# Announcements

- You may bring notes to the exams as follows: One sheet (front and back) to MT 1, 2 sheets to MT2, 4 sheets to the final exam
- Discussion Sections this Friday *are* for credit.
- We will use free software called R Commander. You need to install two programs: R and R Commander. Installation instructions are on website. We will go over it on Friday.
- For homework:
  - Do problems in the order assigned, even if order in the book is different.
  - Put name, Discussion section (1,2,3,4 or 9am,10am, etc.) 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 on due date.
- **Homework (due Mon, 1/14)**:

  2.42, 2.48bcd and five-number summary for data, 2.72

# Sections 2.4 to 2.6 Summarizing quantitative variables

**…including one quantitative and one categorical variable**

# Data used for some examples today:

Dataset "UCDavis1" on website – measured many variables on 173 students in an intro statistics 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 |

# Instructions for Clicker Questions

- 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.
- If you are on waiting list, you will still get credit for today's questions if/when you add.
- Questions will be either:
  - Review questions (on previous lecture) or
  - Questions on topics just covered
- Clicker questions are not in the posted notes

# Clicker Practice
# (to test your clicker, not to test you!)

Which of these are you?

A. Freshman
B. Sophomore
C. Junior
D. Senior
E. Super senior or other

# 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:*** Variable = Exercise hours/week

***Question:*** What is the average exercise per week for UCD students, and how much variability is there in exercise amounts?

# *One Quantitative Variable, continued*

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

***Example:*** Variable = age at death

***Question:*** 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?

So far, oldest was Jeanne Calment, a French woman who lived to be 122 years, 164 days; died 1997.

See http://anson.ucdavis.edu/~wang/calment.html

# *One Categorical and One Quantitative Variable (Comparing across categories)*

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

***Example:***

Categorical variable = sex

Quantitative variables = exercise, alcohol

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

# *One Categorical and One Quantitative Variable, continued*

***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:***

Cat. variable = Education; Quant var. = income

***Question:*** 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:**

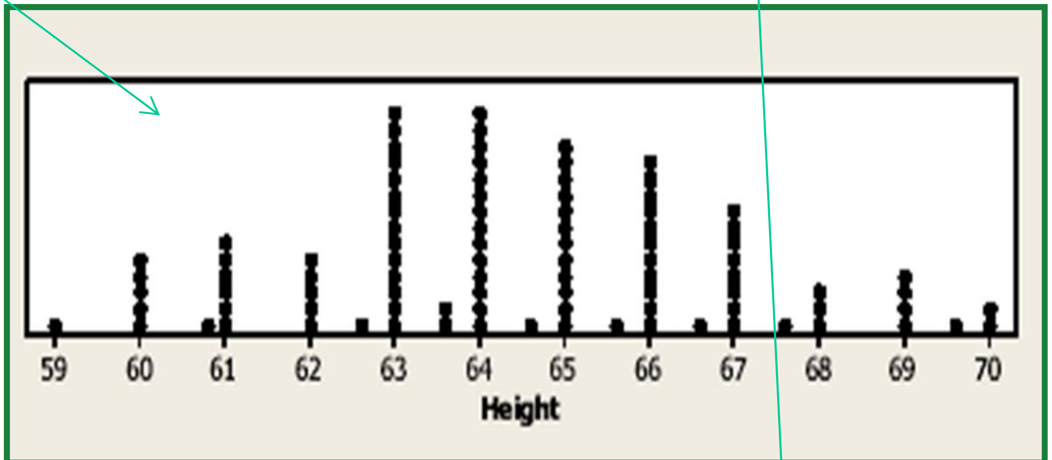
- **Histograms**: similar to bar graphs, used for *any number* of data values.
- **Stem-and-leaf plots** and **dotplots**: present *all individual values*, useful for *small to moderate* sized data sets.
- **Boxplot = box-and-whisker plot**: useful *summary* for *comparing* two or more groups.

