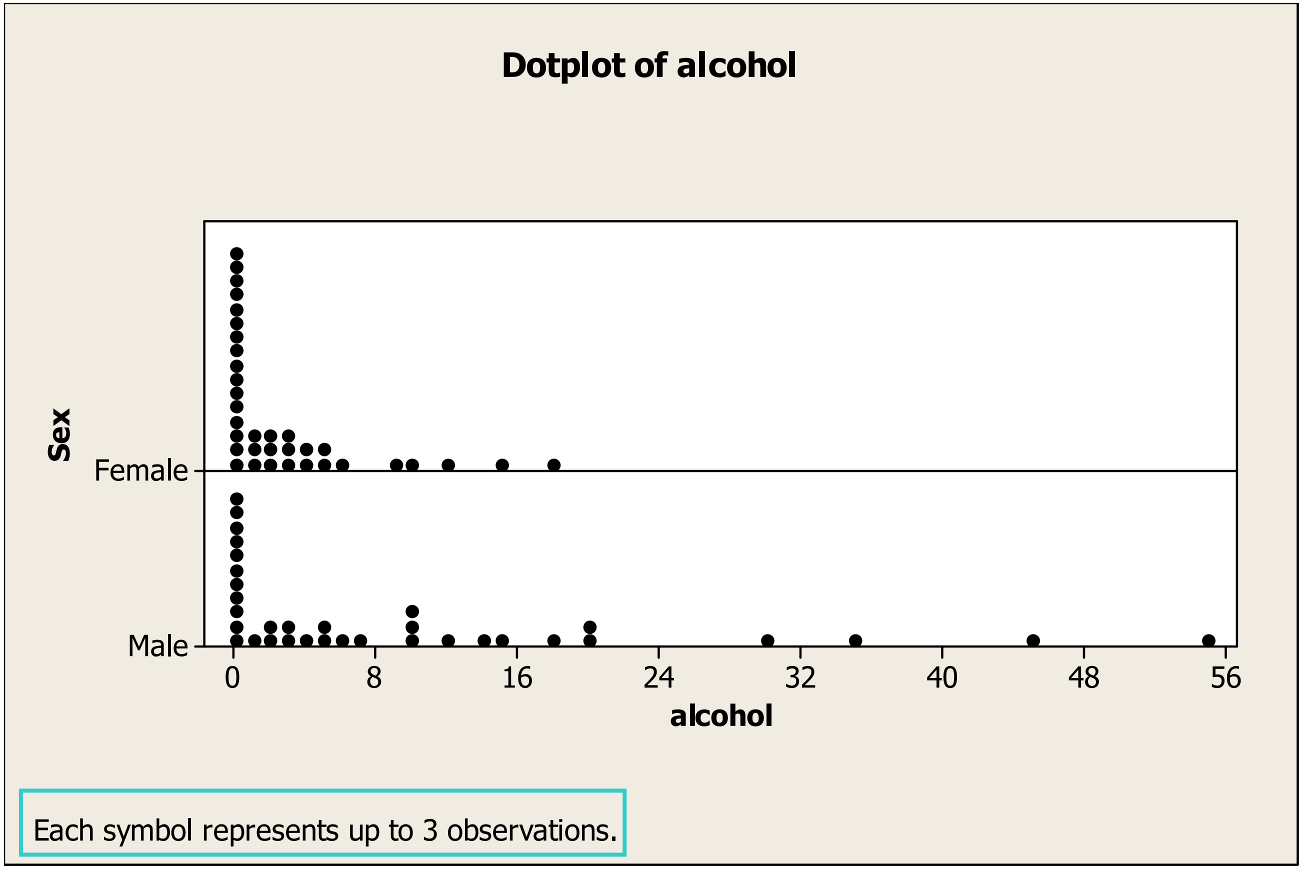
This was done using R Commander, but shows what would be done by hand.

Creating a Dotplot

These can be useful for comparing groups

- Ideally, number line represents all possible values and there is one dot per observation. Not always possible. If dots represent multiple observations, footnote should explain that.
- As with histogram, divide horizontal axis into equal intervals, then put dots on it for each individual in each interval.
- Saw an example last time, “fastest speed driven” comparing males and females.

Example: Alcoholic drinks/week, comparing females and males



Each symbol represents up to 3 observations.

Creating a Stemplot (stem and leaf plot) - Example of 25 pulse rates:

65, 78, 60, 58, 62, 64, 75, 71, 74, 72, 66, 69, 67, 54, 65, 70, 63, 57, 65, 63, 70, 59, 68, 64, 67

Step 1: Create the Stem

Divide range of data into equal units to be used on **stem**. Have 6 to 15 stem values, representing *equally spaced* intervals. Here, we could use 2 or 5 beats/min.

```
5|
5|
6|
6|
7|
7|
```

Example: each of the 6 stem values represents a range of 5 beats of pulse rate

Creating a Stemplot

Step 2: Attach the Leaves

Attach a **leaf** to represent each data point. Next digit in number used as leaf; drop remaining digits, if any.

Step 2: Attaching leaves

5|

5|8

6|0

6|5

7|

7|8

Example: 5|8 = 58

Example: Pulse rates

65, 78, 60, 58 ,. . .

