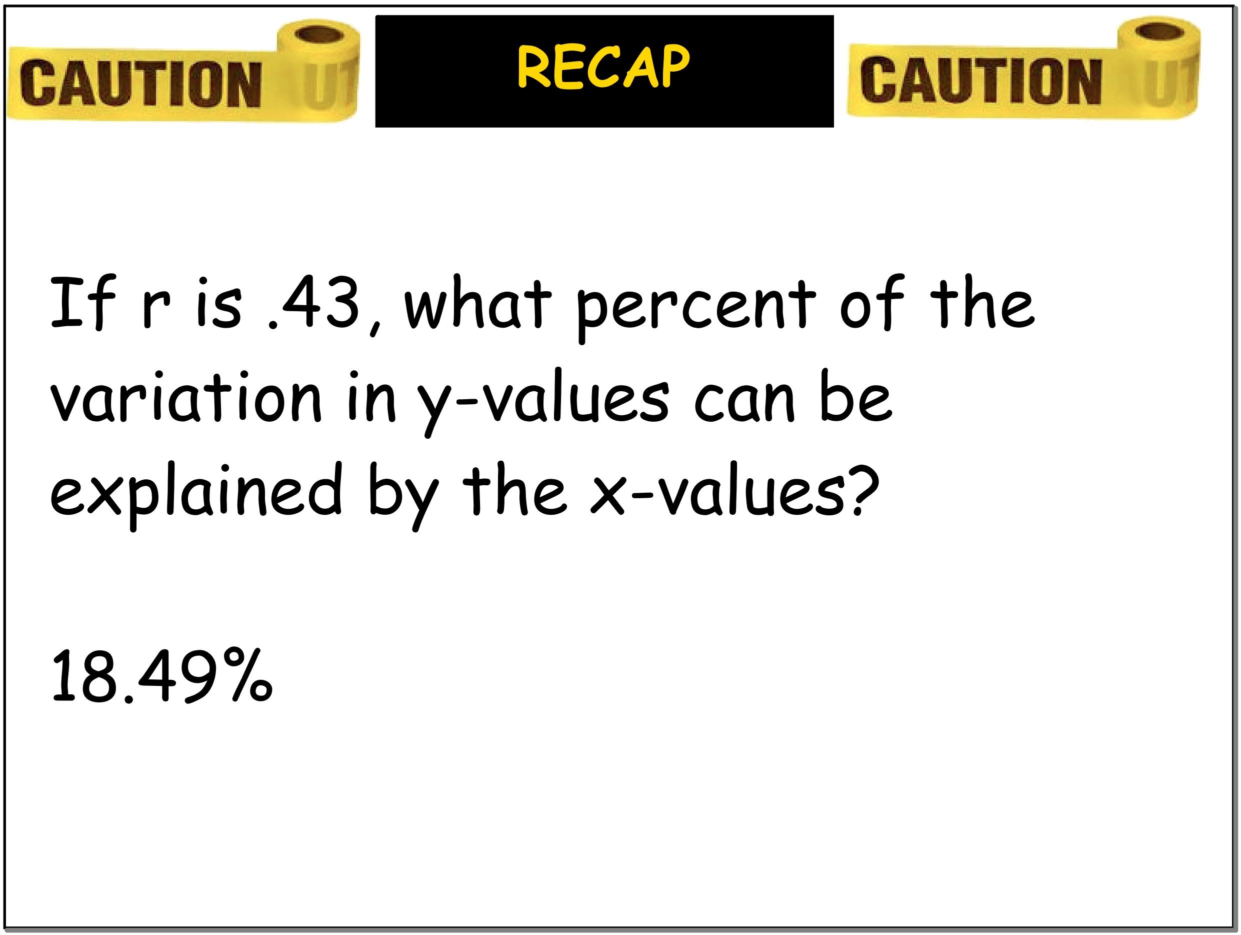
3.4 What are some cautions in analyzing association?