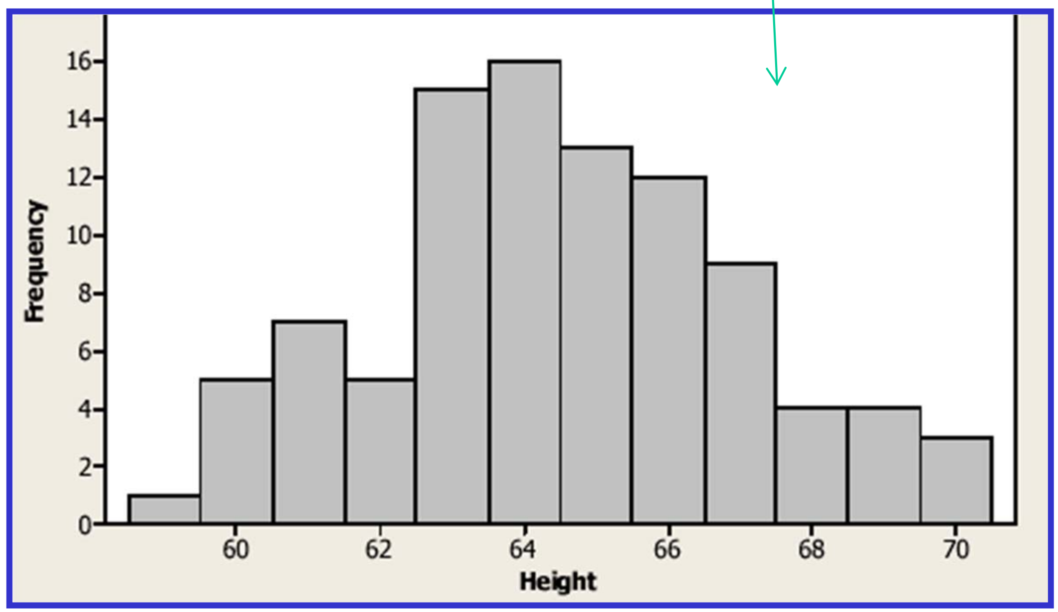
# Stemplots, Dotplots and Histograms

**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.

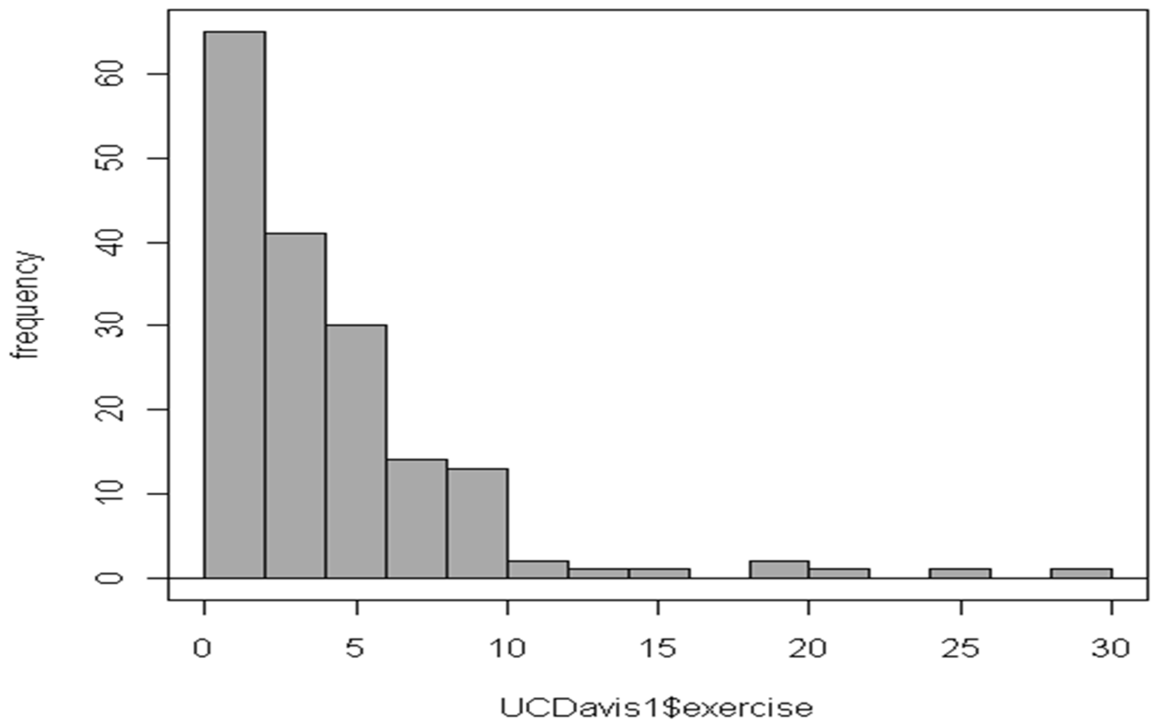
# Notes about histograms

- Intervals are equally spaced.
- 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? Need to be consistent.
- Can use *frequencies* (counts) or *relative frequencies* (proportions) as vertical axis.

*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.

