ST 311 **Chapter 7** *Reiland*

Chapter 7

Scatterplots and Correlation

"The cause of lightning," Alice said very decidely, for she felt quite sure about this, "is the thunder—no, no!" she hastily corrected herself. "I meant it the other way." *Lewis Carroll,* Alice in Wonderland

"Toots Shor's restaurant is so crowded, nobody goes there anymore." Yogi Berra
former NY Yankees catcher

IN THE REAL WORLD (*i*) **Latin's Remarkable Resurrection** (New York) – A real resurrection of Latin is taking place. The upsurge at the high school level is very much a result of student demand. Why are students flocking to Latin, and what do they expect to get out of it? There is, allegedly, a clear correlation between taking Latin at school and doing well on scholastic aptitude exams, and the students are well aware, as they always are, of the statistics. They may not see why Latin, in this high-tech age, should help them reach their goal, but they are clear in the fact that indeed it does.

(*ii*) **Standardized Test Prep Courses** Preparation courses for standardized achievement tests such as the SAT, LSAT, and GMAT are big business. The median score on the Law School Admissions Test (LSAT) is 150 and the quartiles are approximately (145, 160). An LSAT prep course enrolls a group of people who scored 135 the first time they took the test. On their second test their average score increased to 150. The people who run the prep course trumpet the benefits of their course and claim that they have added an average of 15 points to the scores of people in this group. Does the 150 average support the claim made by the prep course administrators?

(iii) **Rating NFL and NCAA Quarterbacks** In 1973 the National Football League arrived at a formula for rating quarterbacks. The formula is based on the percentage of passes completed, average yards gained per completion, percent of passes resulting in touchdowns, and the percent of passes intercepted. The single season record is **Peyton Manning**'s 121.1 rating during the 2004 season. The minimum and maximum possible ratings are 0 and 158.3, respectively. The complexity of the NFL formula resulted in ESPN's 2011 introduction of the Total QBR rating system that ranges between 0 and 100. Since 1979 the NCAA has used a rating system that has a minimum of – 731.6 (if every pass is interecepted for a 99 yard loss) and a maximum of 1,261.6 (if every pass is a 99 yard touchdown pass). In 2011 **Robert Griffin, III** (192.3, Baylor) and **Russell Wilson** (191.6, Wisc) earned the two highest NCAA single-season quarterback ratings of all time. The freshman record belongs to Michael Vick of Virginia Tech, whose rating during the 1999 season was 180.4

Chapter Objectives:

At the end of this chapter you should be able to:

1) create a scatterplot to graphically depict the realtionship between 2 quantitative variables
2) describe the information that a scatterplot conveys about the relationship between 2 quantitative variables: form, direction, strength, points that depart from the overall pattern.
3) calculate the correlation coefficient between 2 quantitative variables using technology.
4) interpret the value of the correlation coefficient
5) describe when it is appropriate to use the correlation to describe the relationship between 2 quantitative variables
6) list the properties of the correlation coefficient
7) apply the properties of the correlation coefficient to determine the correlation when the units of the original variables are changed
8) describe the difference between association, correlation and cause-and-effect.