First 4 value attached.

Optional Step: order leaves on each branch.

Further Details for Creating Stemplots

Splitting Stems:

Reusing digits two or five times.

Stemplot A:

```
5|4
5|789
6|023344
6|55567789
7|00124
7|58
```

Two times:

1st stem = leaves 0 to 4

2nd stem = leaves 5 to 9

Stemplot B:

```
5|4
5|7
5|89
6|0
6|233
6|44555
6|677
6|89
7|001
7|2
7|45
7|
7|8
```

Five times:

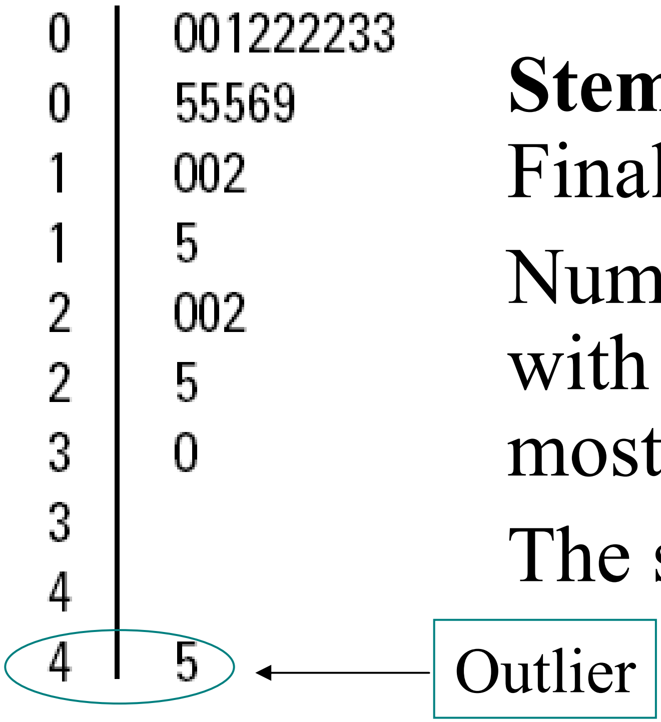
1st stem = leaves 0 and 1

2nd stem =leaves 2 and 3, etc.

Example 2.8 *Big Music Collection*

About how many CDs do you own?

Estimated music CDs owned for $n = 24$ Penn State students



0	001222233
0	55569
1	002
1	5
2	002
2	5
3	0
3	
4	
4	5

Outlier

Ex: 4|5 = 450's

Stem is '100s' and **leaf** unit is '10s'. Final digit is **truncated**.

Numbers ranged from 0 to about 450, with 450 being a clear **outlier** and most values ranging from 0 to 99.

The shape is **skewed right**.

Describing Shape

- **Symmetric**, **bell-shaped**

  (Female heights bell-shaped, also pulse rates)
- **Symmetric**, ***not*** **bell-shaped**
- **Bimodal:** Two prominent "peaks" (modes)
- **Skewed Right**: On number line, values clumped at left end and *extend* to the *right*

  (CDs, alcohol and exercise all skewed to *right*)
- **Skewed Left**: On number line, values clumped at right end and *extend* to the *left* (Ex: Age at death from heart attack.)

Example: How Much Do Students Exercise?

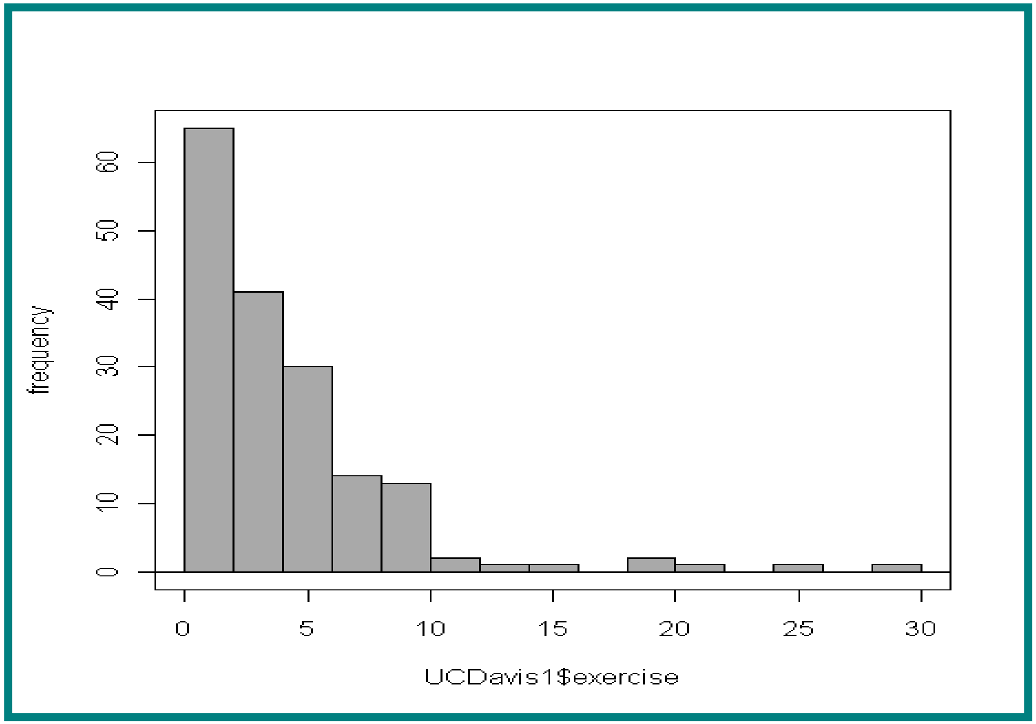
How many hours do you exercise a week (nearest ½ hr)?