# Exercise hours per week, n = 172

## Note that 15 intervals are used; some gaps.



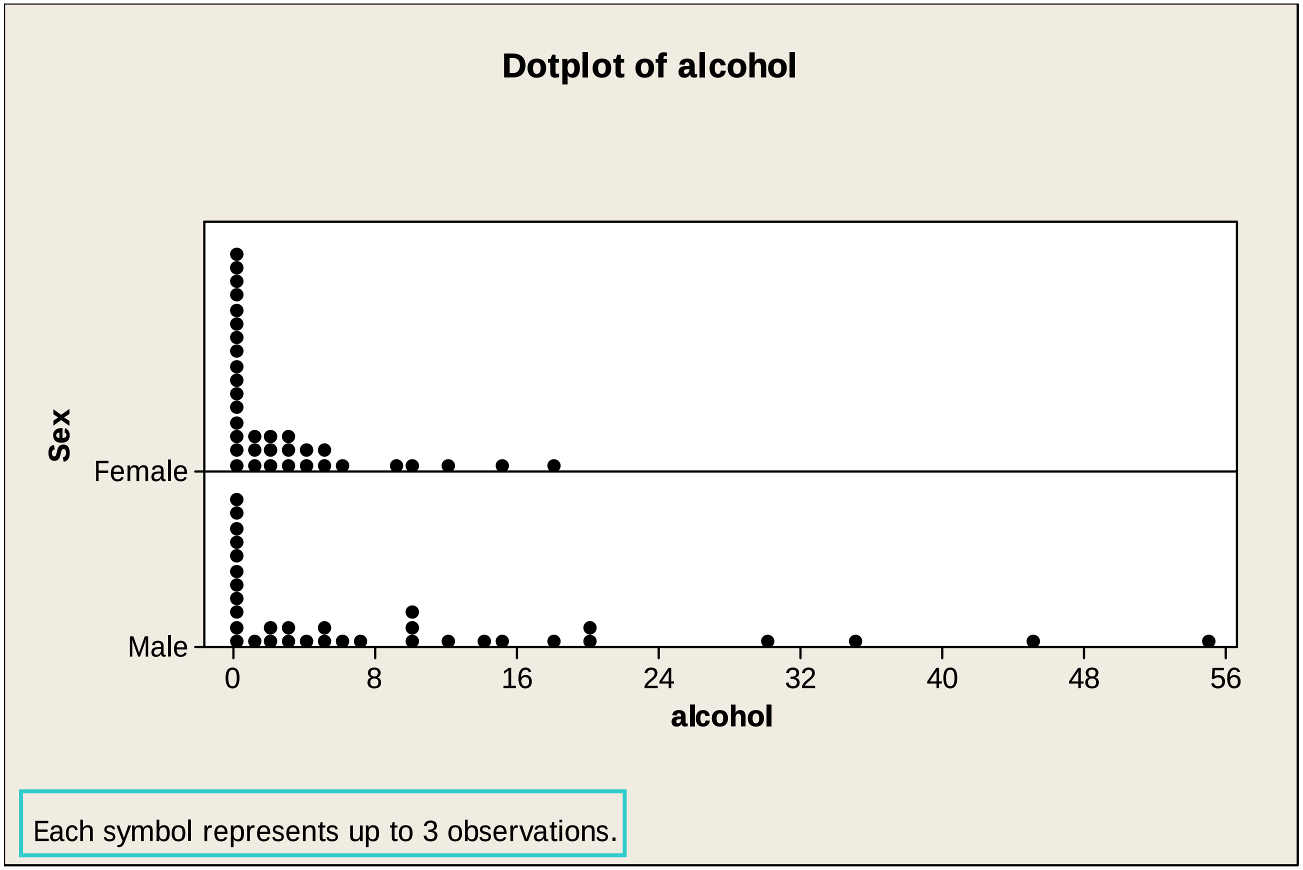

This was done using R Commander, but shows what could be done by hand.

# Creating a Dotplot

## These can be useful for comparing groups

- Saw an example last time, “fastest speed driven” comparing males and females.
- 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.

# Example: 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 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 any remaining digits.

**Example: Pulse rates, process shown in class**

**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**

**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

**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

Two times:
$1^{st}$ stem = leaves 0 to 4
$2^{nd}$ stem = leaves 5 to 9

Five times:
$1^{st}$ stem = leaves 0 and 1
$2^{nd}$ stem =leaves 2 and 3, etc.

# Example 2.9 *Big Music Collection*

## About how many songs on your iPod?

| | |
|---|---|
| 0 | 00 |
| 0 | 555667889 |
| 1 | 123 |
| 1 | 557 |
| 2 | 04 |
| 2 | 56 |
| 3 | 3 |
| 3 | |
| 4 | 0 |
| 4 | |
| 5 | 0 |

Ex: 2|4 = 2400's

**Stem** is '1000s' and **leaf unit** is '100s'. For instance, 500 is 0500, represented as 0|5

Final two digits are **truncated**.

Numbers ranged from 0 to 3305, then 4000 and 5000 as possible **outliers**.

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*

  (iPod, 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?