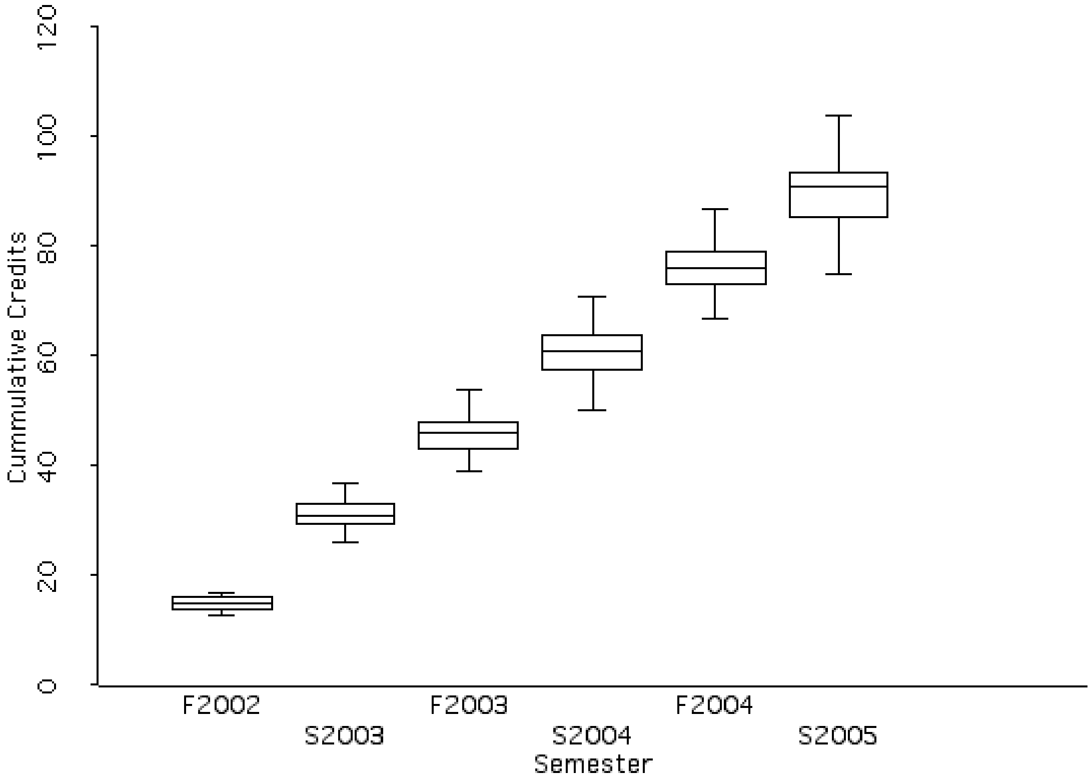
Motivation

1. Autism linked to precipitation levels
2. Thunderstorms linked to asthma attacks
3. Sleep disorders linked to cognitive deficits in children.
4. Osteoarthritis Risk Linked To Finger Length Ratio
5. Fast food restaurants linked to prevalence of strokes
6. Attendance at violent movies linked to reduction in crime.

How are the length of time a college student has been in school and the number of credits he/she has accumulated related?

Suppose Registration and Records at NC State tracked 100 students for 6 consecutive semesters from the Fall 2002 semester through the Spring 2005 semester. The students are all incoming freshmen (no transfer students) for the Fall 2002 semester and all 100 stay enrolled throughout the six semesters (no dropouts).

Shown below are boxplots of the cumulative credits for each semester for these 100 students.



Questions

1. Are the shapes of the distributions consistent over time?

2. Is the center consistent over time? Do we expect that given the data? Why or why not?

3. Is the variability consistent over time? Do we expect that given the data? Why or why not?

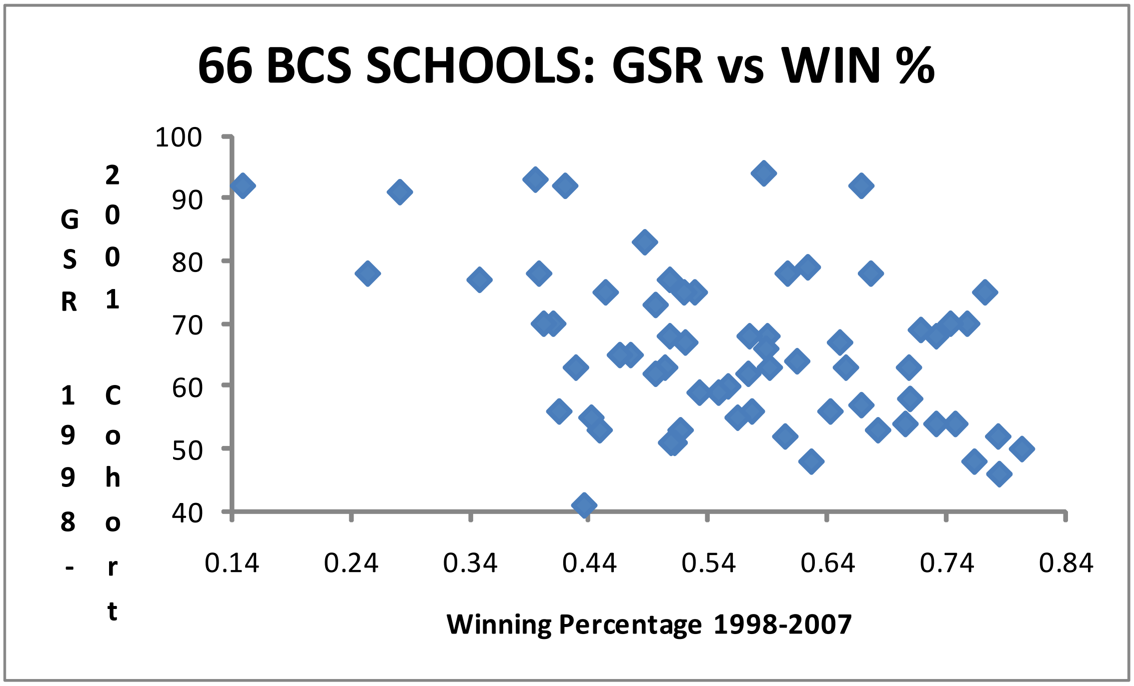
4. By examining the box plots, write down how you would explain to another person how the length of time a college student has been in school and the number of credits he/she has accumulated are related.

Scatterplots

graphical display of bivariate data when both variables are quantitative

bivariate data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ both variables quantitative

EXAMPLE: **x -variable**: winning percentage of each of the 66 football teams in the Bowl Championship Series (BCS) over the 10-year period 1998-2007.
y-variable: mean of the **Graduation Success Rate** (GSR) for football players at each of the BCS universities that entered school in the 4 years 1998-2001 (the GSR is basically the proportion of players that graduated within 6 years).