Shape is *skewed to the right*

172 responses from students in intro statistics class

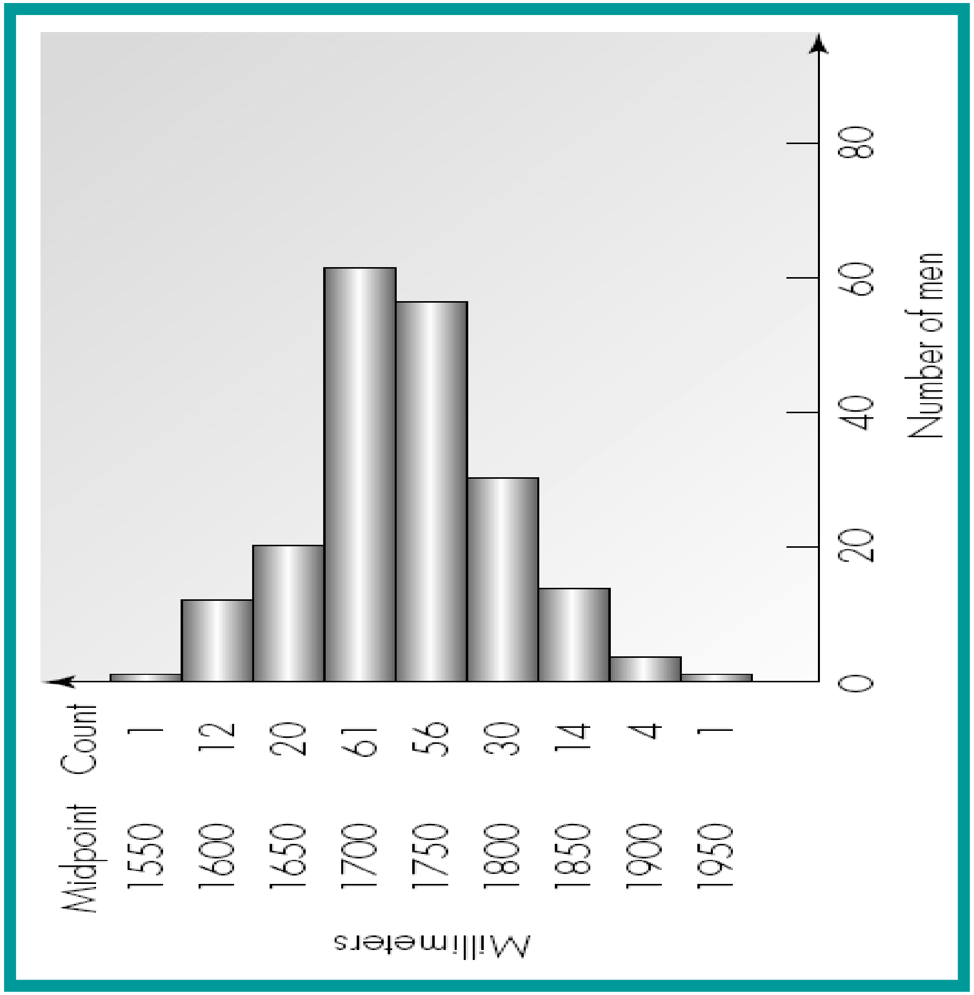
Most range from 0 to 10 hours with mode of 2 hours.

Responses trail out to 30 hours a week.



Bell-shaped example: Heights of British Males

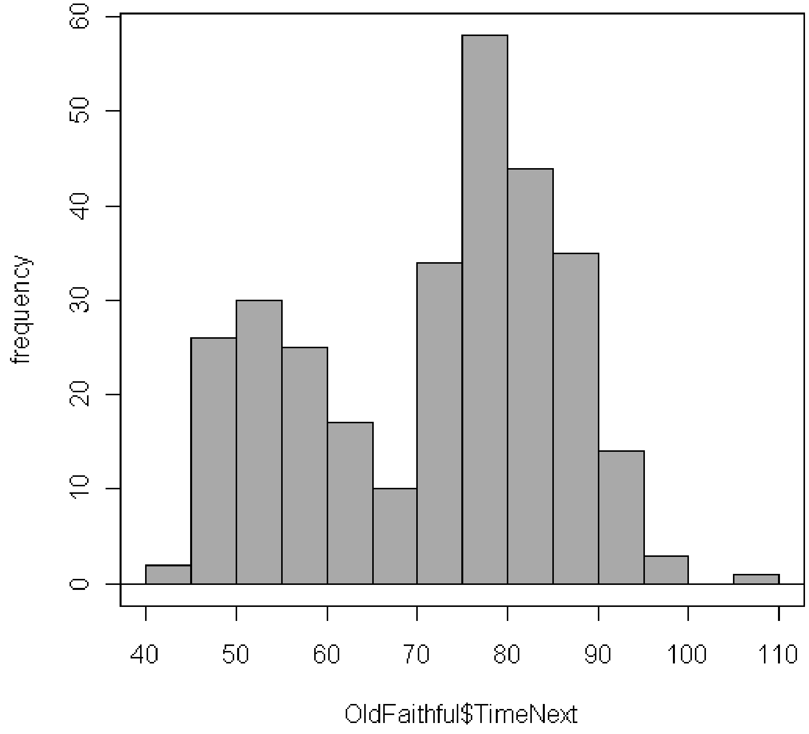
Heights of 199 randomly selected British men, in millimeters. Bell-shaped, centered in the mid-1700s mm with no outliers.