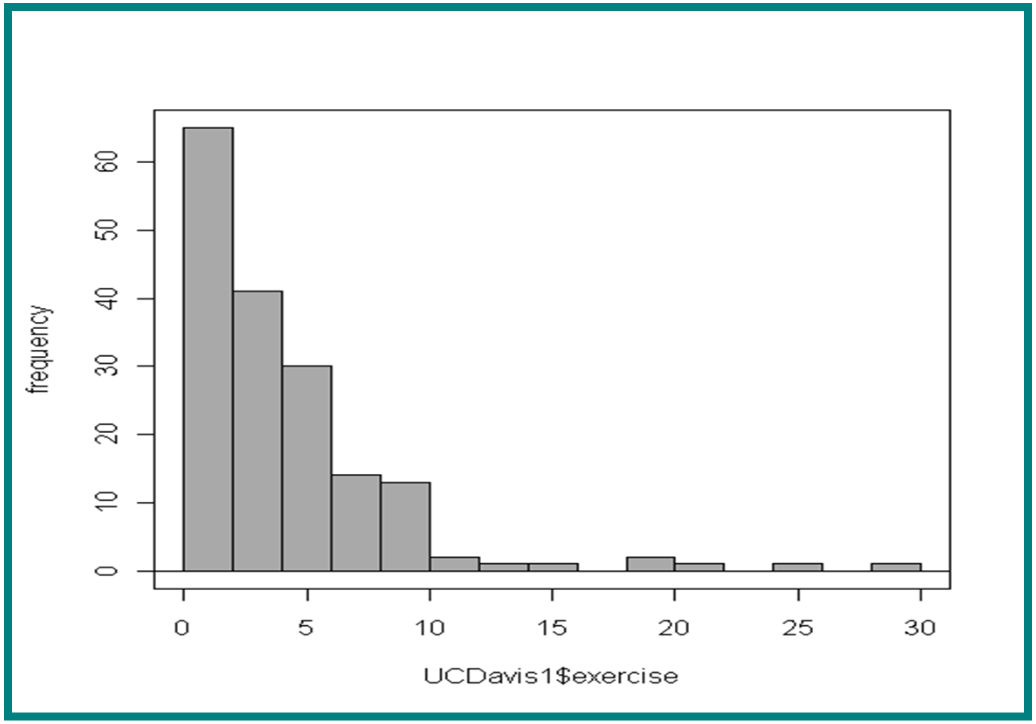
**How many hours do you exercise a week (nearest ½ hr)?**

## Shape is *skewed to the right*

172 responses from students in intro statistics class

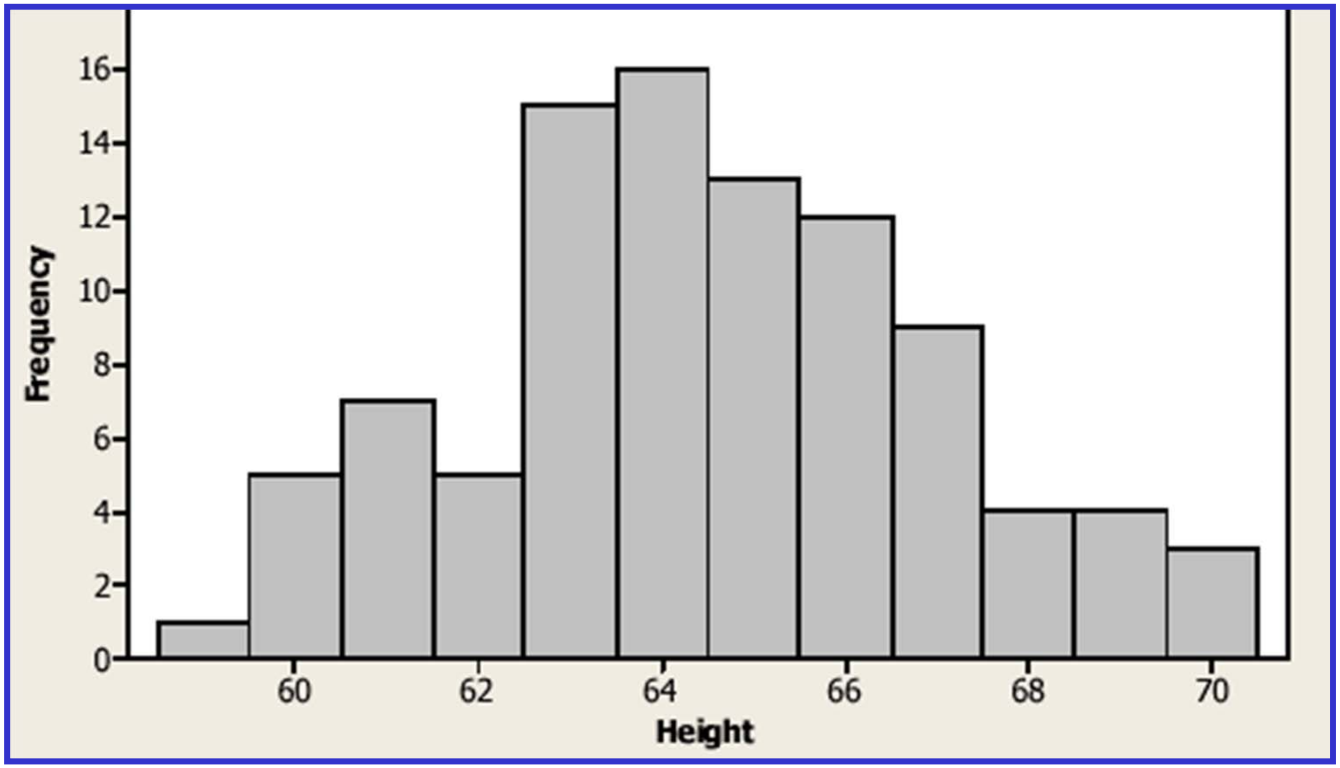
Most range from 0 to 10 hours with mode of 2 hours.