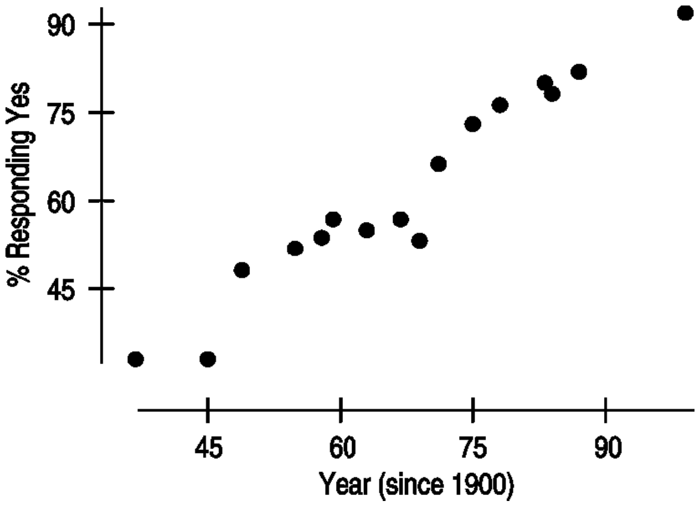
1. Describe the overall **shape** of the relationship
   - Linear/curved?
   - Clusters
   - Outliers

2. Describe the **trend or direction** of the relationship

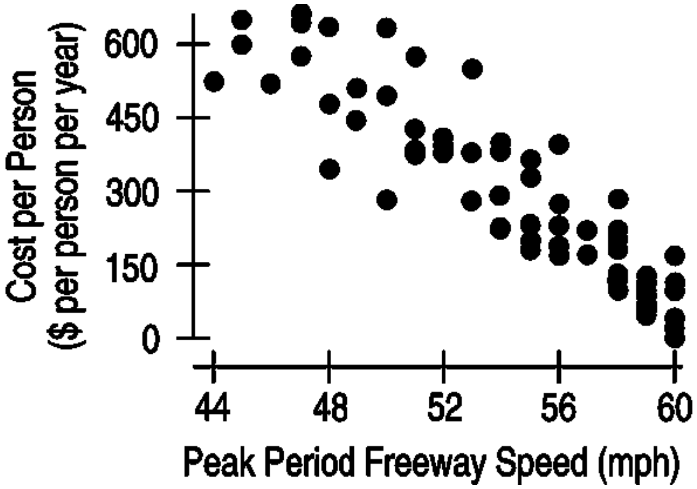
3. Describe the **strength** of the relationship
   Strong/Weak
   Constant/Varying

4. Are there plausible explanations for the pattern? Lurking variables?

Percent of US voters who said they would vote for a woman for US president



Cost per person of traffic delays in 70 US cities

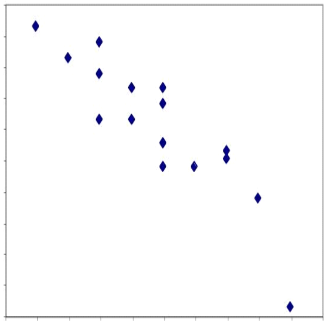


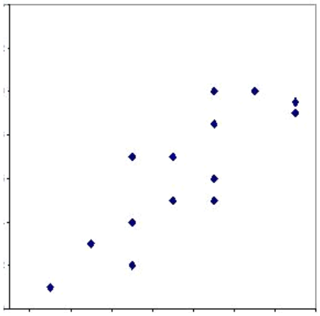
To make a scatterplot: using ti83/84, see p. 170; using Excel, p. 169.

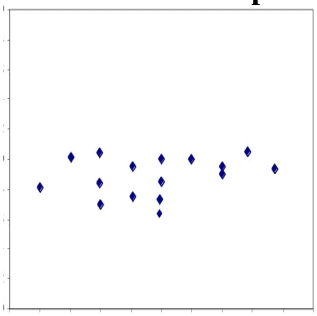
Characteristics on which to focus when examining a scatterplot

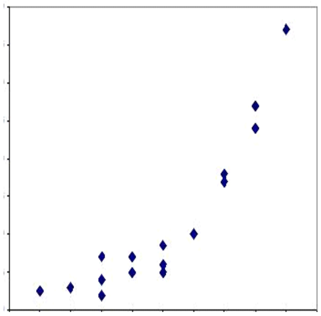
1) Form
2) Direction
3) Strength
4) Outliers

