

# Announcements

- HW due by 6pm Fri. I will collect in class, or you can drop it in slot on wall, across from 2202 Bren Hall. Put your name AND Student ID in upper right of page. Tear off ragged edge if you use notebook paper!
- You will need a calculator for exams, but not a fancy one.
- One sheet of notes for 1<sup>st</sup> midterm, two sheets for 2<sup>nd</sup> midterm, 4 sheets for final.



---

# **Sections 2.4 to 2.6**

## **Summarizing quantitative variables**

**...including one quantitative and one categorical variable**

---

# Examples:



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)

# Clicker Practice

(to test your clicker, not to test you!)

Which discussion section are you enrolled in, or if not enrolled, would you like to be in?

- A. Section 1, meets at 3pm
- B. Section 2, meets at 4pm
- C. Section 3, meets at 5pm
- D. Section 4, meets at 6pm

# Clicker Quiz Question

- Remember you can change your answer, but the *last* answer you enter is what's recorded for you.
- Make sure the green light illuminates when you click your answer.
- I will give you a 5-second warning when time is almost up.

# Data for the first 6 students:

<b>Sex</b>	<b>Height (inches)</b>	<b>Exercise (hours/week)</b>	<b>Alcohol (drinks/week)</b>
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.**



# Asking the right questions (p. 18)

## *One Quantitative Variable*

**Question 1:** What are the interesting summary measures, like the average or the range of values, that 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 measurements similar across categories?

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

**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

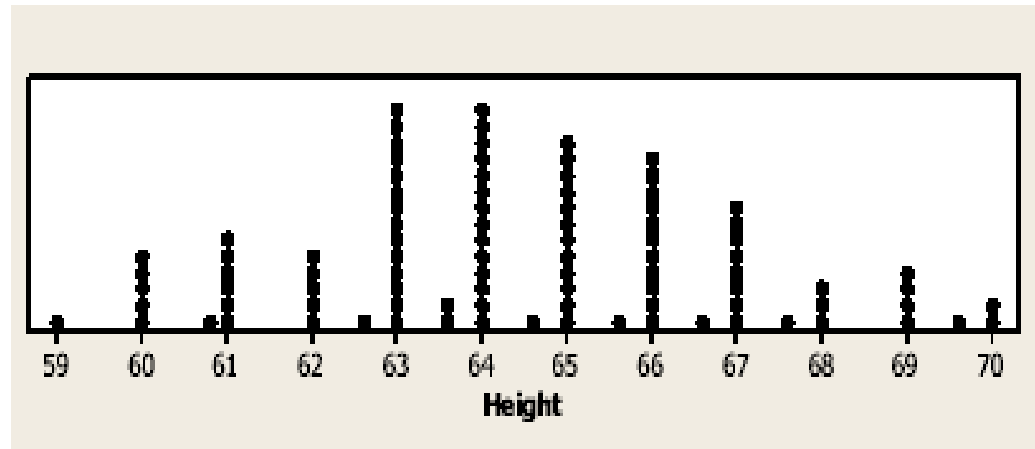


- **Histograms:** similar to bar graphs, used for *any number* of data values.
- **Stem-and-leaf plots** and **dotplots:** present *all individual values*, useful for *small to moderate* sized data sets.
- **Boxplot** or **box-and-whisker plot:** useful *summary* for *comparing* two or more groups.