Responses trail out to 30 hours a week.



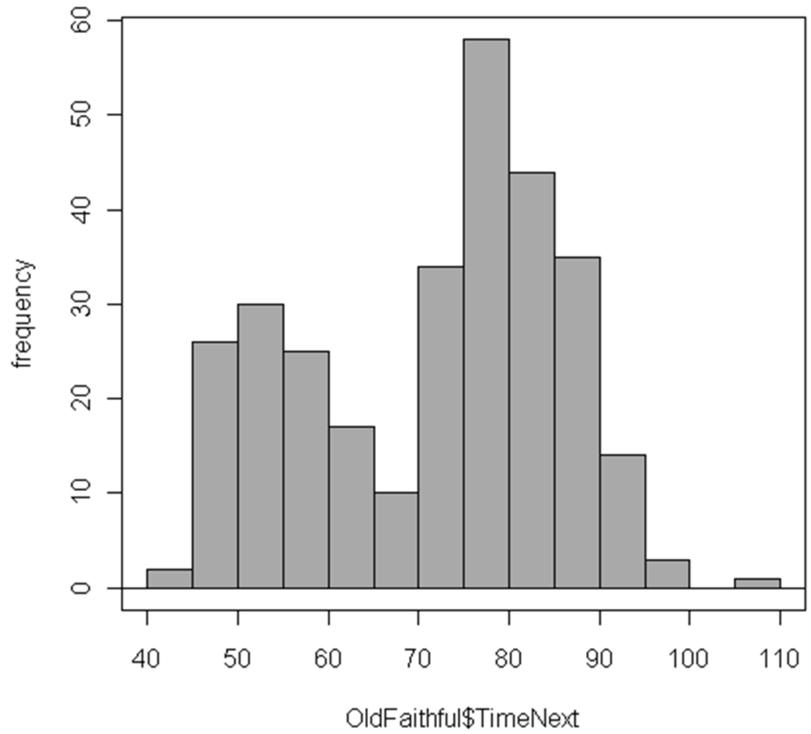
# Bell-shaped example: Women's heights

**Heights** of 94 female college students. Bell-shaped, centered around 64 or 65 inches, with no outliers.



# Bimodal Example: The Old Faithful Geyser – time *between* eruptions (in book, Fig. 2.13 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 (details of how to find these a little later)***

| | | |
|---|---|---|
| | **7** | |
| **6** | | **8** |
| **3** | | **16** |

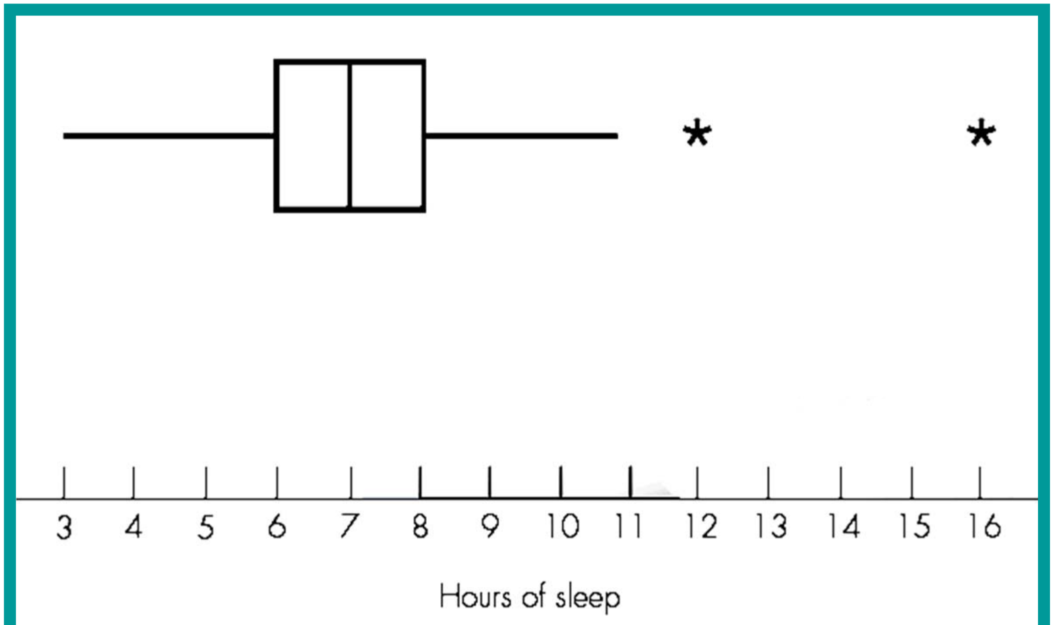
Two students reported 16 hours; the max for the remaining 188 students was 12 hours.

