Lecture 9/Chapter 7
Summarizing and Displaying Measurement (Quantitative) Data

- □Five Number Summary
- □Boxplots
- □Mean vs. Median
- □Standard Deviation

Definitions *(Review)*

Summarize values of a quantitative (measurement) variable by telling center, spread, shape.

- **Center**: measure of what is typical in the distribution of a quantitative variable
- **Spread:** measure of how much the distribution's values vary
- **Shape:** tells which values tend to be more or less common

Definitions

- **Quartiles:** measures of spread:
  - **Lower quartile** has one-fourth of data values at or below it (middle of smaller half)
  - **Upper quartile** has three-fourths of data values at or below it (middle of larger half)

  (By hand, for odd number of values, omit median to find quartiles.)
- **Interquartile range (IQR):** tells spread of middle half of data values
  = upper quartile - lower quartile

Ways to Measure Center and Spread

- **Five Number Summary:**
  1. Lowest value
  2. Lower quartile
  3. Median
  4. Upper quartile
  5. Highest value

Sometimes displayed as

#3

#2 #4

#1 #5

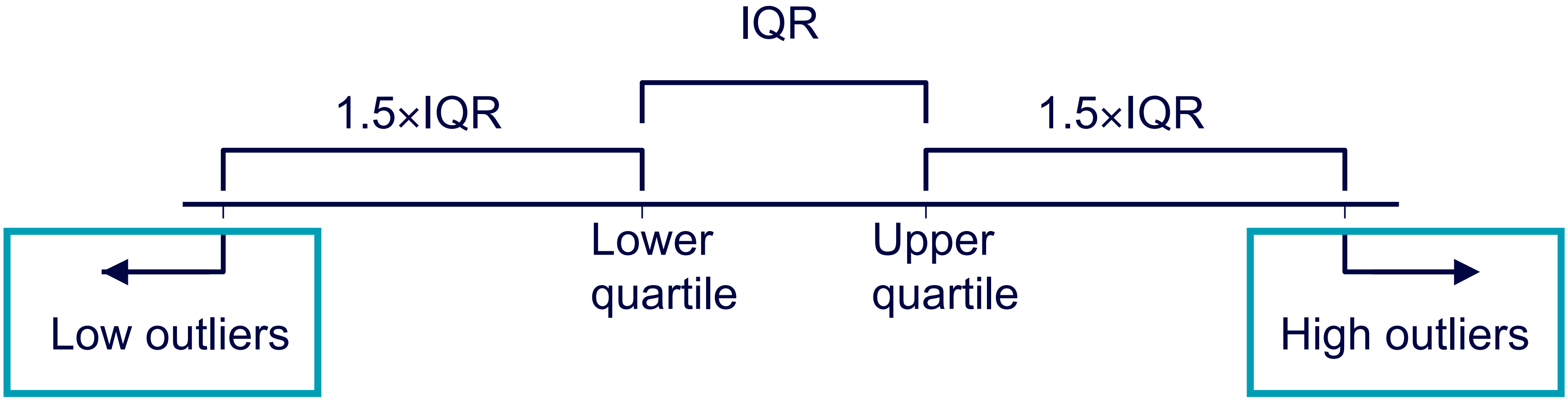
- **Mean** and **Standard Deviation**

  (we'll discuss standard deviation later)

Definition

The **1.5-Times-IQR Rule** identifies outliers:

- □ below lower quartile - 1.5(IQR) called low outlier
- □ above upper quartile +1.5(IQR) called high outlier