# Stemplots, Dotplots and Histograms

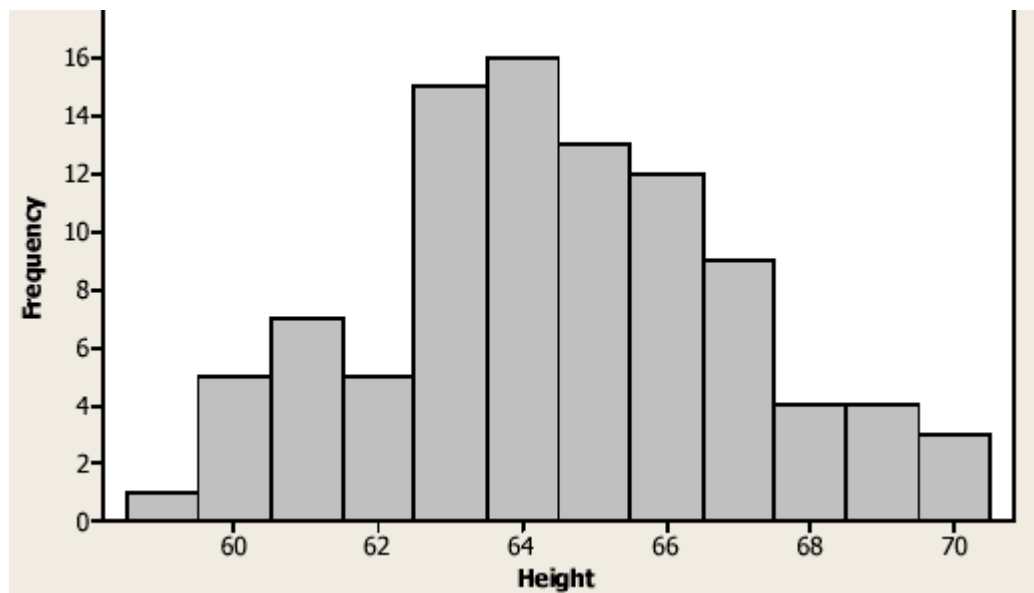
EX: Heights for 94 Females

```
|5| 9
|6| 0000001111111
|6| 22222233333333333333
|6| 444444444444444455555555555555
|6| 66666666666677777777
|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



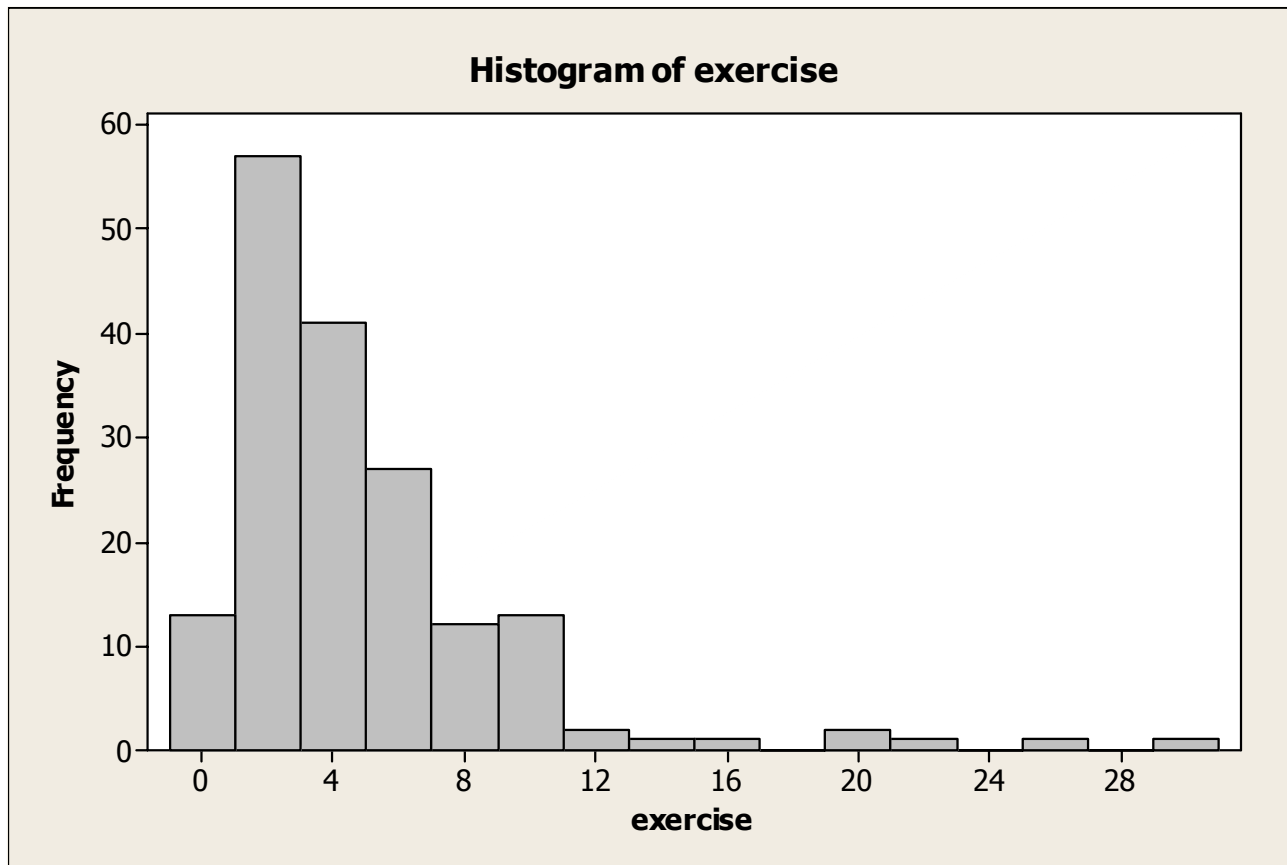
**Step 1:** Decide **how many** *equally spaced* (same width) **intervals** to use for the horizontal axis. Between 6 and 15 intervals is a good number (more if there are gaps and/or outliers).

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

**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. Decide rule to use for values that fall on the border between two intervals.

# Exercise hours per week, $n = 172$

Note that 16 intervals are used; some gaps.  
Intervals cover (-1 to .0.9), (1 to 2.9), 3 to 4.9), etc.



# Creating a Dotplot

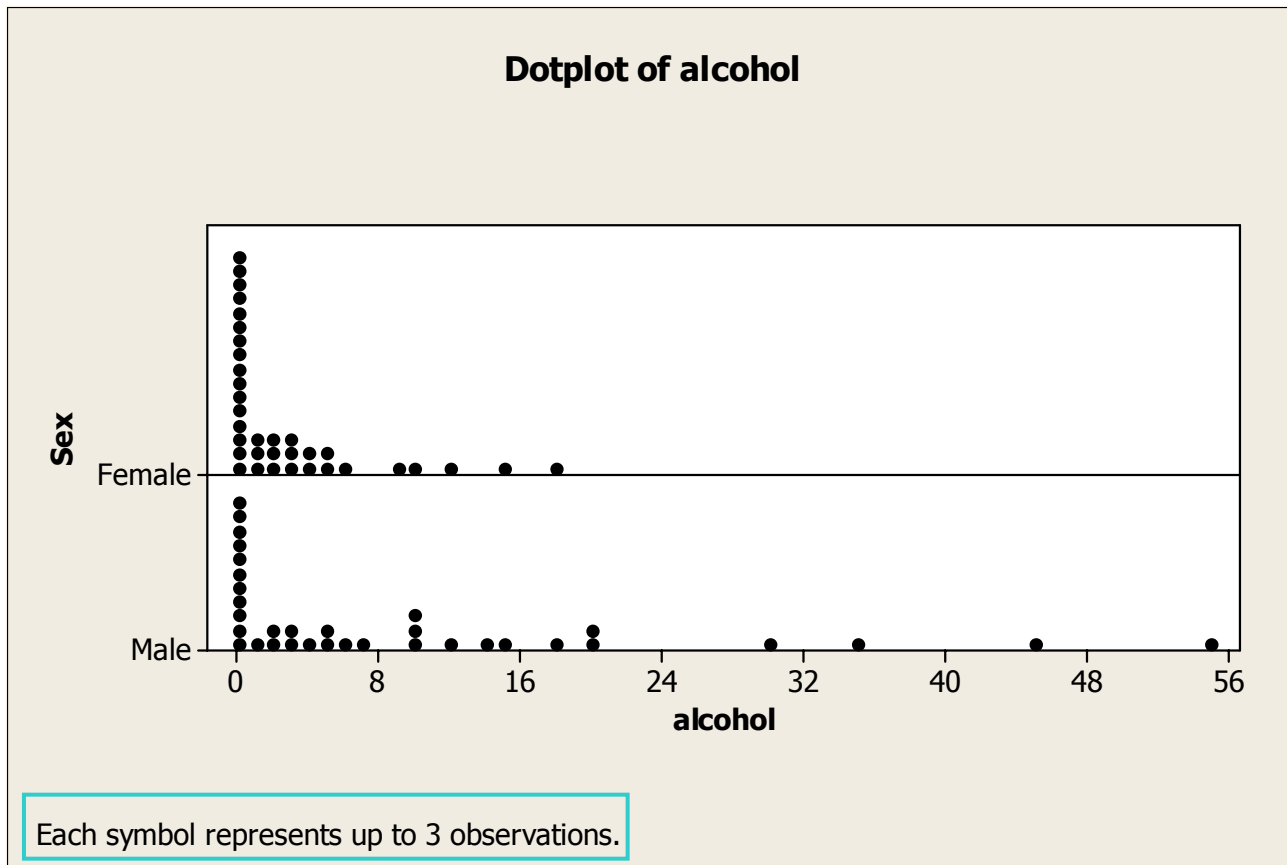