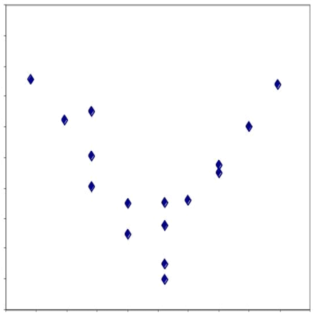
Form and Direction of an association

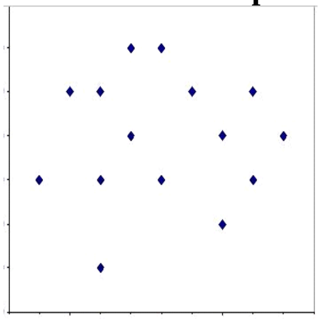

Linear


Linear

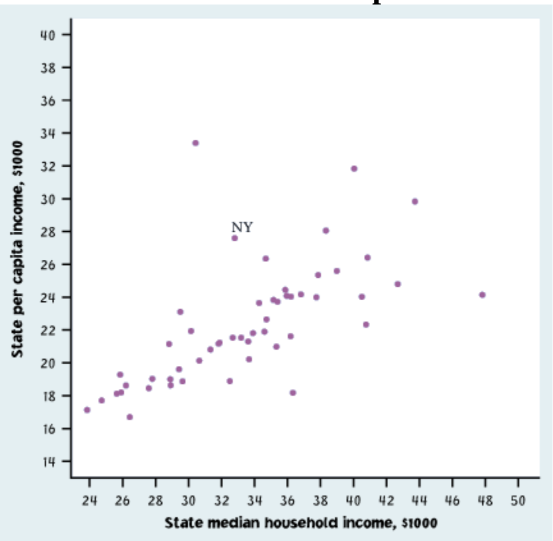

No relationship


Nonlinear

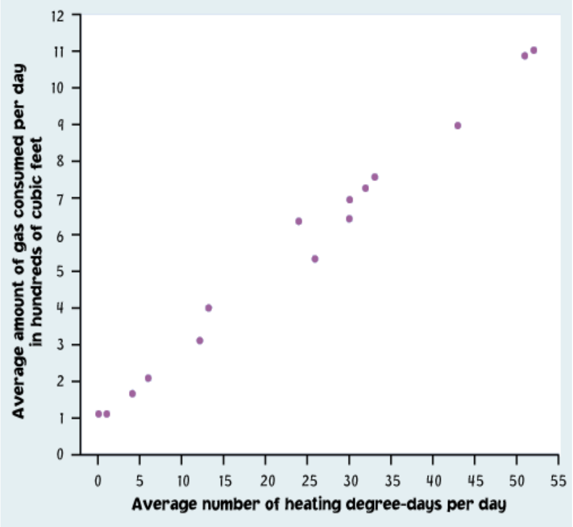

Nonlinear


No relationship

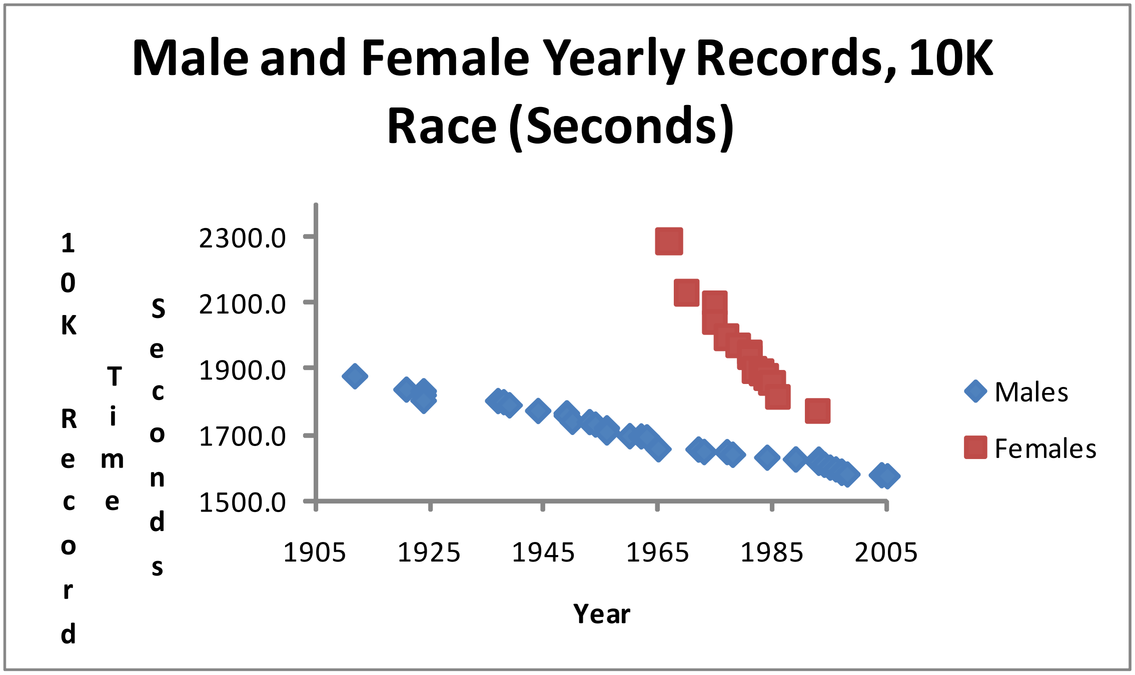
Strength


Weak relationship

For a particular state median household income, you can't predict the state per capita income very well.


Strong relationship

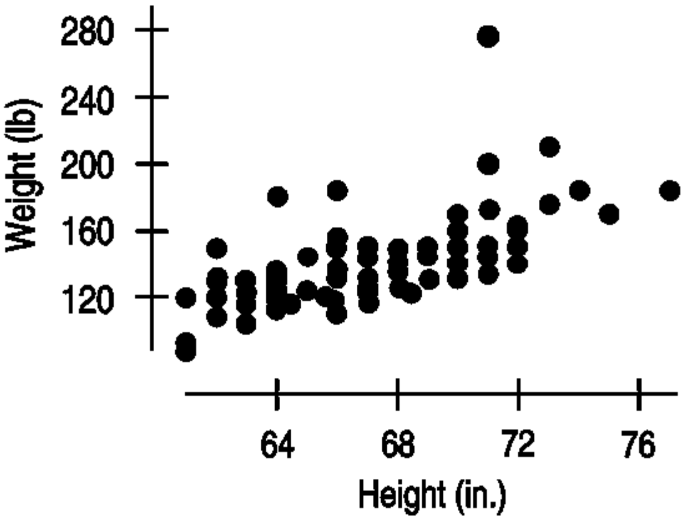
The daily amount of gas consumed can be predicted quite accurately for a given temperature value.