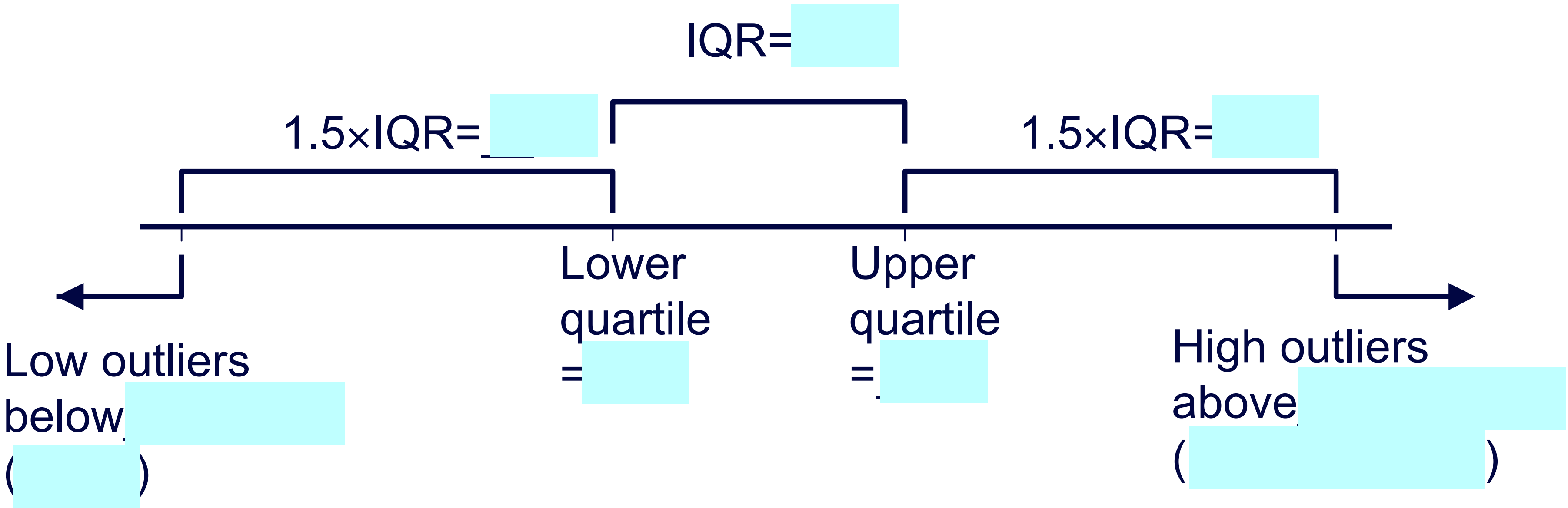
Example: *5 No. Summary, IQR, Outliers*

- **Background:** *Male earnings*

0 2 2 3 3 3 3 4 4 5 5 5 5 5 5

6 6 6 6 7 8 8 10 10 12 15 20 25 42

- **Question:** *What are 5. No. Sum. & IQR? Outliers?*
- **Response:** so IQR=



IQR=

1.5×IQR=___

1.5×IQR=

Lower quartile =

Upper quartile =_

Low outliers below

( )

High outliers above

( )

Displays of a Quantitative Variable

Displays help see the shape of the distribution.

- **Stemplot**
  - Advantage: most detail
  - Disadvantage: impractical for large data sets
- **Histogram**
  - Advantage: works well for any size data set
  - Disadvantage: some detail lost
- **Boxplot**
  - Advantage: shows outliers, makes comparisons
  - Disadvantage: much detail lost

Definition

A **boxplot** displays median, quartiles, and extreme values, with special treatment for outliers:

1. Lower whisker to lowest non-outlier
2. Bottom of box at lower quartile
3. Line through box at median
4. Top of box at upper quartile
5. Upper whisker to highest non-outlier

Outliers denoted “*”.