Sep 28-5:12 PM

Objectives

- Extrapolation
- Outliers and Influential Observations
- Correlation does not imply causation
- Lurking variables and confounding
- Simpson's Paradox

Sep 28-5:12 PM

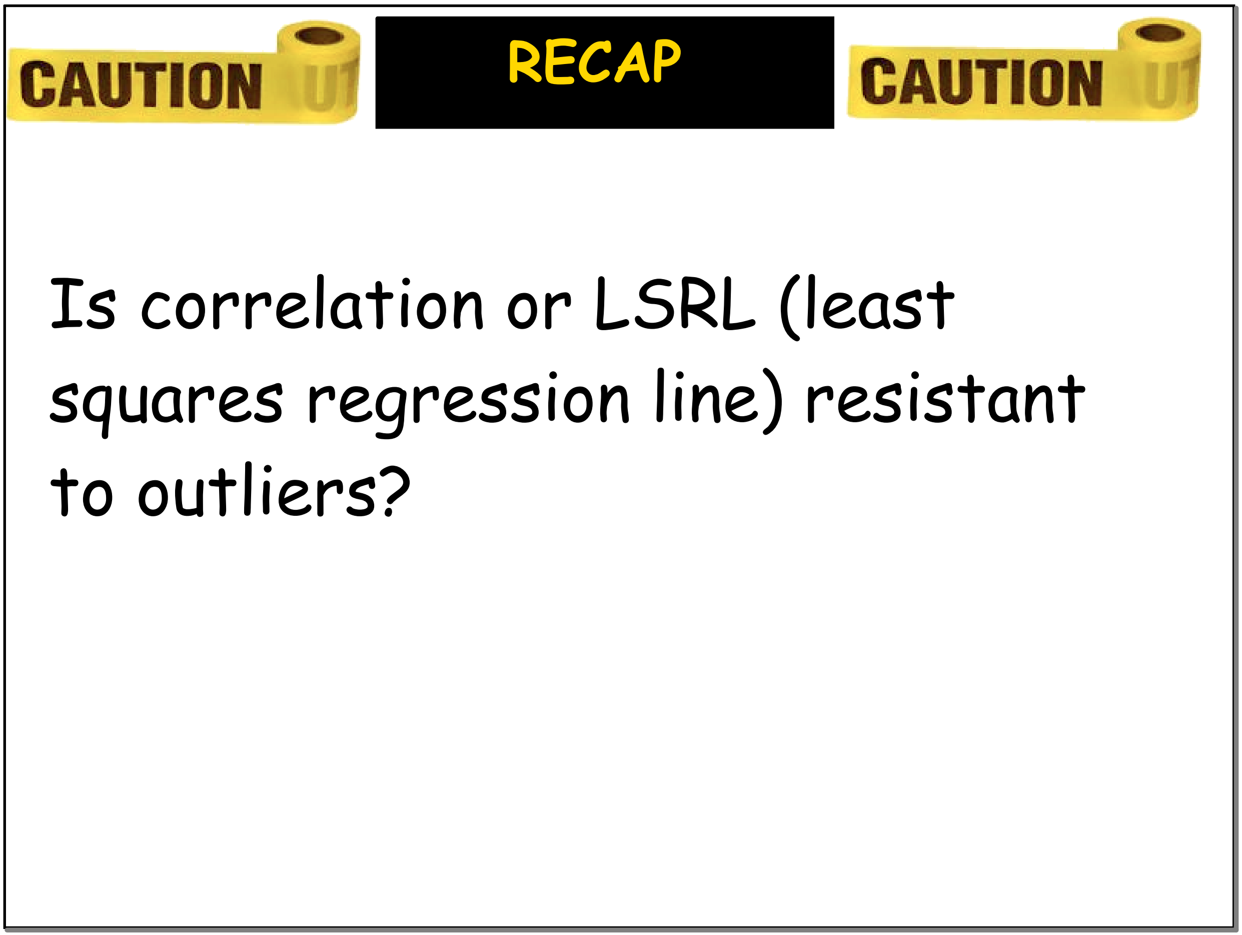

CAUTION
CAUTION

RECAP

If r is .43, what percent of the variation in y-values can be explained by the x-values?