# 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

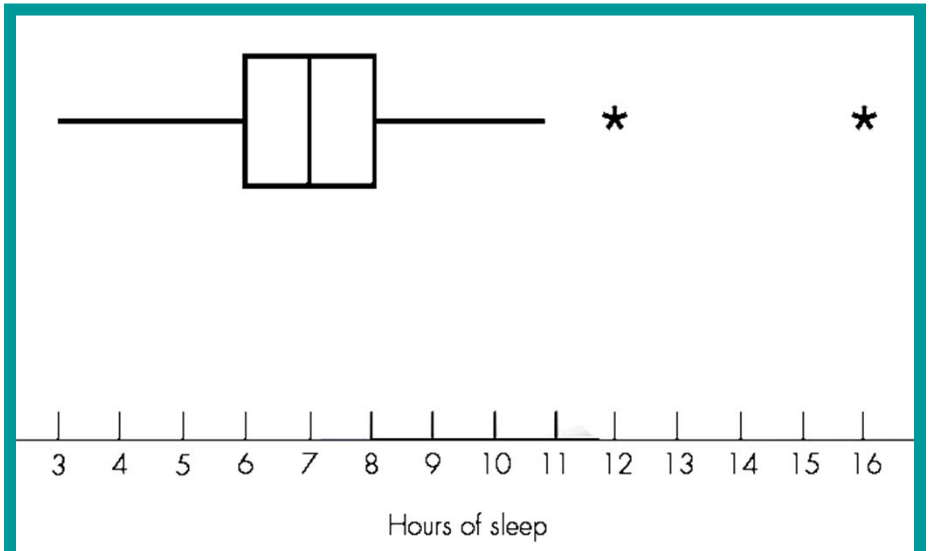
1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8 (quartiles).
3. Draw line in box at median of 7.
4. Compute IQR (interquartile range) = 8 – 6 = 2.
5. Compute 1.5(IQR) = 1.5(2) = 3; outlier is any value below 6 – 3 = 3, or above 8 + 3 = 11.
6. Draw line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



# Interpreting Boxplots

- Divides 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.



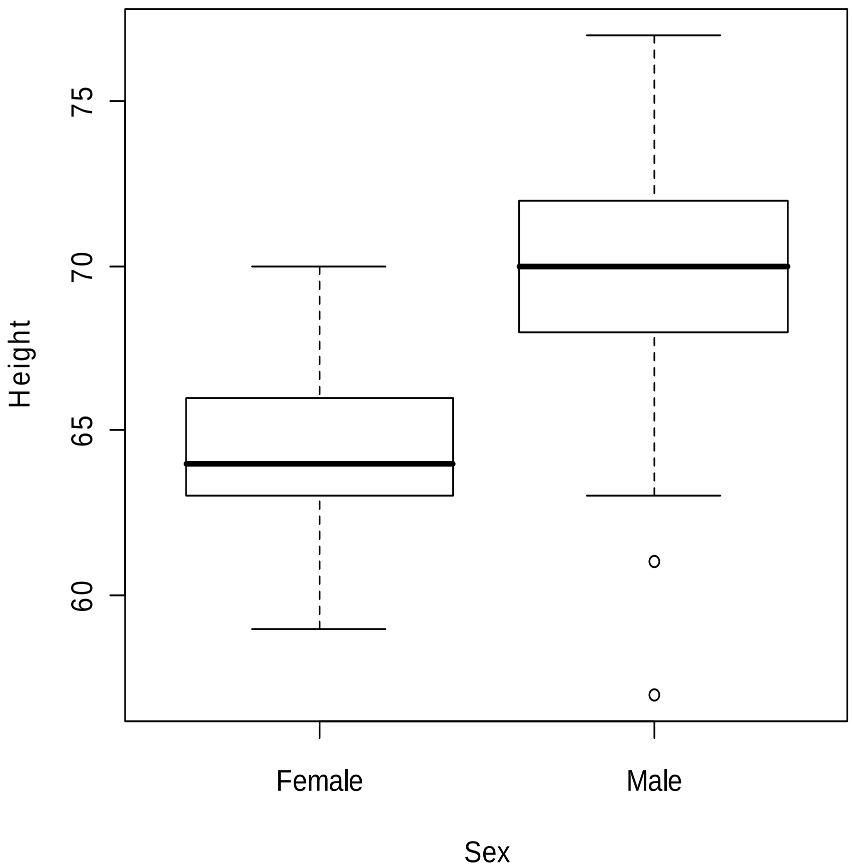

¼ of students slept between 3 and 6 hours

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

Note the two outliers for 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.