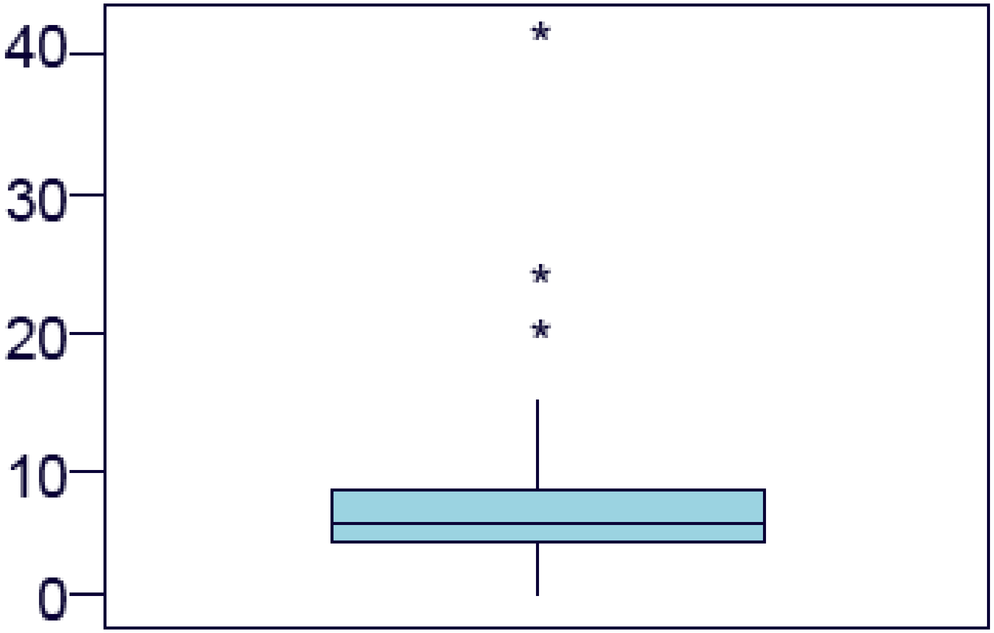
Example: *Constructing Boxplot*

- **Background**: 29 male students' earnings had 5 No. Summary: 0, 3, 5, 9, 42 and three outliers (above 18)

0 2 2 3 3 3 3 4 4 5 5 5 5 5 5
6 6 6 6 7 8 8 10 10 12 15 20 25 42

- **Question:** How do we sketch boxplot?
- **Response:**
- Lower whisker to
- Bottom of box at
- Line through box at
- Top of box at
- Upper whisker to

Outliers marked "*"



Example: *Mean vs. Median (Symmetric)*

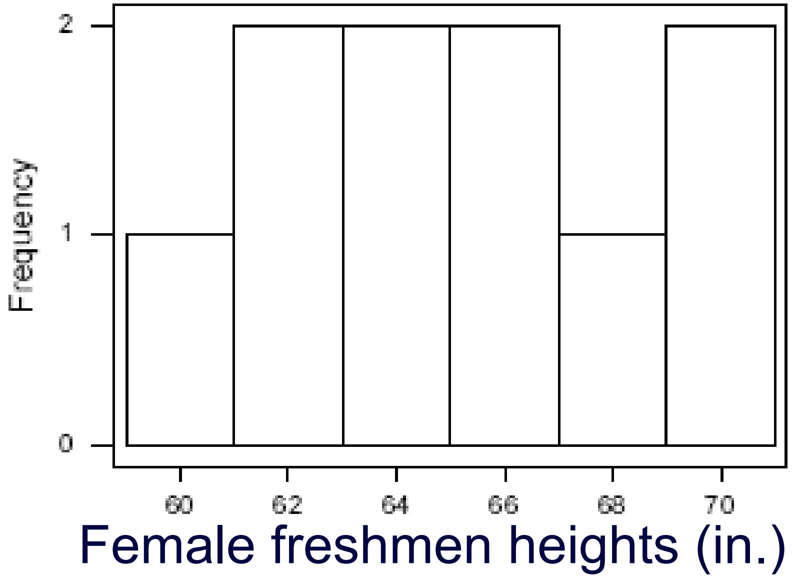
- **Background**: *Heights of 10 female freshmen:*

  59 61 62 64 64 66 66 68 70 70

- **Question:** How do mean and median compare?
- **Response:**
  - Mean =
  - Median =

  Mean_ Median.