Source: Marsh, 1988, p. 315; data reproduced in Hand et al., 1994, pp. 179-183

Bimodal Example: The Old Faithful Geyser – time *between* eruptions (in book, Fig. 2.10 shows *duration* of eruptions), histogram from R Commander.

Times between eruptions of the Old Faithful geyser, shape is **bimodal**. Two clusters, one around 50 min., other around 80 min.



Source: Hand et al., 1994

Five Number Summary – a simple quantitative summary:

The five-number summary display

	Median	
Lower Quartile		Upper Quartile
Lowest		Highest

- **Lowest** = Minimum
- **Highest** = Maximum
- **Median** = number such that half of the values are at or above it and half are at or below it (middle value or average of two middle numbers in ordered list).
- **Quartiles** = medians of the two halves.

Boxplots

Visual picture of the five-number summary

Example: How much do statistics students sleep?

190 statistics students asked how many hours they slept the night before (a Tuesday night).

Five-number summary for number of hours of sleep

	7	
6		**8**
3		**16**

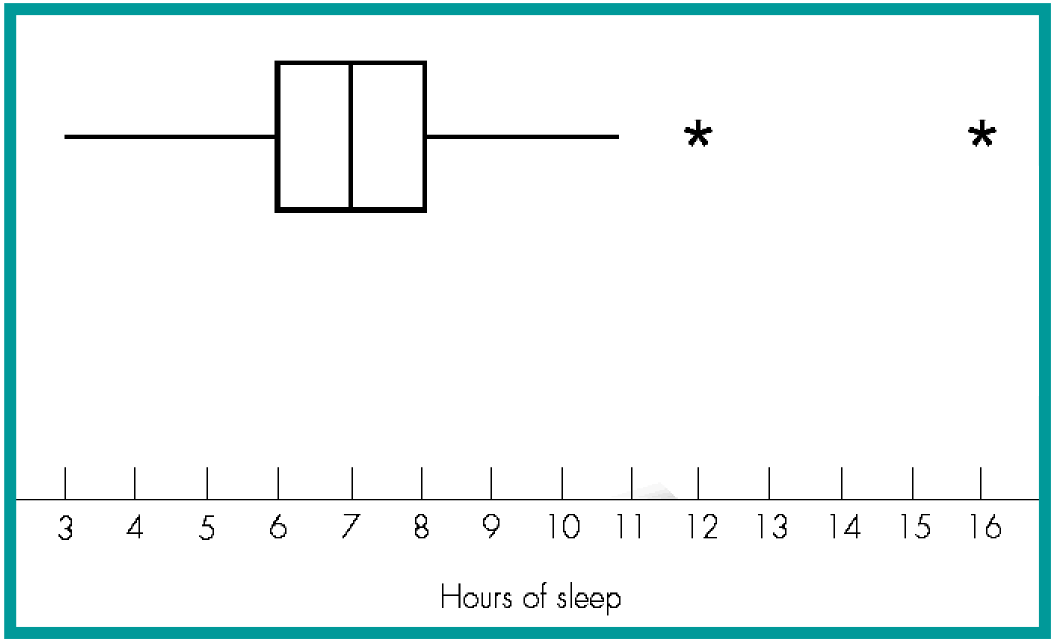
Two students reported 16 hours; the max for the remaining 188 students was 12 hours.

Creating a Boxplot

1. Draw horizontal (or vertical) line, label it with values from lowest to highest in data.
2. Draw rectangle (box) with ends at quartiles.
3. Draw line in box at value of median.
4. Compute IQR = distance between quartiles.
5. Compute 1.5(IQR); *outlier* is any value more than this distance from closest quartile. Draw line (whisker) from each end of box extending to farthest data value that is not an outlier. (If no outlier, then to min and max.)
6. Draw asterisks to indicate the outliers.