- Ideally, number line represents all possible values and there is one dot per observation.

Not always possible. From Minitab, Version 15:

- The x-axis for a dotplot is divided into many small intervals, or bins. Data values falling within each bin are represented by dots.
- If possible, Minitab displays a dot for each observation. Otherwise, a dot represents multiple observations with a footnote indicating the maximum number of observations represented by each dot.

# Useful for comparisons: Alcoholic drinks/week comparing females and males

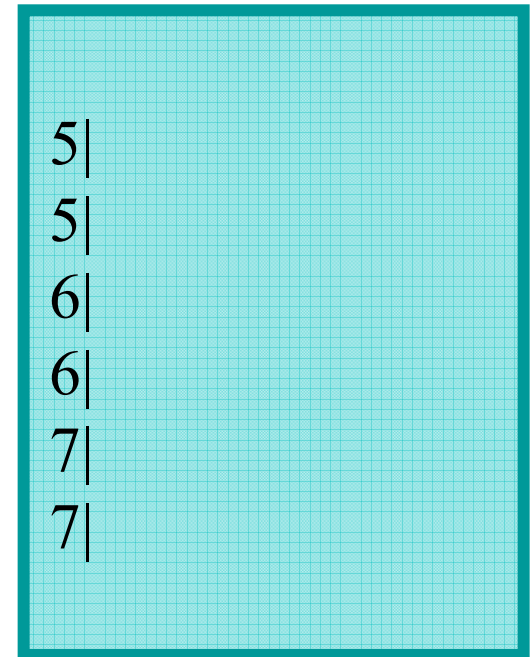


# Creating a Stemplot (stem and leaf plot) - Example of 25 pulse rates:

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 – 15 stem values, representing *equally spaced* intervals. Here, we could use 2 or 5 beats.



**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.

### Example: Pulse rates

65, 78, 60, 58, ...

First 4 value attached.

### Step 2: Attaching leaves