Correlation:

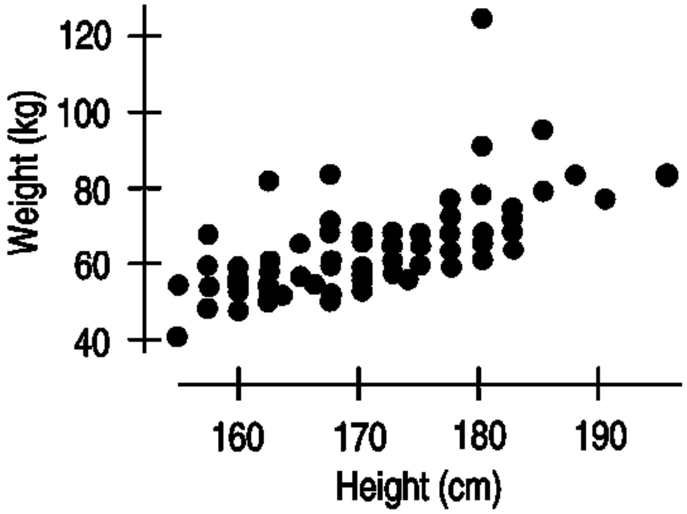
A Quantitative Measure of the Linear Relationship Between Two Quantitative Variables

Student weight (in pounds) vs height (in inches)



the choice of the units for the variables should not matter

Student weight (in kilograms) vs height (in centimeters)
the shape of the pattern is not changed

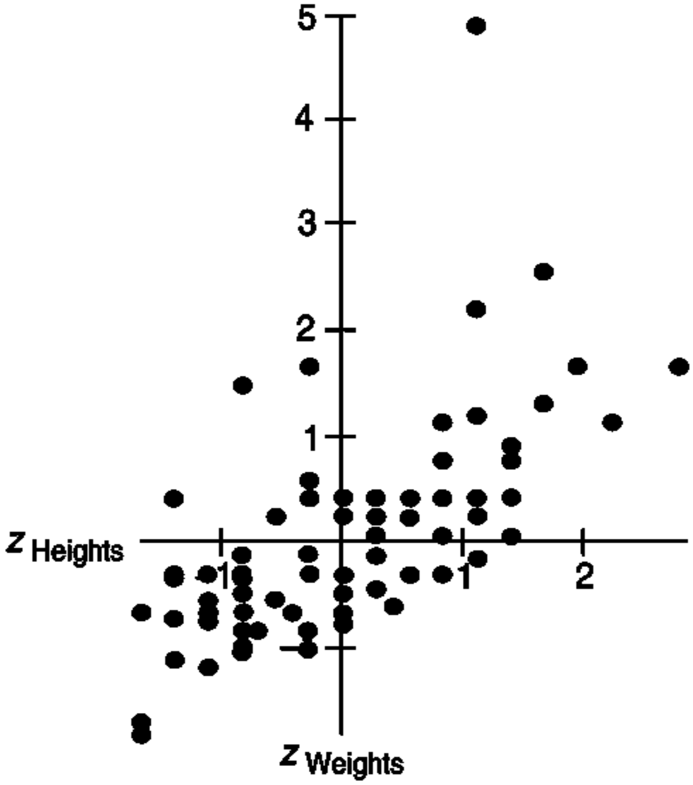


Remove the units! HOW? z-scores.

standardize both variables

observation $(x_i,\, y_i)$ becomes

$$(z_{x_i},\, z_{y_i}) \;=\; \left(\frac{x_i - \overline{x}}{s_x},\, \frac{y_i - \overline{y}}{s_y}\right)$$



In this standardized plot:
⇒ the center of this scatterplot is at the origin
⇒ the scales on both axes are now standard deviation units
⇒ the linear pattern seems steeper since the length of one standard deviation is the same vertically and horizontally

WARNING: scatterplot axes can be manipulated to give a distorted visual impression of the strength of the linear relationship between x and y.