  Note that shape is



Example: *Mean vs. Median (Skewed)*

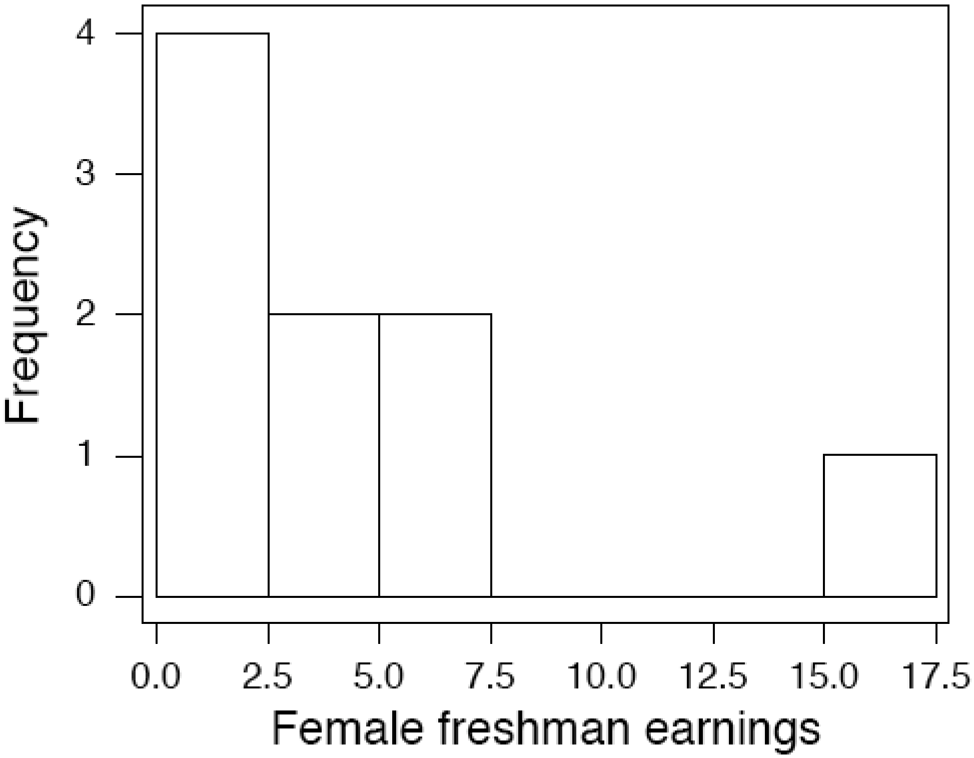
- **Background**:*Earnings ($1000) of 9 female freshmen:*

  1 2 2 2 3 4 7 7 17

- **Question:** How do mean and median compare?
- **Response:**
  - Mean = ____
  - Median =

Mean Median;

note that shape is



Mean vs. Median

- **Symmetric:**

  mean approximately equals median
- **Skewed left / low outliers:**

  mean less than median
- **Skewed right / high outliers:**

  mean greater than median
- **Pronounced skewness / outliers→**

  Report median.
- **Otherwise, in general→**

  Report mean (contains more information).

Definitions *(Review)*

Measures of Center

- **mean=**average= $\frac{\text{sum of values}}{\text{number of values}}$
- **median:**
  - ***the*** middle for odd number of values
  - average of middle two for even number of values
- **mode:** most common value

Measures of Spread

- **Range:** difference between highest & lowest
- **Standard deviation**

Definition/Interpretation

- **Standard deviation**: square root of "average" squared distance from mean.
- **Mean:** typical value
- **Standard deviation:** typical distance of values from their mean

Having a feel for how standard deviation measures spread is much more important than being able to calculate it by hand.

Example: *Guessing Standard Deviation*

- **Background:** Household size in U.S. has mean approximately 2.5 people.
- **Question:** Which is the standard deviation?