```
5|  
5|8  
6|0  
6|5  
7|  
7|8
```

Example:  $5|8 = 58$

**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:

1<sup>st</sup> stem = leaves 0 to 4

2<sup>nd</sup> 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:

1<sup>st</sup> stem = leaves 0 and 1

2<sup>nd</sup> 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

**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**.

Ex:  $4|5 = 450$ 's

# Describing Shape



- **Symmetric, bell-shaped**  
(Female heights bell-shaped)
- **Symmetric, not bell-shaped**
- **Bimodal:** Two prominent “peaks” (modes)
- **Skewed Right:** values clumped at left end and *extend to the right*  
(CD’s, alcohol and exercise skewed right)
- **Skewed Left:** values clumped at right end and *extend to the left*

# Example: How Much Do Students Exercise?

How many hours do you exercise a week (nearest  $\frac{1}{2}$  hr)?

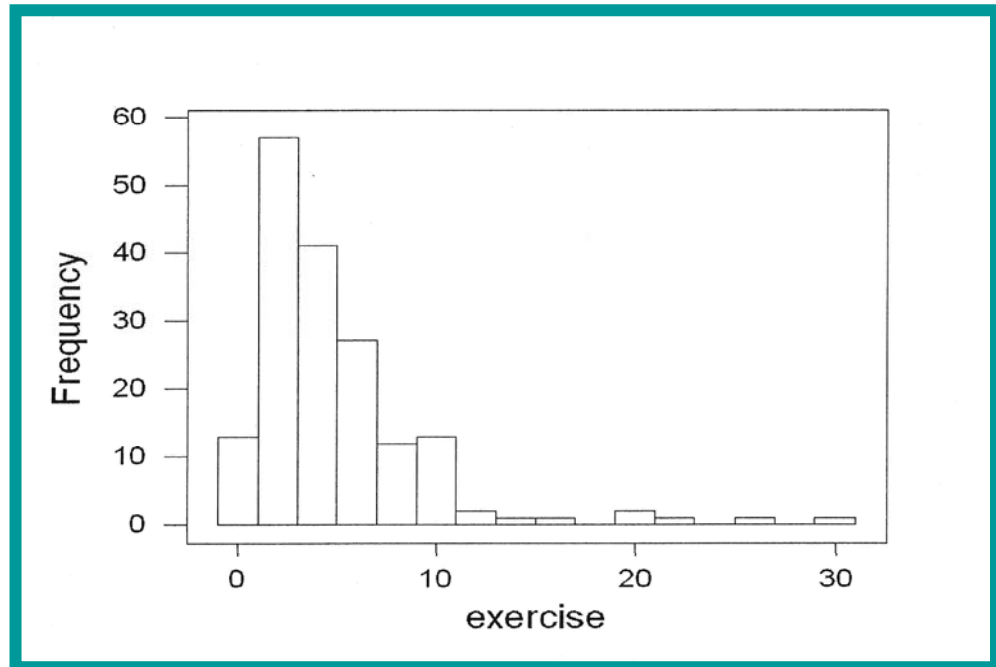
Shape is *skewed to the right*



172 responses from students in intro statistics class

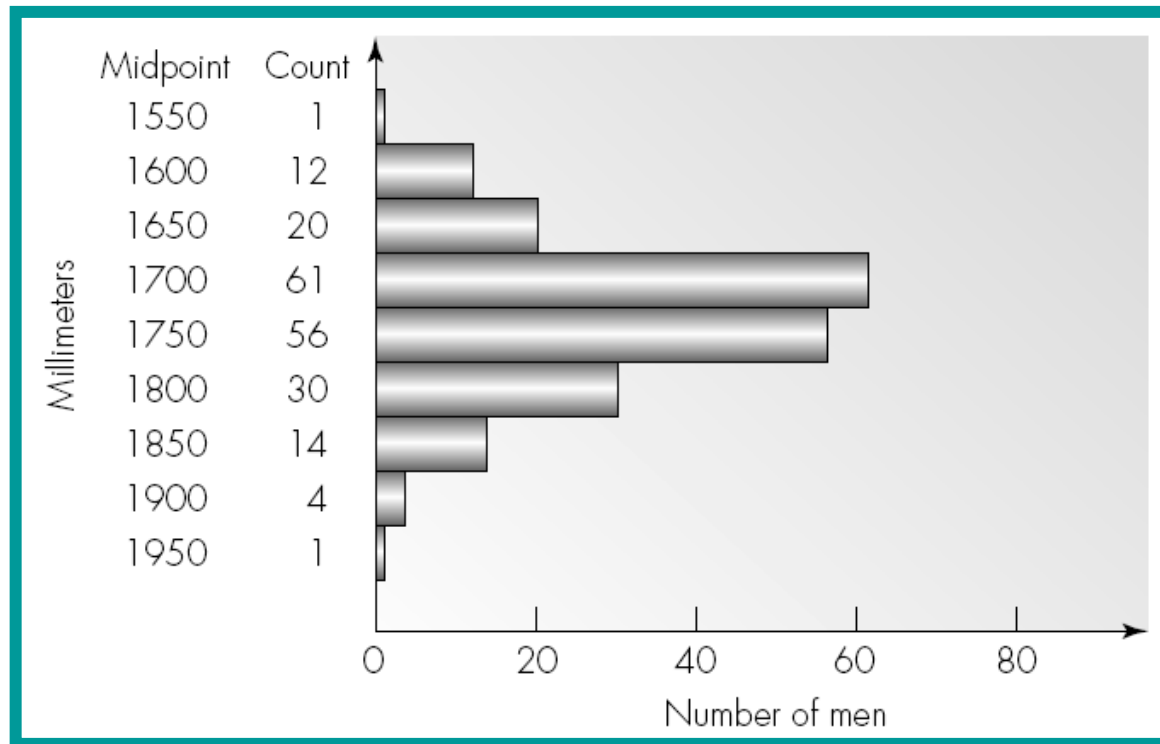
Most range from 0 to 10 hours with mode of 2 hours.

Responses trail out to 30 hours a week.



# 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.



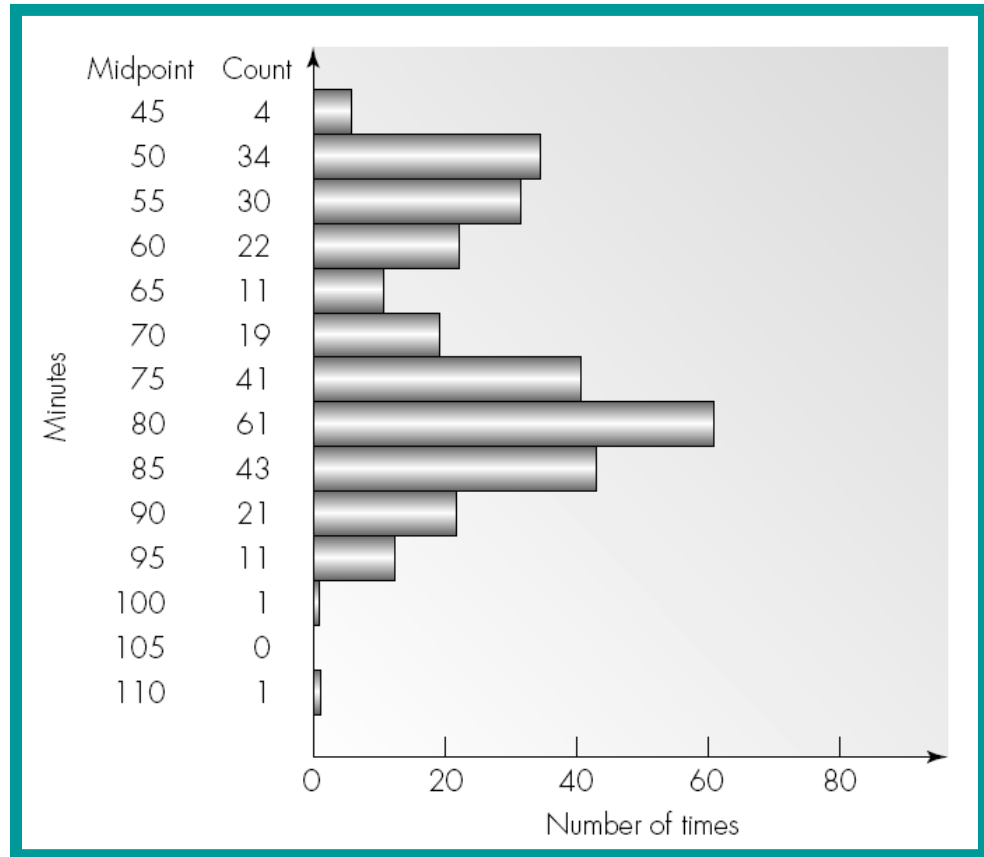