Correlation coefficient

$$\boldsymbol{r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}}$$

Correlation coefficient in terms of original x and y

$$\boldsymbol{r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}$$

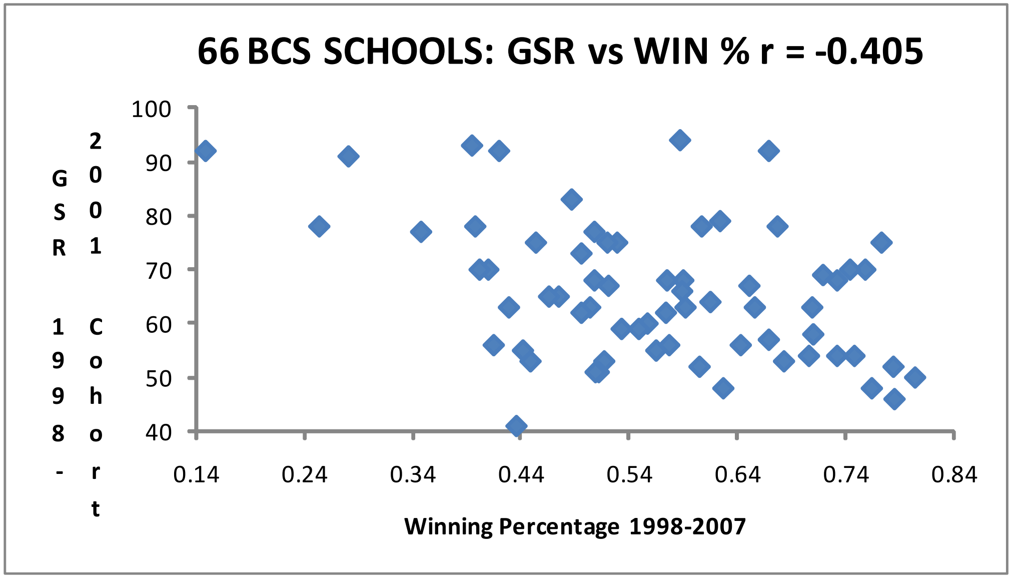
The correlation coefficient measures the strength of the *linear* association between two quantitative variables

DO NOT calculate correlation by hand!
To calculate correlation: using the ti83/84: see p. 170 in text; using Excel, DataDesk and other technology: see p. 169.

Properties of the correlation:

1) scaleless (no units)
2) $-1 \ \leq \ r \leq \ 1$
   $r = \ -1$ only if $y = a + bx$ with slope $b < 0$
   $r = \ +1$ only if $y = a + bx$ with slope $b > 0$
3) It doesn't make any difference which variable you call "x" and which variable you call "y"
4) r is not affected by a linear transformation on x or y if the linear transformation has slope $b > 0$.

NOTE! correlation does not measure causation



Examples:

i) second-hand smoke

ii) coffee & heart disease

iii) <u>polio in 1950's:</u> high positive correlation between incidence of polio and soft drink consumption; both of these variables are affected by a 3rd variable: weather (**covariate**)

iv) many newspaper articles report evidence of correlations, **but not necessarily causation.**

STUDY FINDS WEIGHT GUIDELINES TOO GENEROUS

Middle-age bulge called risky

Associated Press

CHICAGO Middle-aged women should weigh far less than people think — and less than the government recommends — in order to have healthy hearts, Harvard researchers say.

"We found that about 40 percent of all heart attacks that occur in middle-aged women are due to overweight," said Dr. JoAnn E. Manson, co-director of women's health at Harvard-affiliated Brigham and Women's Hospital in Boston. She said similar results are found in men.

The study showed that women of average weight had about a 38 percent higher risk of heart attack than women who were 15 percent less than average U. S. weights.

Women who gained 10 or fewer pounds in early to middle adulthood had the lowest risk of heart attacks, the researchers report in today's issue of The Journal of the American Medical Association.

Weight risk guidelines

An easy rule of thumb for estimating ideal body weight, consistent with findings of the Harvard study:

■For women, 100 pounds for a height of 5 feet, with five additional pounds for each added inch of height.

■ For men, 106 pounds for a height of 5 feet, and six addit-ional pounds for every added inch of height.

These ideals may vary by plus or minus 10 percent.

For example, a 5-foot-4-inch woman had the lowest risk if she weighed less than 120 pounds. At the same height, a weight of 120 to 142 pounds carried a 20 percent higher risk. At 142 to 156 pounds, it was 50 percent higher; at 156 to 180 pounds it was double; and at more than 180 pounds, it was $3\frac{1}{2}$ times higher than for the 120 pound woman.

"I don't want to be scaring people with these findings, but we have been overly complacent about obesity and weight gain in adults," Manson said by telephone Monday.

The federal government in 1990 revised its guidelines for desirable weights upward, saying Americans over age 35 could be significantly heavier than under 1965 guidelines.

"The current federal weight guidelines are in a sense encouraging the fattening of Americans," Manson said, noting that one in three adults is overweight.

While cautioning against over-reaction to the findings, she recommended increasing physical activity, lowering the fat and calorie content of the diet and eating more fruits, vegetables, and grains.

Important: before you use the correlation to describe the realtionship between 2 variables, check the following conditions:

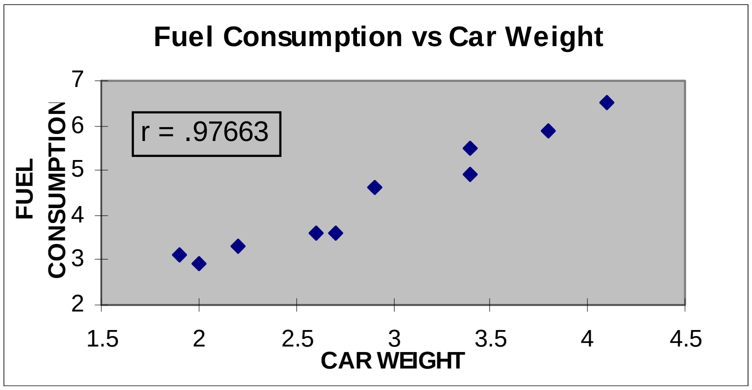
1) <u>Quantitative variables condition</u>.
correlation applies only to quantitative variables; don't apply correlation to categorical variables masquerading as quantitative variables

2) <u>Straight enough condition</u>
correlation measures the strength only of the linear association, and will be misleading if the relationship is not linear. What is "straight enough"? That's a judgment call.

