Interpreting the Standard Deviation

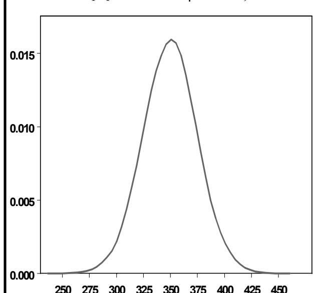
- Given two samples from a population, the sample with the larger standard deviation (SD) is the more variable
  - Say we have $s_x = 21.4; \quad s_y = 29.6$
- We are using the SD as a relative or comparative measure—*Y* is …?
- How does the SD provide a measure of variability for a single sample or, what does 29.6 really mean?

The Empirical Rule

A rule of thumb that applies to data sets that have a mound shaped, symmetric distribution

- Approximately 68% of the measurements will fall within 1 SD of the mean
- Approximately 95% of the measurements will fall within 2 SDs of the mean
- Approximately 99.7% of the measurements will fall within 3 SDs of the mean

Distribution for measurements from a normal population with $\mu = 350; \quad \sigma = 25$



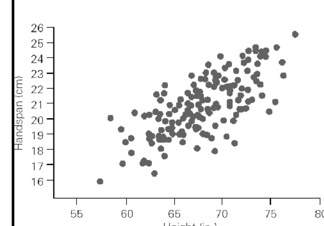
Application of Empirical Rule for Mound-Shaped Distributions of Data (continued)

$$(\mu-\sigma, \mu+\sigma) = (350-25,\ 350+25) = (325, 375)$$
$$(\mu-2\sigma, \mu+2\sigma) = (350-50,\ 350+50) = (300, 400)$$
$$(\mu-3\sigma, \mu+3\sigma) = (350-75,\ 350+75) = (275, 425)$$

Relationships Between Quantitative Variables

- **Scatterplot**, a two-dimensional graph of data values.
- Use a ***scatterplot*** to look at the relationship between two quantitative variables
- Plot has one variable's values along the vertical axis and the other variable's values along the horizontal axis
- **Correlation**, a statistic that measures the *strength* and *direction* of a linear relationship
- **Regression equation**, an equation that describes the average relationship between a response and explanatory variable---we will not get to this

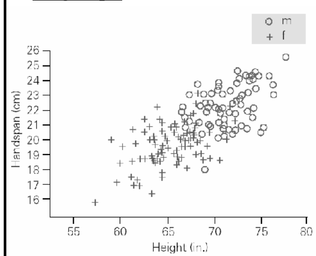
heights (in inches) and fully stretched handspans (in centimeters) of 167 college students.



Questions that might be asked

- What is the *average* pattern? Does it look like a straight line or is it curved?
- What is the direction of the pattern?
- How much do individual points vary from the average pattern?
- Are there any unusual data points?

Use different plotting symbols or colors to represent different subgroups.



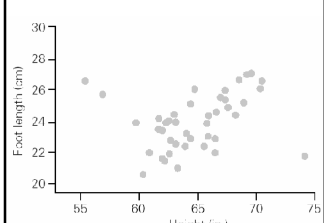
Positive/Negative Association

- Two variables have a **positive association** when the values of one variable tend to increase as the values of the other variable increase.
- Two variables have a **negative association** when the values of one variable tend to decrease as the values of the other variable increase.

For the handspan/height data

- Taller people tend to have greater handspan measurements than shorter people do (**positive association**)
- The handspan and height measurements
  may have a **linear relationship**.

Look for outliers: points that have an usual combination of data values.



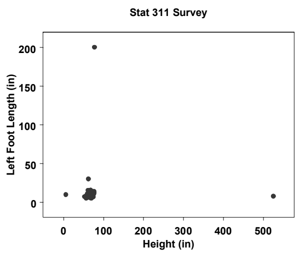
Outliers

Outlier--- an unusually large or small measurement relative to the other observations

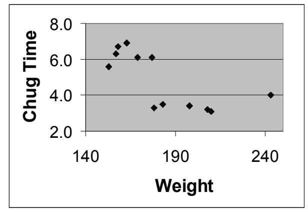
Common causes:

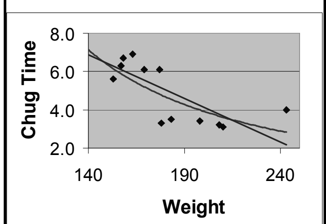
- Measurement incorrectly observed or recorded (including data entry)
- Measurement comes from a different population
- Measurement is correct, but represents a rare event

heights (in inches) and fully left foot length (in inches) of Stat 311 students



Body weights and the time it takes to chug a 12-ounce beverage for n=13 college students. The data were submitted by a student for a class project. (*Source:* William Harkness, Pennsylvania State University)





Methods for Detecting Outliers

1.5 IQR Rule* and *Box plots

- Based on quartiles of a data set
- ***Quartiles*** partition the data set into 4 groups, each containing 25% of the measurements
- The lower quartile, Q_1 , is the 25th percentile; the middle quartile, M, is the median (50th percentile); the upper quartile, Q_3 , is the 75th percentile

Methods for Detecting Outliers

Example: sample of 5000 data values from the normal population with $\mu = 350$; σ=25

```
R summary output
   Min.:257.0
1st Qu.:333.3
 Median:350.0
   Mean:349.8
3rd Qu.:366.0
   Max.:439.9
```


Methods for Detecting Outliers

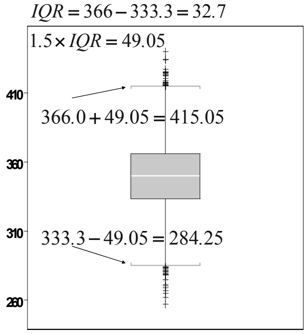
Interquartile Range (IQR)--- distance between the upper and lower quartiles

$$IQR = Q_3 - Q_1$$

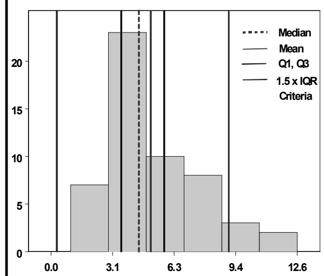
Lower fence = $Q_1 - (1.5 \times \text{IQR})$
Upper fence = $Q_3 + (1.5 \times \text{IQR})$

Methods for Detecting Outliers

$$IQR = 366 - 333.3 = 32.7$$



Methods for Detecting Outliers



Methods for Detecting Outliers