  (a) 0.014 (b) 0.14 (c) 1.4 (d) 14.0
- Response: ______

Example: *Calculating a Standard Deviation*

- □ **Background**: Female hts 59, 61, 62, 64, 64, 66, 66, 68, 70, 70
- □ **Question:** What is their standard deviation?
- □ **Response:** sq. root of "average" squared deviation from mean:

mean=65

deviations=

squared deviations=

av sq dev=

=.

Standard deviation=sq. root of "average" sq. deviation =

(This is the typical distance from the average height 65; units are inches.)

Example: *Calculating another Standard Deviation*

- **Background**: Female earnings 1, 2, 2, 2, 3, 4, 7, 7, 17
- **Question:** What is their standard deviation?
- **Response:** sq. root of “average” squared deviation from mean:

mean=5

deviations=

squared deviations=

av sq dev=(

=

standard deviation=sq. root of “average” sq. deviation =

Is this really the typical distance from the typical earnings?

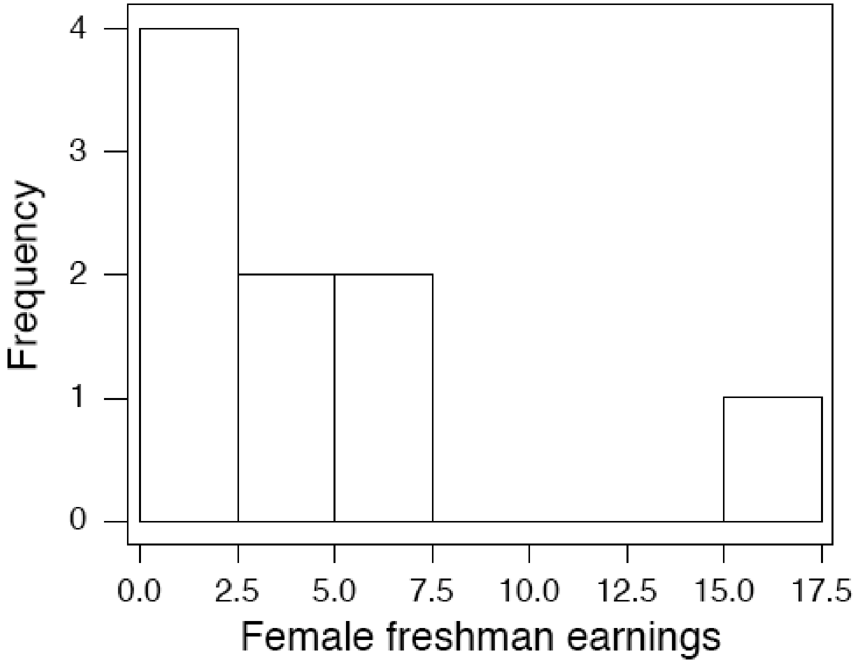
Example: *Calculating another Standard Deviation*

- **Response:** mean=5, standard deviation=5

Is 5 thousand really typical for earnings?

Is 5 thousand really typical distance of earnings from average?

Two thirds earned K or less; all but one were within K of 4 K. If the outlier 17 is omitted, mean= sd=



The mean and, to an even greater extent, the standard deviation are distorted by outliers or skewness in a distribution. Although they are not ideal summaries for such distributions, we will see later that the normal distribution actually applies if we take a large enough sample from a non-normal population and use inference to draw conclusions about the population mean or proportion, based on our sample mean or proportion. We will begin to study the normal curve next (Chapter 8).

EXTRA CREDIT (Max. 5 pts.) Summarize data for a survey variable; include mention of center, spread, and shape, and at least 2 of the 3 displays (stemplot, histogram, boxplot). Survey data is linked from my Stat 800 website www.pitt.edu/~nancyp/stat-0800/index.html and MINITAB can be used in any Pitt computer lab to produce displays and summaries. Alternatively, you can process the data by hand.