18.49%

Sep 28-5:12 PM

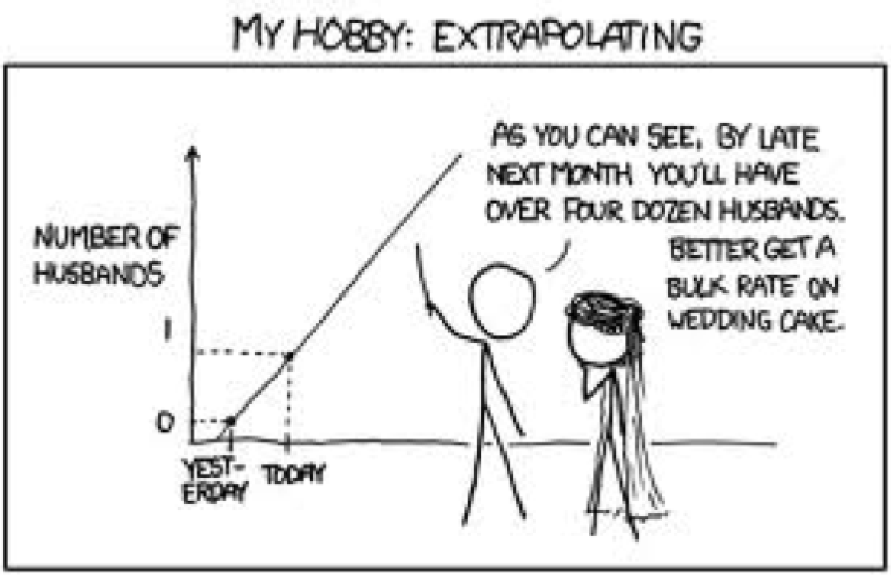

CAUTION
CAUTION

RECAP

Is correlation or LSRL (least squares regression line) resistant to outliers?

Sep 28-5:12 PM

Extrapolation


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH, YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

Sep 28-5:12 PM

Extrapolation

Extrapolation: Using a regression line to predict y-values for x-values outside the observed range of the data

Riskier the farther we move from the range of the given x-values

There is no guarantee that the relationship given by the regression equation holds outside the range of sampled x-values

Sep 28-5:12 PM

Extrapolation

For example, take the price of a car and its age. The older a car is the less it costs. This is only true up to a certain point though – then the car can become a classic and increase in value.

Sep 28-5:12 PM

Extrapolation

Think about the size of your shoe...if you had a least squares regression line (LSRL) for the size of your shoe from the time you were born to age 12, would you use the same line to predict your shoe size when you are 45?

Sep 28-5:12 PM

Outliers & Influential Points

When you construct a scatterplot, search for data points that are well outside of the trend that the remainder of the data points follow

Sep 28-5:12 PM

Regression Outliers

A **regression outlier** is an observation that lies far away from the trend that the rest of the data follows

An observation is influential if:

Its x value is relatively low or high compared to the remainder of the data

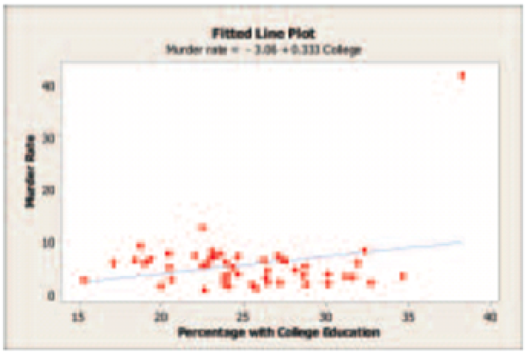
The observation is a regression outlier

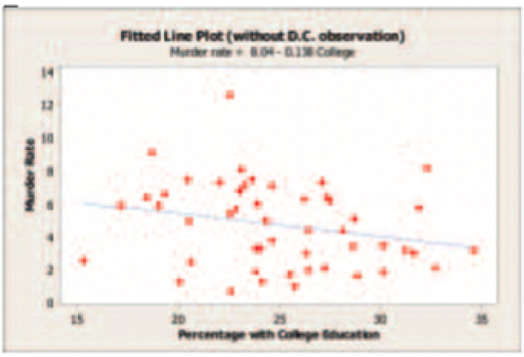
Influential observations tend to pull the regression line toward that data point and away from the rest of the data

Sep 28-5:12 PM

Regression Outliers

Impact of removing an Influential data point





Sep 28-5:12 PM

Correlation does not imply causation

WARNING

Correlation does not imply causation!!!!