Creating a Boxplot for Sleep Hours

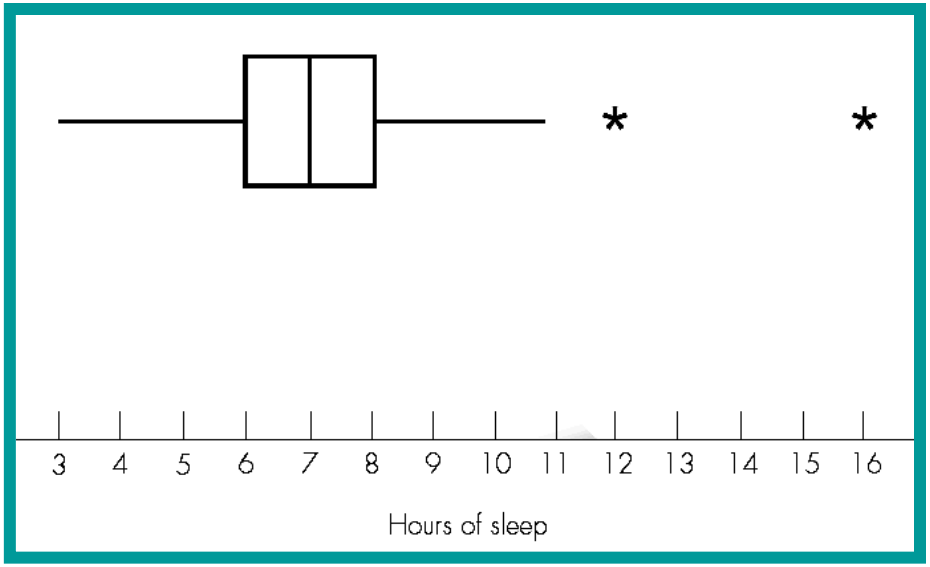
1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8 (quartiles).
3. Draw line in box at median of 7.
4. Compute IQR = 8 – 6 = 2.
5. Compute 1.5(IQR) = 1.5(2) = 3; outlier is any value below 6 – 3 = 3, or above 8 + 3 = 11.
6. Draw line from each end of box extending down to 3 but up to 11.
7. Draw asterisks at outliers of 12 and 16 hours.



Interpreting Boxplots

- Divide the data into fourths.
- Easily identify outliers.
- Useful for comparing two or more groups.

Outlier: any value more than 1.5(IQR) beyond closest quartile.



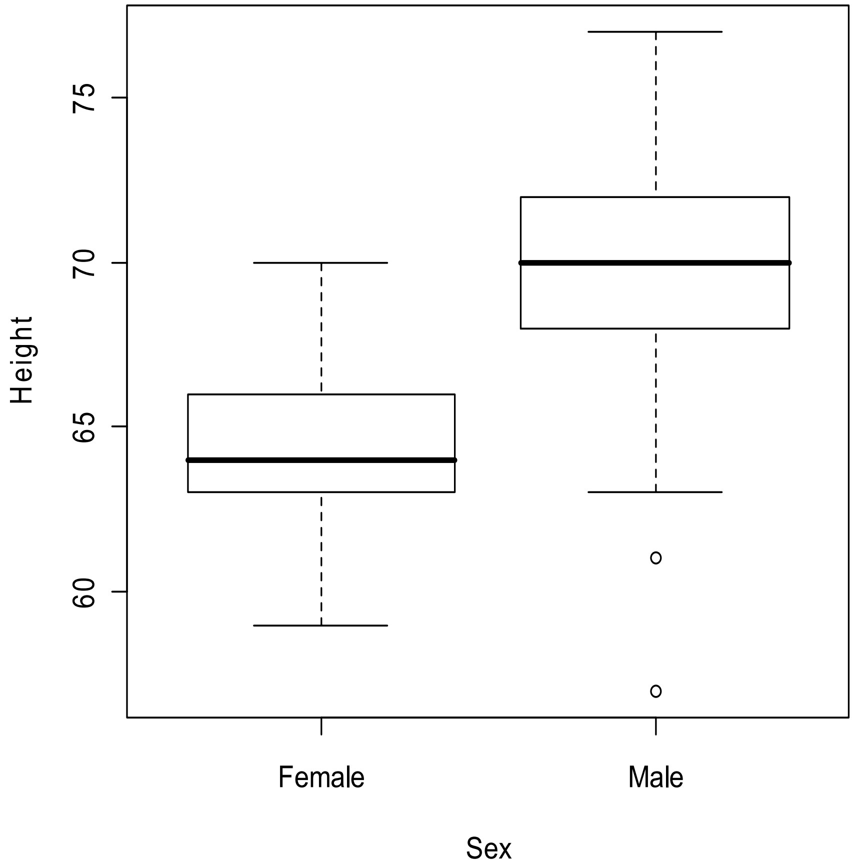
¼ of students slept between 3 and 6 hours

¼ slept between 6 and 7 hours

¼ slept between 7 and 8 hours

¼ slept between 8 and 16 hours

Sometimes boxplots are vertical instead of horizontal



Example: Boxplot of female and male heights, created using R Commander

2.6 Outliers and How to Handle Them

Outlier: a data point that is not consistent with the bulk of the data.