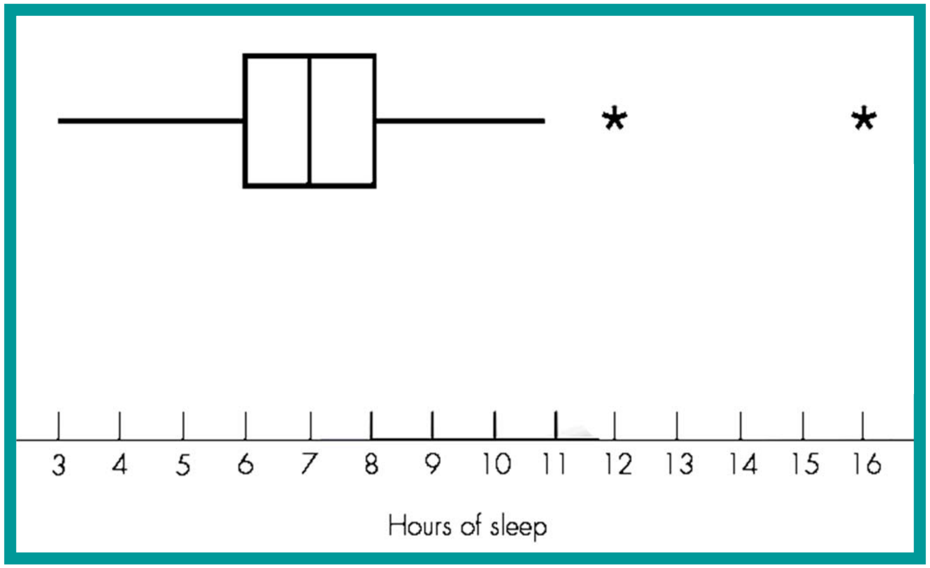
***Example***: 5000 songs on iPod

# Possible reasons for outliers and what to do about them:

1. *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.
2. *Mistake made while taking measurement or entering it into computer.* If verified, should be discarded or corrected.
3. *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.