Sep 28-5:12 PM

Correlation does not imply causation

A strong correlation between x and y means that there is a strong linear association that exists between the two variables

A strong correlation between x and y, does not mean that x causes y

Sep 28-5:12 PM

Correlation does not imply causation

Data are available for all fires in Chicago last year on x = number of firefighters at the fires and y = cost of damages due to fire

Would you expect the correlation to be negative, zero, or positive?

If the correlation is positive, does this mean that having more firefighters at a fire causes the damages to be worse? Yes or No

Sep 28-5:12 PM

Association does not imply causation

Data are available for all fires in Chicago last year on x = number of firefighters at the fires and y = cost of damages due to fire

Identify a third variable that could be considered a common cause of x and y:

a. Distance from the fire station
b. Intensity of the fire
c. Size of the fire

Sep 28-5:12 PM

Lurking Variable

A **lurking variable** is a variable, usually unobserved, that influences the association between the variables of primary interest

Ice cream sales and drowning

temperature →

Reading level and shoe size

age →

Sep 28-5:12 PM

Lurking Variable

Most car accidents happen close to home!

Taller people are better at math!

Sun block sales are associated with higher murder rates!

Risk of heart attack is associated with race!

You get better gas mileage with a heavier car!

Students with bigger heads have higher reading abilities!

Sep 28-5:12 PM

Confounding

A coach wants his players to do better, so he has them run 2 miles at every practice. Without knowing it, the players also start taking vitamins.

Two months later, they are playing better.

But is that from the running or the vitamins?
This is confounding.

When two explanatory variables are both associated with a response variable but are also associated with each other, there is said to be **confounding**

Sep 28-5:12 PM

Confounding

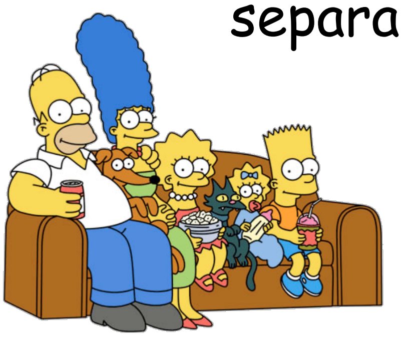
A group of people is offered either a low deductible and a high interest rate on their insurance or a high deductible and a low interest rate.

We'll never know if they picked their plan based on deductible, interest rate, or a combination of both... confounding.

Sep 28-5:12 PM

Simpson's Paradox

When the direction of an association between two variables changes after we include a third variable and analyze the data at separate levels of that variable



Sep 28-5:12 PM

Simpson's Paradox

Example:

Is Smoking Actually Beneficial to Your Health?

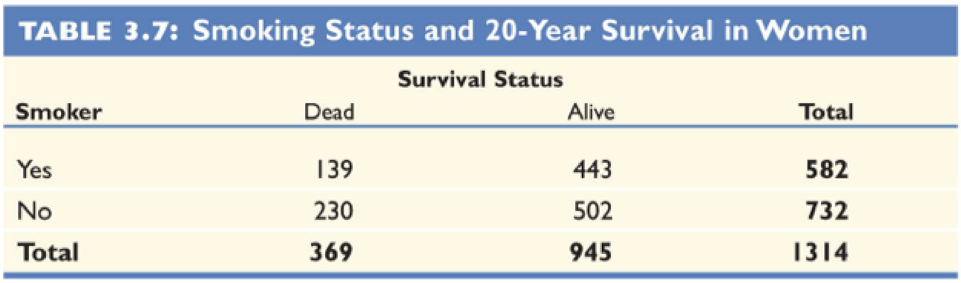


TABLE 3.7: Smoking Status and 20-Year Survival in Women

	Survival Status		
Smoker	Dead	Alive	**Total**
Yes	139	443	**582**
No	230	502	**732**
Total	**369**	**945**	**1314**

Probability of Death of Smoker = 139/582=24%

Probability of Death of Nonsmoker=230/732=31%

This can't be true that smoking improves your chances of living! What's going on!