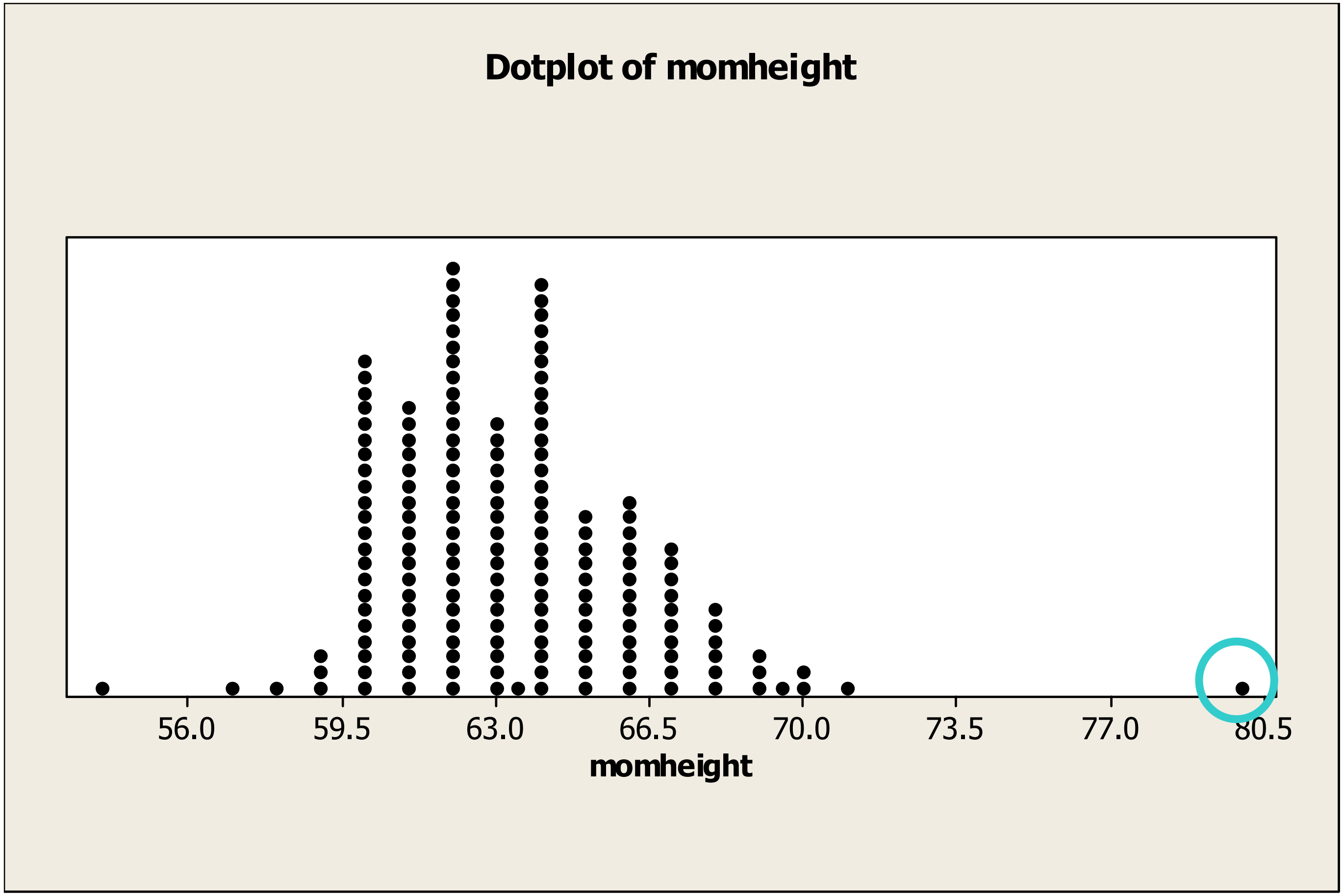
- Look for them via graphs.
- Can have big influence on conclusions.
- Can cause complications in some statistical analyses.
- Cannot discard without justification.

Example: 450 CDs

Possible reasons for outliers and what to do about them:

1. *Mistake made while taking measurement or entering it into computer*. If verified, should be discarded or corrected.
2. *Individual in question belongs to a different group than bulk of individuals measured*. Values may be discarded if summary is desired and reported for the majority group only.
3. *Outlier is legitimate data value and represents natural variability for the group and variable(s) measured*. Values may <u>not</u> be discarded. They provide important information about location and spread.

Example: *Students gave mother's height*



Height of 80 inches = 6 ft 8 inches, almost surely an error!
Reason #1, investigate and try to find error; remove value.

Example 2.16 *Tiny Boatmen*

Weights (in pounds) of 18 men on crew team:

Cambridge: 188.5, 183.0, 194.5, 185.0, 214.0,
203.5, 186.0, 178.5, **109.0**

Oxford: 186.0, 184.5, 204.0, 184.5, 195.5,
202.5, 174.0, 183.0, **109.5**

Note: last weight in each list is unusually small. ???

Example 2.16 *Tiny Boatmen*

Weights (in pounds) of 18 men on crew team:

Cambridge: 188.5, 183.0, 194.5, 185.0, 214.0,
203.5, 186.0, 178.5, **109.0**

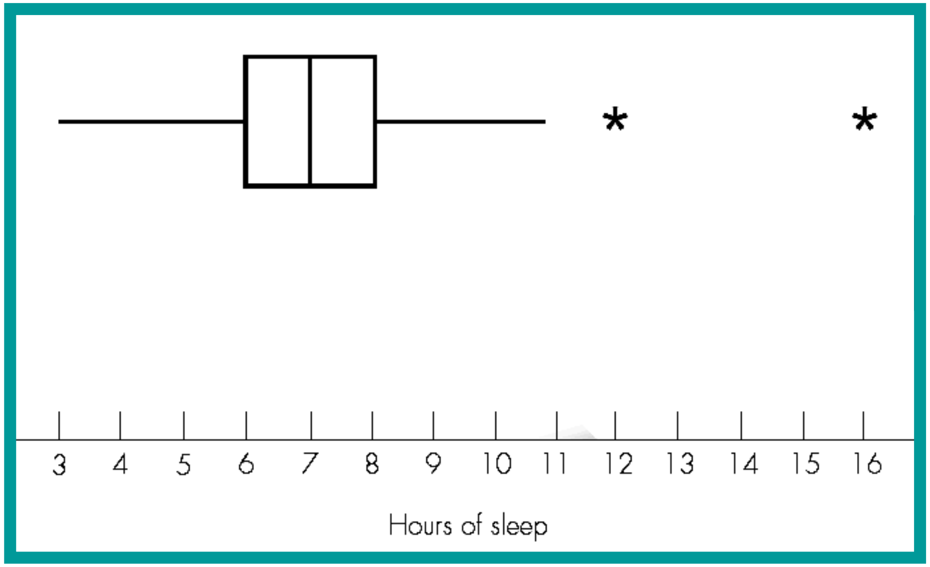
Oxford: 186.0, 184.5, 204.0, 184.5, 195.5,
202.5, 174.0, 183.0, **109.5**

Note: last weight in each list is unusually small. ???

They are the *coxswains* for their teams, while others are *rowers*. **Reason 2**: different group, okay to remove if only interested in rowers.

Example: *Sleep hours*

Two students were outliers in amount of sleep, but the values were not mistakes.



Reason 3: Natural variability, it is *not* okay to remove these values.

Real life example of the use of picture of quantitative data:

Detecting Exam Cheating with a Dotplot

Details:

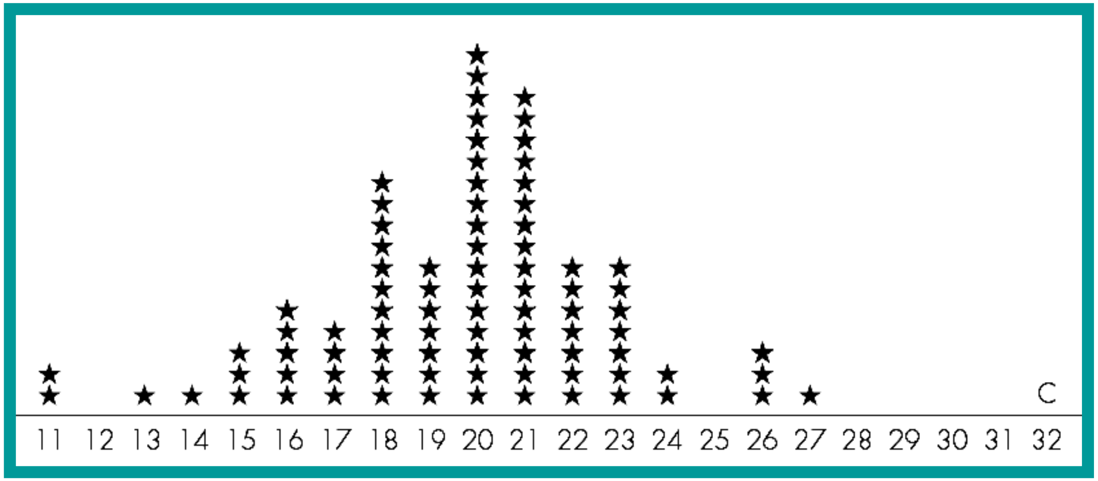
- Class of 88 students taking 40-question multiple-choice exam.
- Student C accused of copying answers from Student A.
- Of 16 questions *missed* by both A and C, both made same wrong guess on 13 of them. So they matched on 37 Q's (24 correct and 13 incorrect), and didn't match on 3 Q's.
- Prosecution argued that a match that close by chance alone is very unlikely; Student C found guilty.
- Case challenged because the Prosecution unreasonably assumed any of four wrong answers on a missed question were equally likely to be chosen.

Source: Boland and Proschan, Summer 1991, pp. 10-14.

Example, cont.: *Detecting Exam Cheating with a Histogram*

Second Trial:

For each student (except A), counted how many of his or her 40 answers matched the answers on A’s paper. Dotplot shows Student C as obvious outlier. Quite unusual for C to match A’s answers so well without some explanation other than chance.



Defense argued based on dotplot, A could have been copying from C. Guilty verdict overturned. However, Student C was seen looking at Student A’s paper – jury forgot to account for that.

2.5 More Numerical Summaries of Quantitative Data

Notation for Raw Data:

n = number of individuals in a data set

$x_1, x_2, x_3, \ldots, x_n$ represent individual raw data values

Example: A data set consists of heights for the first 6 students in the UCDavis1 dataset.

So $n = 6$, and

$x_1 = 66$, $x_2 = 64$, $x_3 = 72$, $x_4 = 68$, $x_5 = 68$, and $x_6 = 64$

Describing the “Location” of a Data Set

- **Mean:** the numerical average
- **Median:** the middle value (if *n* odd) or the average of the middle two values (*n* even)

Symmetric: mean = median
Skewed Left: usually mean < median
Skewed Right: usually mean > median

Determining the Mean and Median

The Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

where $\sum x_i$ means "add together all the values"

The Median

If n is odd: $Median$ = middle of ordered values. Count $(n + 1)/2$ down from top of ordered list.