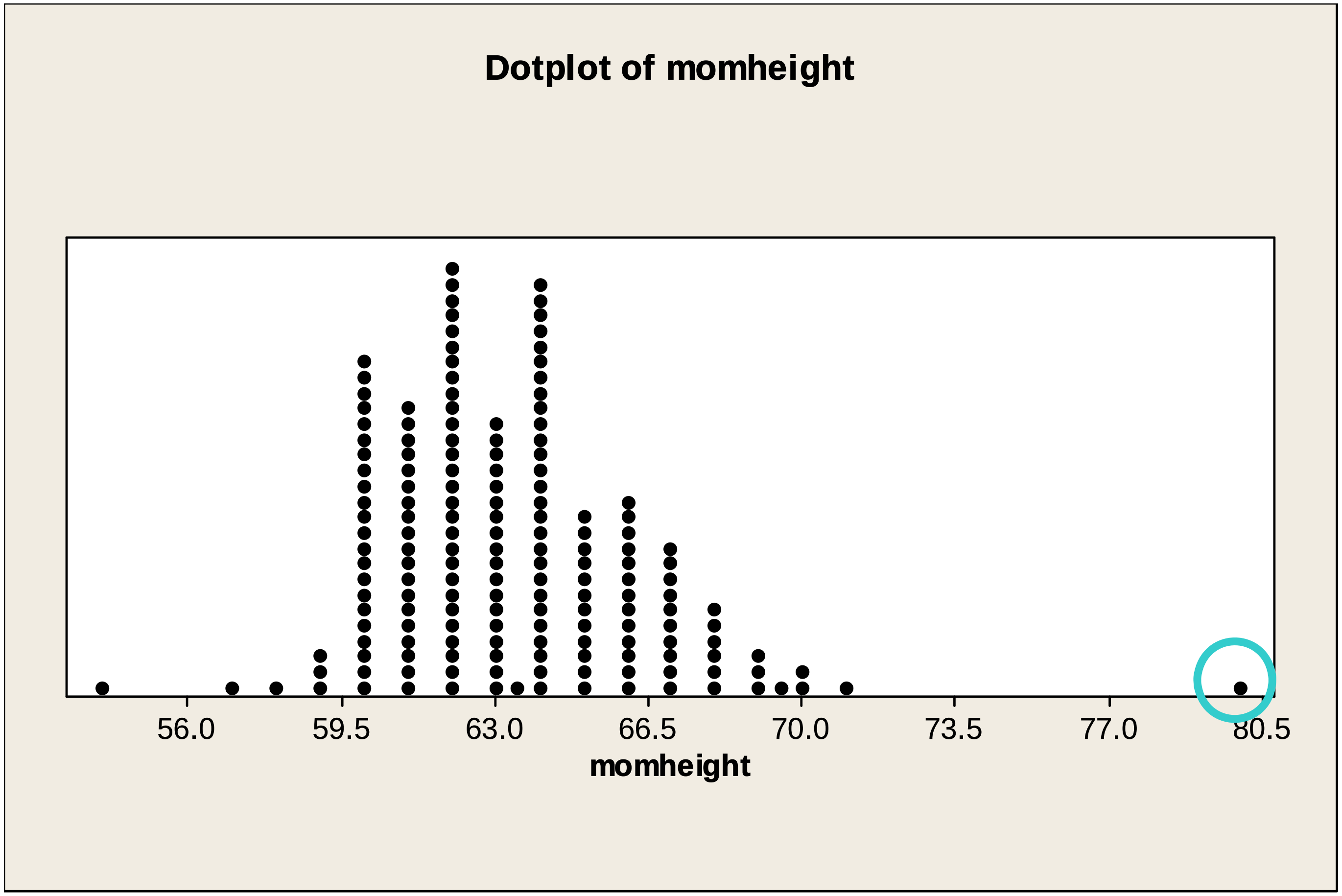
# Example: *Sleep hours*

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




**Reason 1**: Natural variability, it is *not* okay to remove these values.

# Example 2: *Students gave mother's height*




Height of 80 inches = 6 ft 8 inches, almost surely an error!

**Reason #2**, investigate and try to find error; remove value.

## **Example 2.16** ***Tiny Boatmen***

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

*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 3**: different group, okay to remove if only interested in rowers.

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

## *Detecting Exam Cheating with a Dotplot*

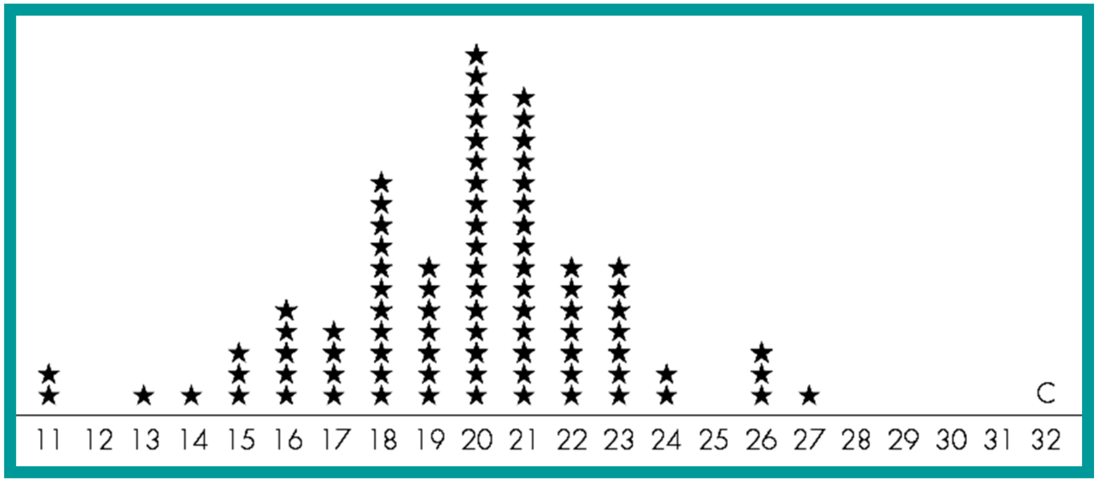
- Class of 88 students taking 40-question multiple-choice exam.
- Student C accused of copying answers from Student A.
- Of 16 questions *missed* by both A and C, both made same wrong guess on 13 of them. So they matched on 37 Q's (24 correct and 13 incorrect), and didn't match on 3 Q's.
- Prosecution 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 Dotplot*

**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 4 students in the UCDavis1 dataset.

So $n = 4$, and

$x_1 = 66, \quad x_2 = 64, \quad x_3 = 72, \quad x_4 = 68$

# 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 eight 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: *Normal* does not mean *Average*

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 at median:

- half at or below the median ("at" if ties)
- half 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 = 44$^{th}$ value in the list = 110 mph
- $\boldsymbol{Q_1}$ = median of the 43 values below the median = (43+1)/2 = 22$^{nd}$ value from the start of the list = 95 mph
- $\boldsymbol{Q_3}$ = median of the 43 values above the median = (43+1)/2 = 22$^{nd}$ 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: $25^{th}$ percentile
- Median: $50^{th}$ percentile
- Upper quartile: $75^{th}$ percentile

# Homework (due Mon, 1/14)

- Read Sections 2.4 to 2.6
- Problems in Chapter 2:
  2.42
  2.48bcd and create a 5 number summary for the data in this exercise
  2.72