Simpson's Paradox

Break out Data by Age

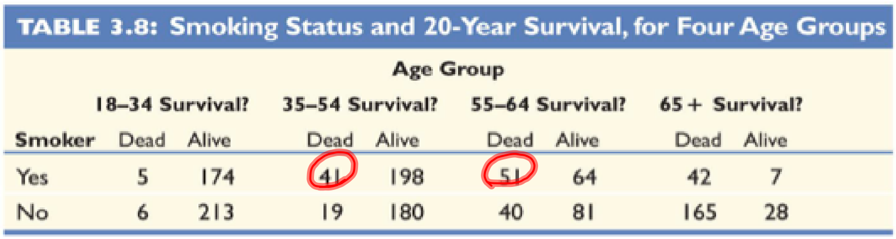


TABLE 3.8: Smoking Status and 20-Year Survival, for Four Age Groups

	Age Group							
	18–34 Survival?		35–54 Survival?		55–64 Survival?		65 + Survival?	
Smoker	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28

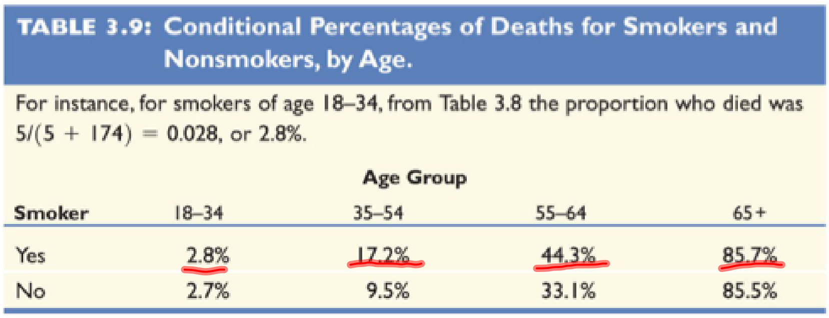
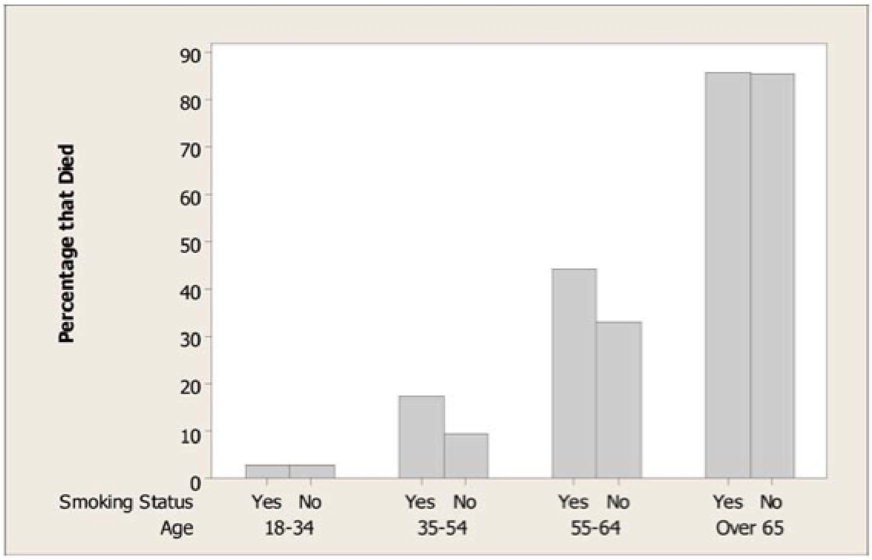


TABLE 3.9: Conditional Percentages of Deaths for Smokers and Nonsmokers, by Age.

For instance, for smokers of age 18–34, from Table 3.8 the proportion who died was 5/(5 + 174) = 0.028, or 2.8%.

	Age Group			
Smoker	18–34	35–54	55–64	65+
Yes	2.8%	17.2%	44.3%	85.7%
No	2.7%	9.5%	33.1%	85.5%

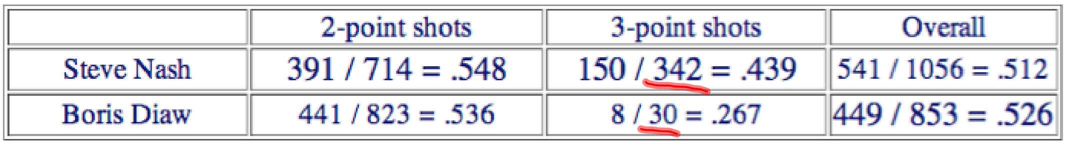


▲ **FIGURE 3.23: MINITAB bar graph comparing percentage of deaths for smokers and nonsmokers, by age.** This side-by-side bar graph shows the conditional percentages from Table 3.9.