**Source: Marsh, 1988, p. 315; data reproduced in Hand et al., 1994, pp. 179-183**

# Example: The Old Faithful Geyser – time between eruptions (in book, *duration of eruptions*)



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

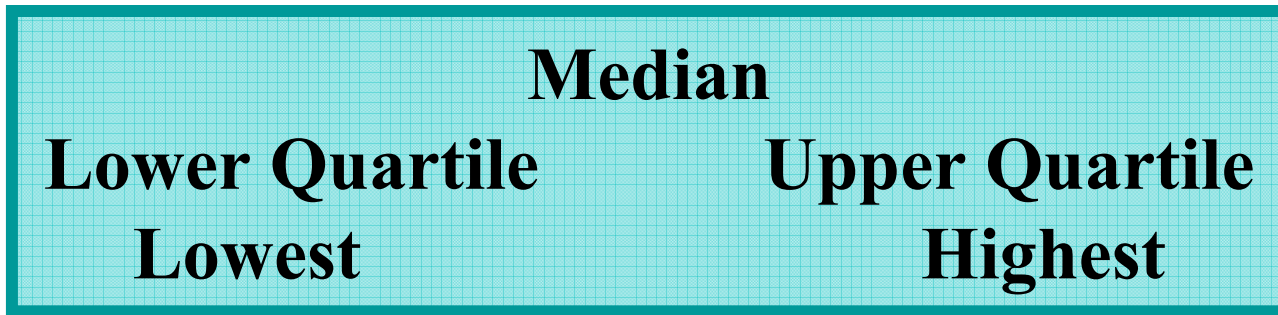
Source: Hand et al., 1994



# Five Number Summary:



## *The five-number summary display*



- **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.

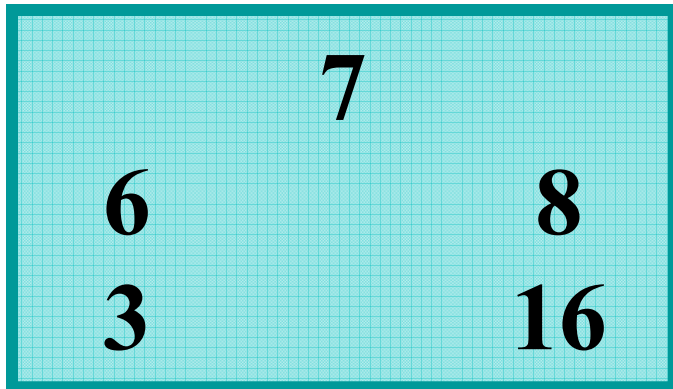
# Boxplots

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*



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

# Creating a Boxplot

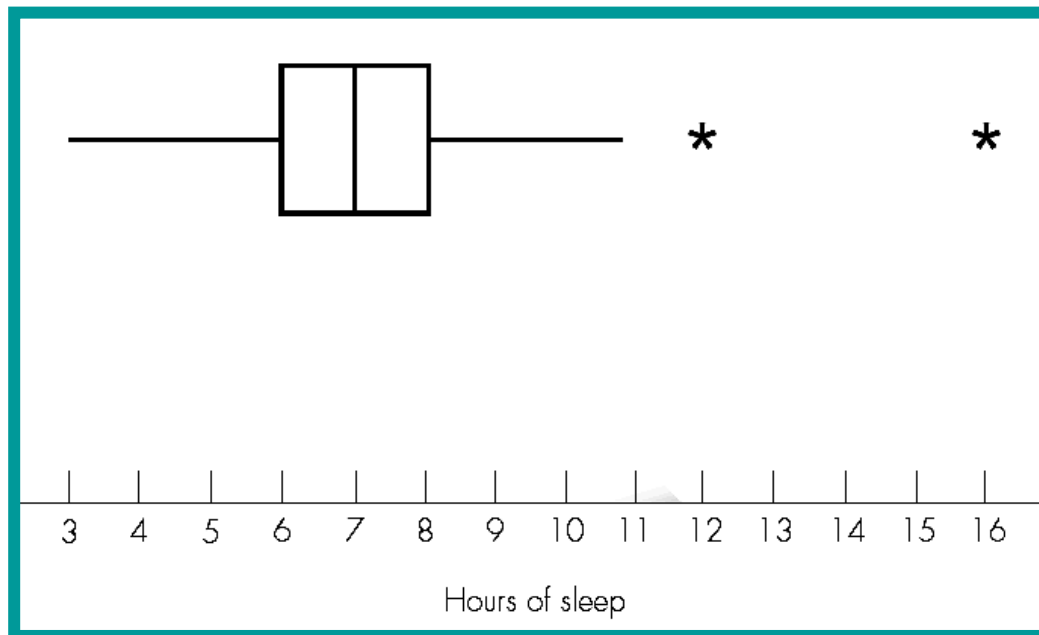


1. Draw horizontal (or vertical) line, label it with values from lowest to highest in data.
2. Draw rectangle (box) with ends at quartiles.
3. Draw line in box at value of median.
4. Compute  $IQR = \text{distance between quartiles}$ .
5. Compute  $1.5(IQR)$ ; *outlier* is any value more than this distance 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



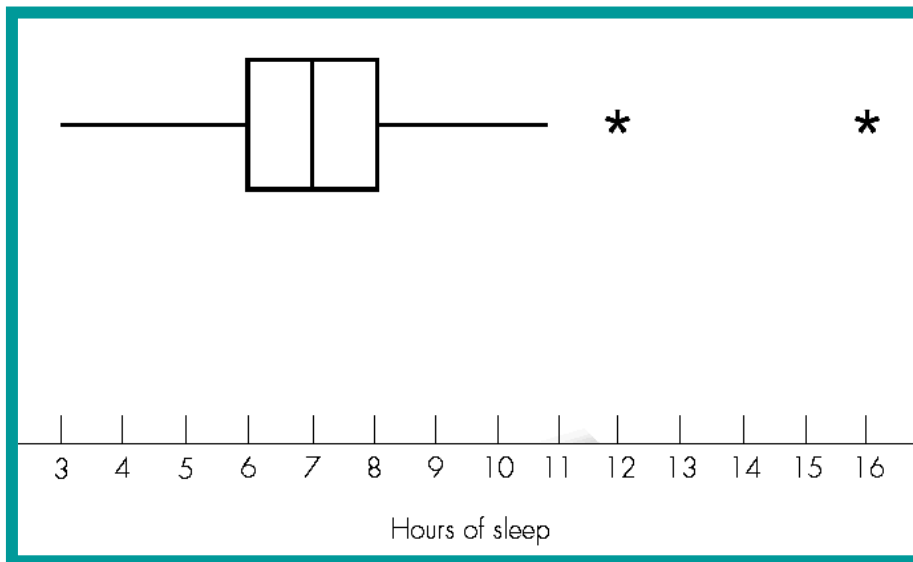
1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8.
3. Draw line in box at median of 7.