If n is even: $Median$ = average of middle two ordered values. Average the values that are $(n/2)$ and $(n/2) + 1$ down from top of ordered list.

The Mean, Median, and Mode

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

- **Mean (numerical average): 76.04**
- **Median: 78.5 (halfway between 78 and 79)**
- **Mode (most common value): no single mode exists, many occur twice.**

The Influence of Outliers on the Mean and Median

- **Larger influence on mean** than median.
- High outliers and data skewed to the *right* will increase the mean.
- Low outliers and data skewed to the *left* will decrease the mean.

Ex: Suppose ages at death of your great-grandparents are 28, 40, 75, 78, 80, 80, 81, 82.

Mean age is 544/8 = 68 years old

Median age is (78 + 80)/2 = 79 years old

Caution: Being *Average* Isn't *Normal*

Common mistake to confuse "average" with "normal".

Is woman 5 ft. 10 in. tall 5 inches taller than normal??

Example: How much hotter than normal is normal?

> "October came in like a dragon Monday, hitting 101 degrees in Sacramento by late afternoon. That temperature tied the record high for Oct. 1 set in 1980 – and was 17 degrees *higher than normal for the date*. (Korber, 2001, italics added.)"

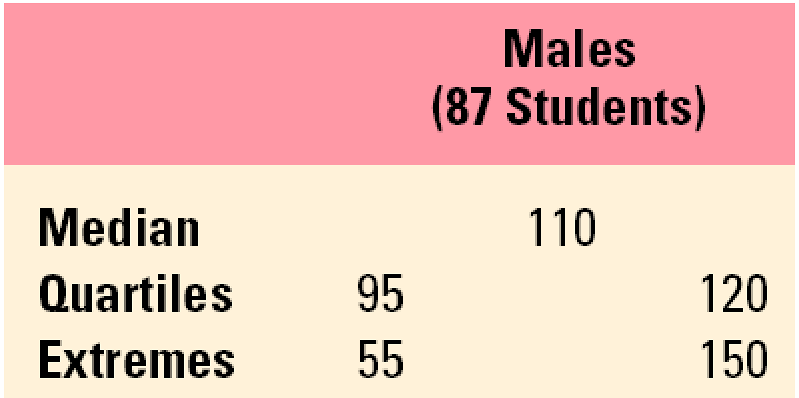
Article had thermometer showing "normal high" for the day was 84 degrees. High temperature for Oct. 1st is quite variable, from 70s to 90s. While 101 was a record high, it was not "17 degrees higher than normal" if "normal" includes the range of possibilities likely to occur on that date.

Describing Spread (Variability): Range, Interquartile Range and Standard deviation

- **Range** = high value – low value
- **Interquartile Range (IQR)** = upper quartile – lower quartile = Q_3 - Q_1 (to be defined)
- **Standard Deviation**
  (covered next time, in Section 2.7)

Example 2.13 *Fastest Speeds Ever Driven*

Five-Number Summary for 87 males



		Males (87 Students)	
Median		110	
Quartiles	95		120
Extremes	55		150

- ***Median*** = 110 mph measures the center of the data (there were many values of 110, see page 42)
- Two ***extremes*** describe spread over 100% of data
  Range = 150 – 55 = 95 mph
- Two ***quartiles*** describe spread over middle 50% of data
  Interquartile Range = 120 – 95 = 25 mph

Notation and Finding the Quartiles

Split the ordered values into the half that is (at or) below the median and the half that is (at or) above the median.

Q_1 = **lower quartile**
= median of data values
that are (at or) *below* the median

Q_3 = **upper quartile**
= median of data values
that are (at or) *above* the median

Example 2.13 *Fastest Speeds (cont)*

Ordered Data (in rows of 10 values) for the 87 males:

55	60	80	80	80	80	85	85	85	85
90	90	90	90	90	92	94	95	95	95
95	***95***	95	100	100	100	100	100	100	100
100	100	101	102	105	105	105	105	105	105
105	105	109	**110**	110	110	110	110	110	110
110	110	110	110	110	112	115	115	115	115
115	115	120	120	120	***120***	120	120	120	120
120	120	124	125	125	125	125	125	125	130
130	140	140	140	140	145	150			

- ***Median*** = (87+1)/2 = 44th value in the list = 110 mph
- ***Q_1*** = median of the 43 values below the median = (43+1)/2 = 22nd value from the start of the list = 95 mph
- ***Q_3*** = median of the 43 values above the median = (43+1)/2 = 22nd value from the end of the list = 120 mph

Percentiles

The **_k_th percentile** is a number that has *k*% of the data values at or below it and (100 – *k*)% of the data values at or above it.

- Lower quartile: 25th percentile
- Median: 50th percentile
- Upper quartile: 75th percentile

Homework (due Friday)

- Read Sections 2.4 to 2.6
- Problems in Chapter 2:

  2.42ac

  2.51

  2.61

  2.75