An association can look quite different after adjusting for the effect of a third variable by grouping the data according to the values of the third variable

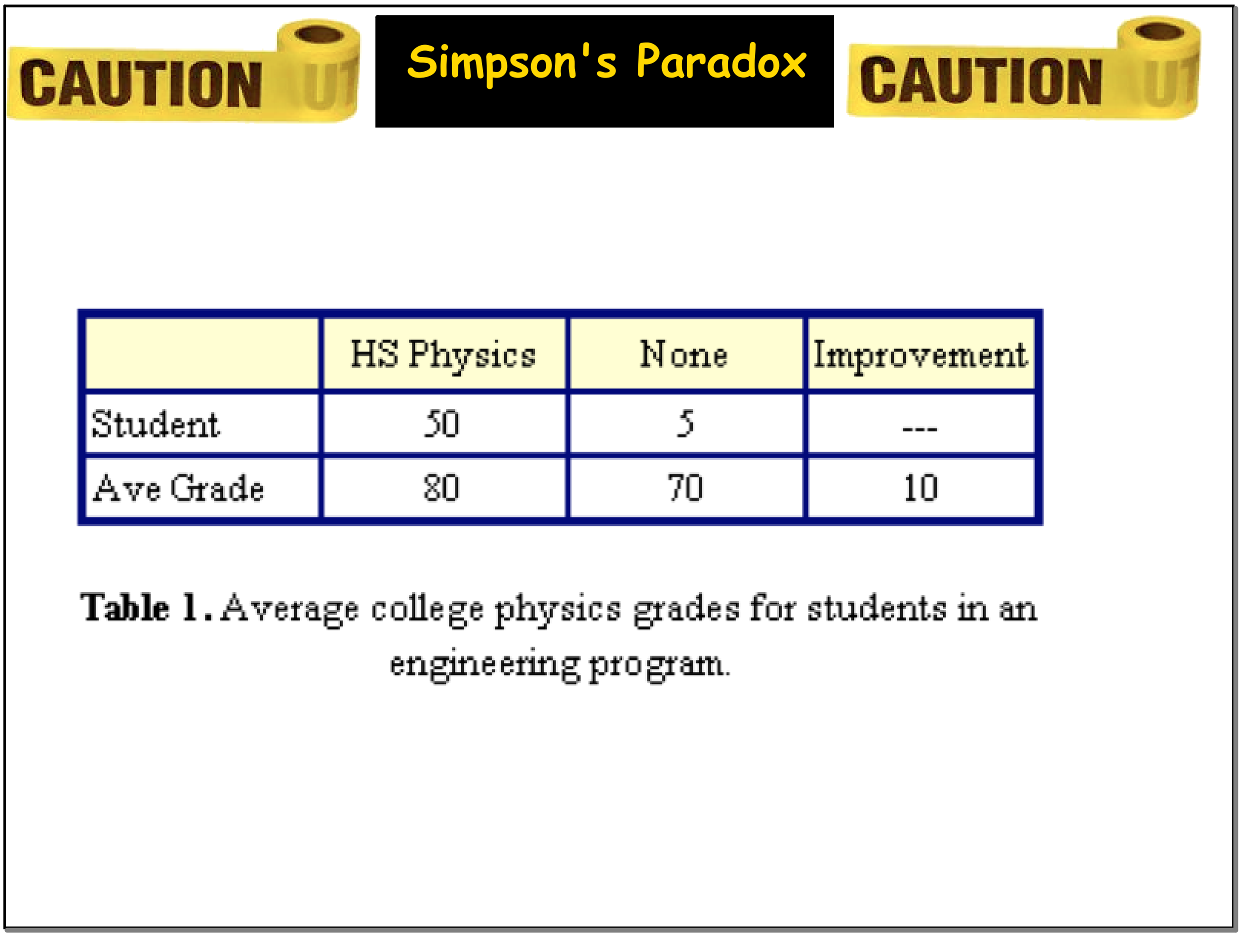
Sep 28-5:12 PM

Simpson's Paradox



	2-point shots	3-point shots	Overall
Steve Nash	391 / 714 = .548	150 / 342 = .439	541 / 1056 = .512
Boris Diaw	441 / 823 = .536	8 / 30 = .267	449 / 853 = .526