4. Compute  $IQR = 8 - 6 = 2$ .
5. Compute  $1.5(IQR) = 1.5(2) = 3$ ; outlier is any value below  $6 - 3 = 3$ , or above  $8 + 3 = 11$ .
6. Draw line from each end of box 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.



$\frac{1}{4}$  of students slept between 3 and 6 hours

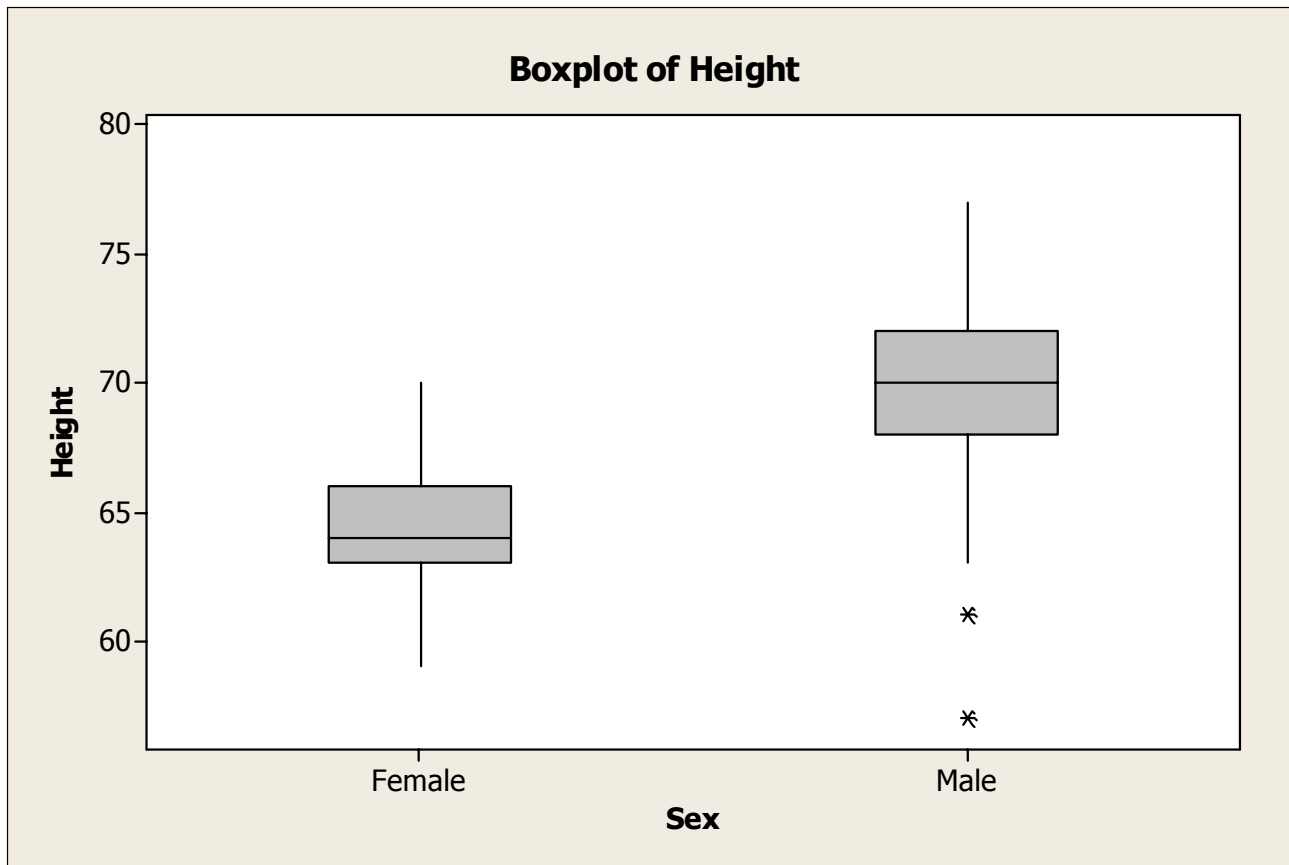
$\frac{1}{4}$  slept between 6 and 7 hours

$\frac{1}{4}$  slept between 7 and 8 hours

$\frac{1}{4}$  slept between 8 and 16 hours

# Sometimes boxplots are vertical

Ex: heights of females and males



## 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.

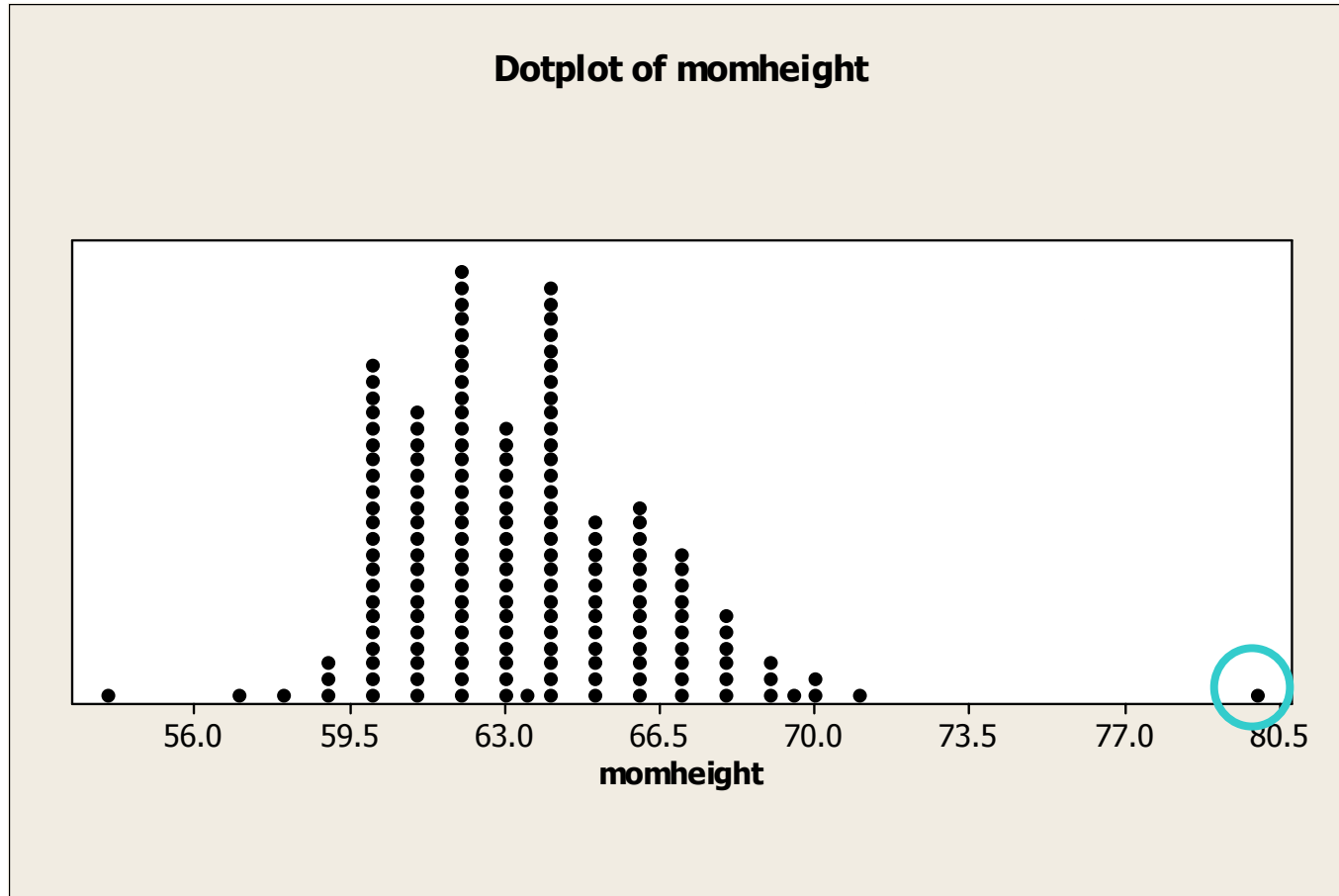
Example: 450 CDs

# Possible Reasons for Outliers and Reasonable Actions



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 not 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)

# 2.5 Numerical Summaries of Quantitative Data



## Notation for Raw Data:

$n$  = number of individuals in a data set

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

**Example:** A data set consists of heights for only the first 6 students in the UC Davis1 dataset.

Then,  $n = 6$ , and

$x_1 = 66, x_2 = 64, x_3 = 72, x_4 = 68, x_5 = 68, \text{ and } x_6 = 64$