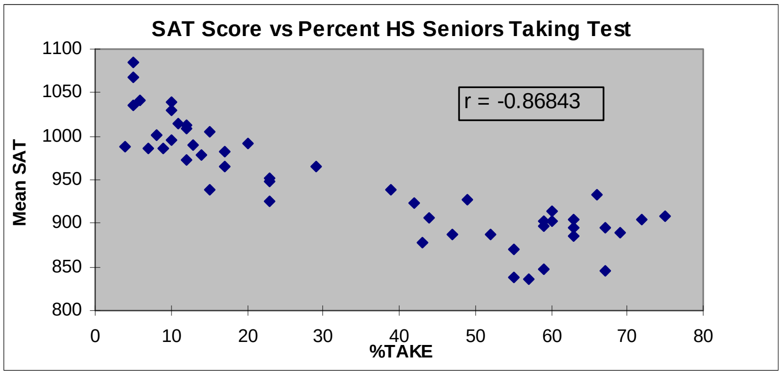
3) <u>Outlier condition.</u>
Outliers can dramatically distort the correlation. If there is an outlier, report the correlation both with and without the outlier.

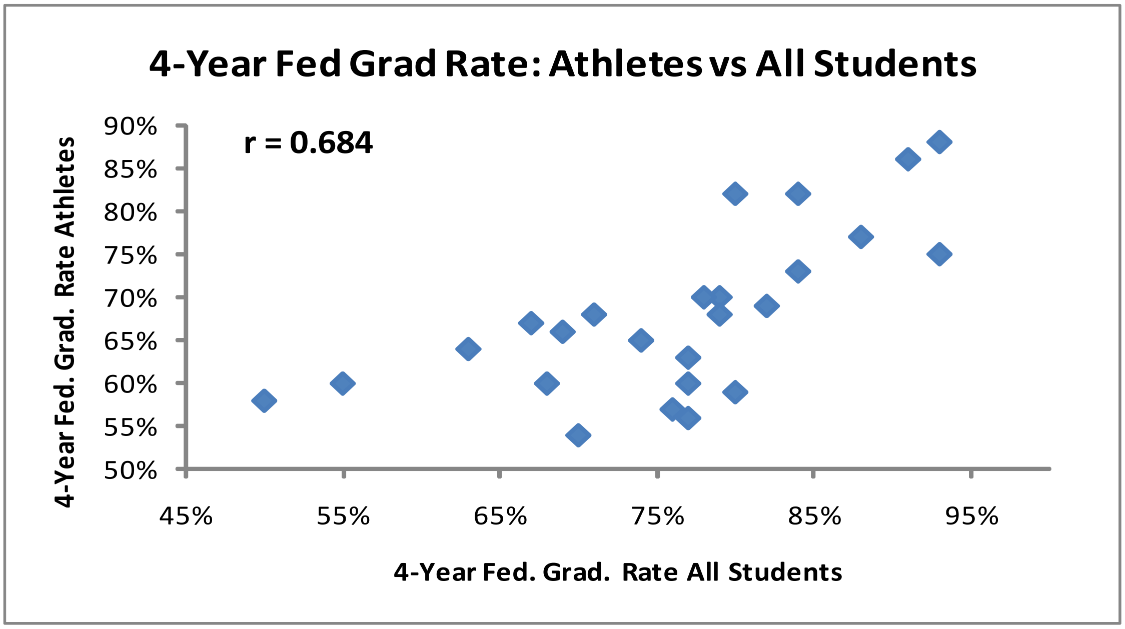
Examples:

1) Car weight and fuel consumption
x = car weight (in thousands of pounds)
y = fuel consumption (gallons needed to go 100 miles)



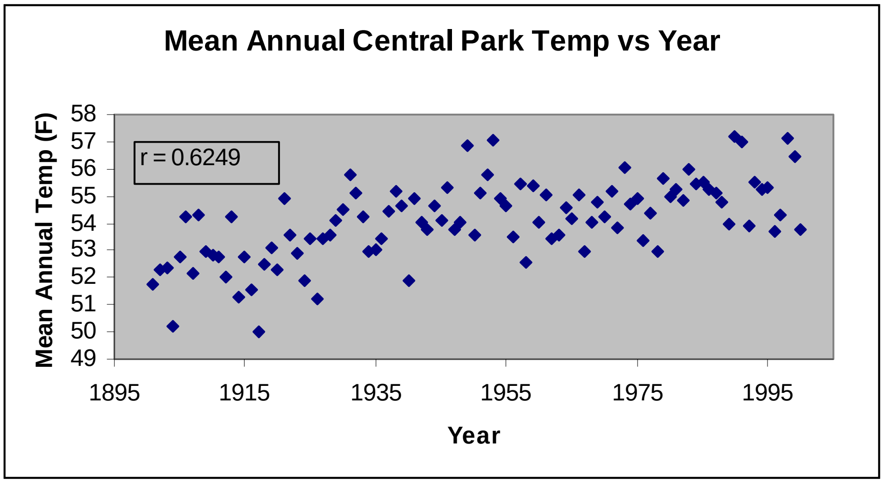
2) SAT scores and percentage of hs seniors taking the test
x = percentage of high school seniors in a state taking the test
y = mean SAT score in a state