# 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: mean < median

Skewed Right: 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

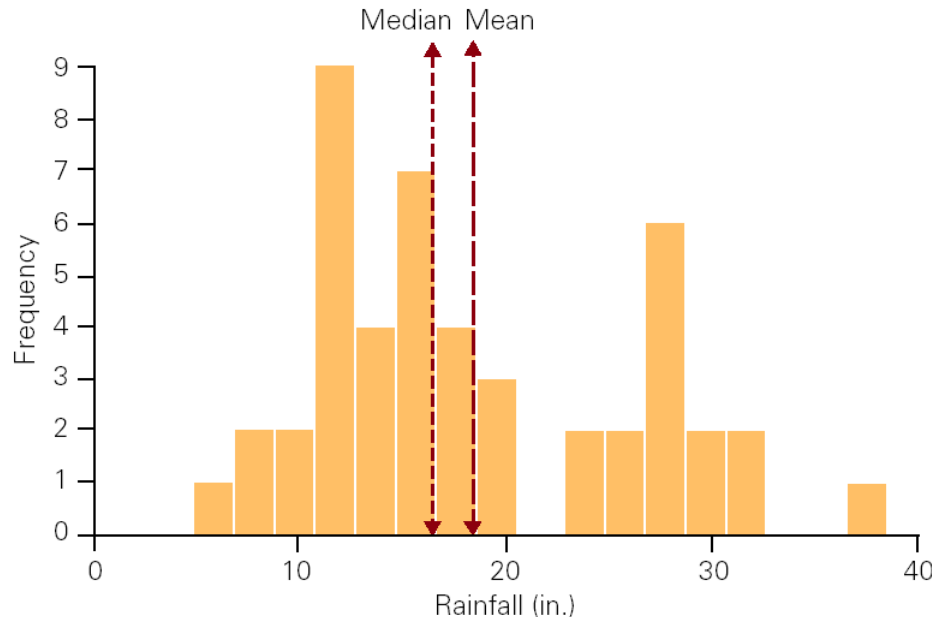
32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

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

# Example 2.11 Rainfall in Davis



**Data:** Average rainfall (inches)  
for Davis, California for 47 years



**FIGURE 2.7** Annual rainfall in Davis, California

**Mean** = 18.69 inches

**Median** = 16.72 inches

Because  $n = 47$ , median is  $(48/2) = 24^{\text{th}}$  value in *ordered* list (by values, not by year), i.e.  $24^{\text{th}}$  value from top or bottom of list. There are 23 values above it and 23 values below it.

# 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 you are taking 4 classes, and number of students in them are 20, 35, 45, 300.

**Mean** class size is  $400/4 = 100$  students

**Median** class size is  $(35+45)/2 = 40$  students

# Caution:

## Being *Average* Isn't *Normal*



Common mistake to confuse “average” with “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. 1<sup>st</sup> 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 later, 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	<b>95</b>	95	100	100	100	100	100	100	100
100	100	101	102	105	105	105	105	105	105
105	105	109	<b>110</b>	110	110	110	110	110	110
110	110	110	110	110	112	115	115	115	115
115	115	120	120	120	<b>120</b>	120	120	120	120
120	120	124	125	125	125	125	125	125	130
130	140	140	140	140	145	150			

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

# Percentiles



The  $k^{\text{th}}$  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: 25<sup>th</sup> percentile
- Median: 50<sup>th</sup> percentile
- Upper quartile: 75<sup>th</sup> percentile

# Clicker Quiz Review Question

- Get out your clickers.
- This question is a review from Friday's material.
- Most days, there will be some review questions and some based on material covered that day.

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

## *Detecting Exam Cheating with a Dotplot*



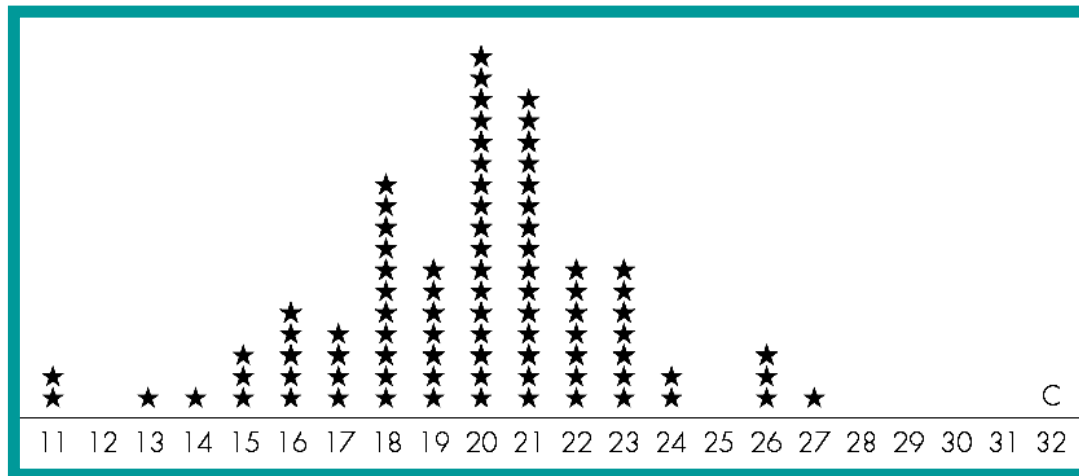
### **Details:**

- 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.
- 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.

# 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.

# Homework (due Friday)

- Read Sections 2.4 to 2.6
- Problems in Chapter 2:
  - 2.40a (p. 62)
  - 2.54 (p. 63)