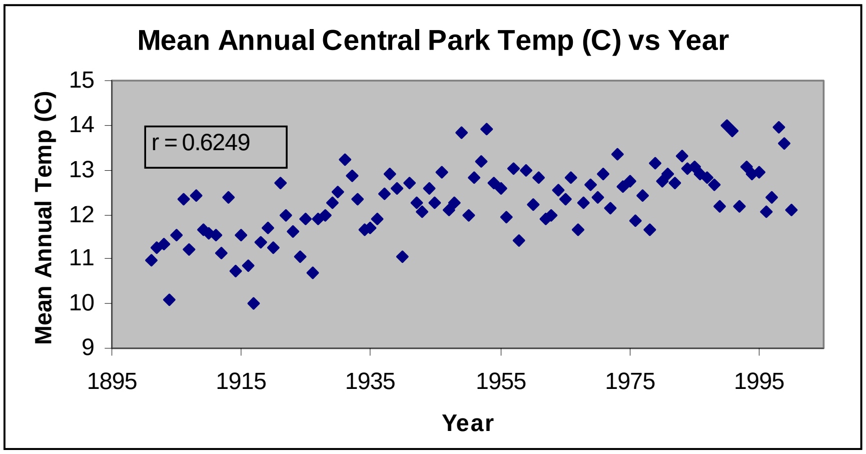


Example (linear transformation does not affect correlation)

temperature is in F°



same scatterplot but temperature is in C°

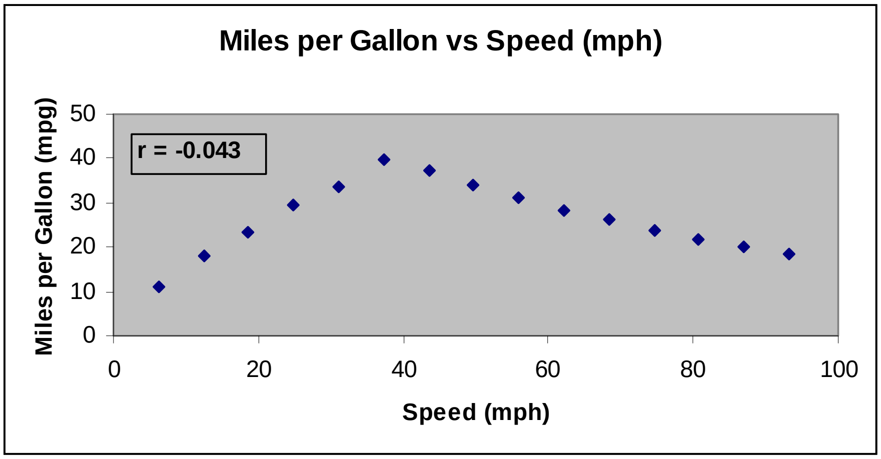


What can Go Wrong

1. **Don't say "correlation" when you mean "association"**
   Association is a deliberately vague term.
   Correlation is a precise term describing the strength and direction of the **linear** relationship between 2 quantitative variables.

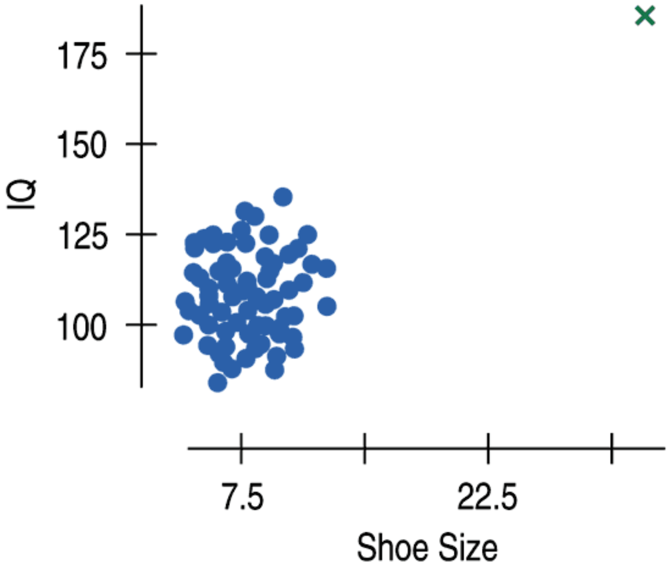
2. **Don't correlate categorical variables**
   The correlation between gender of worker and salary is 0.75.

3. **Be sure the association is linear**
   Plot the data!
   In the graph below the small correlation could give the impression that there is no relationship between speed and miles per gallon.
   The relationship is **nonlinear**.



4. **Beware of outliers.**

Relationship between IQ and shoe size among comedians: surprisingly strong positive correlation of 0.50



The outlier is Bozo the Clown, known for his large shoes and widely acknowledged to be a comic "genius". Without Bozo, the correlation is near 0.

5. D**on't confuse correlation with causation.**