Sep 28-5:12 PM

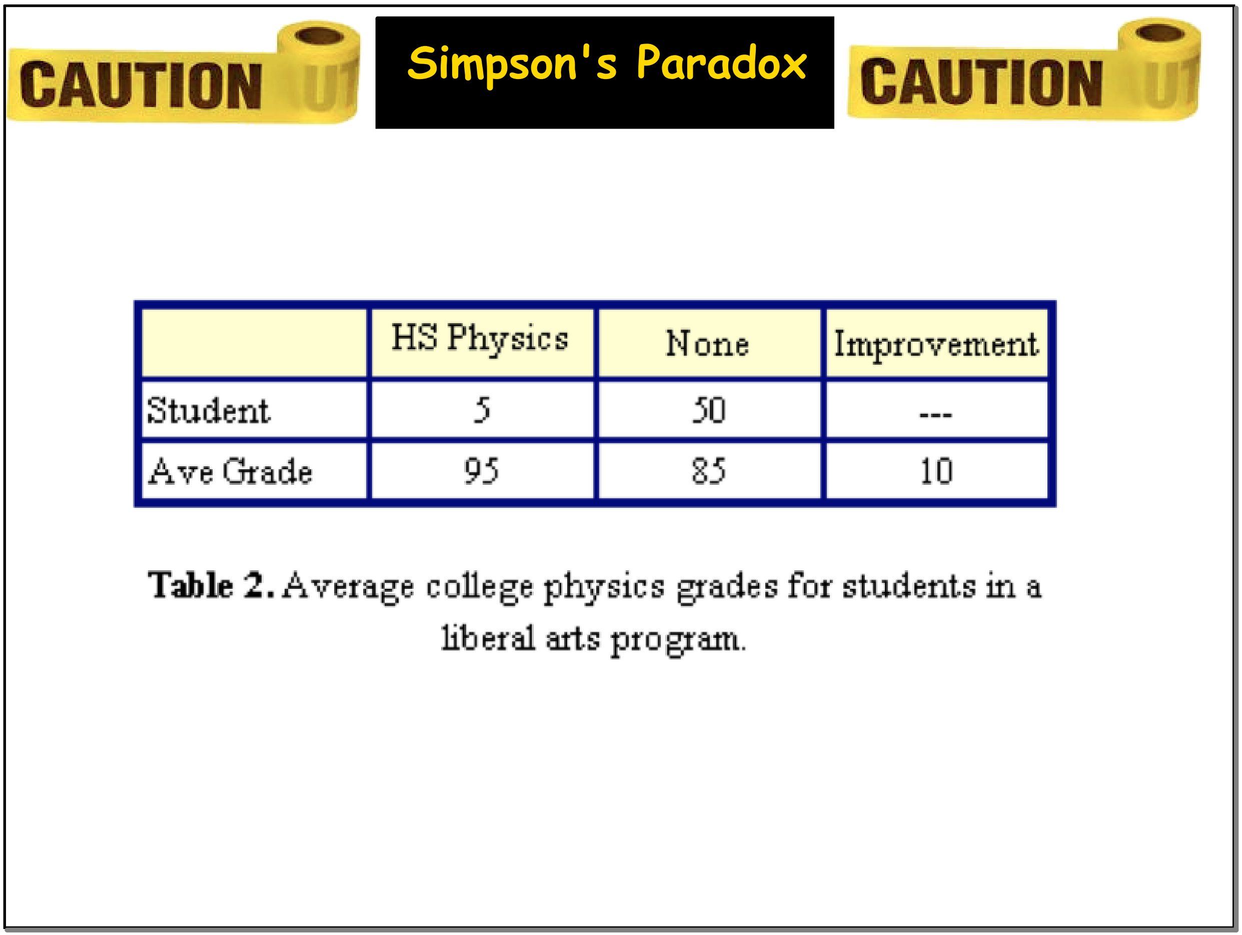

CAUTION
CAUTION

Simpson's Paradox

	HS Physics	None	Improvement
Student	50	5	---
Ave Grade	80	70	10

Table 1. Average college physics grades for students in an engineering program.

Sep 28-5:12 PM

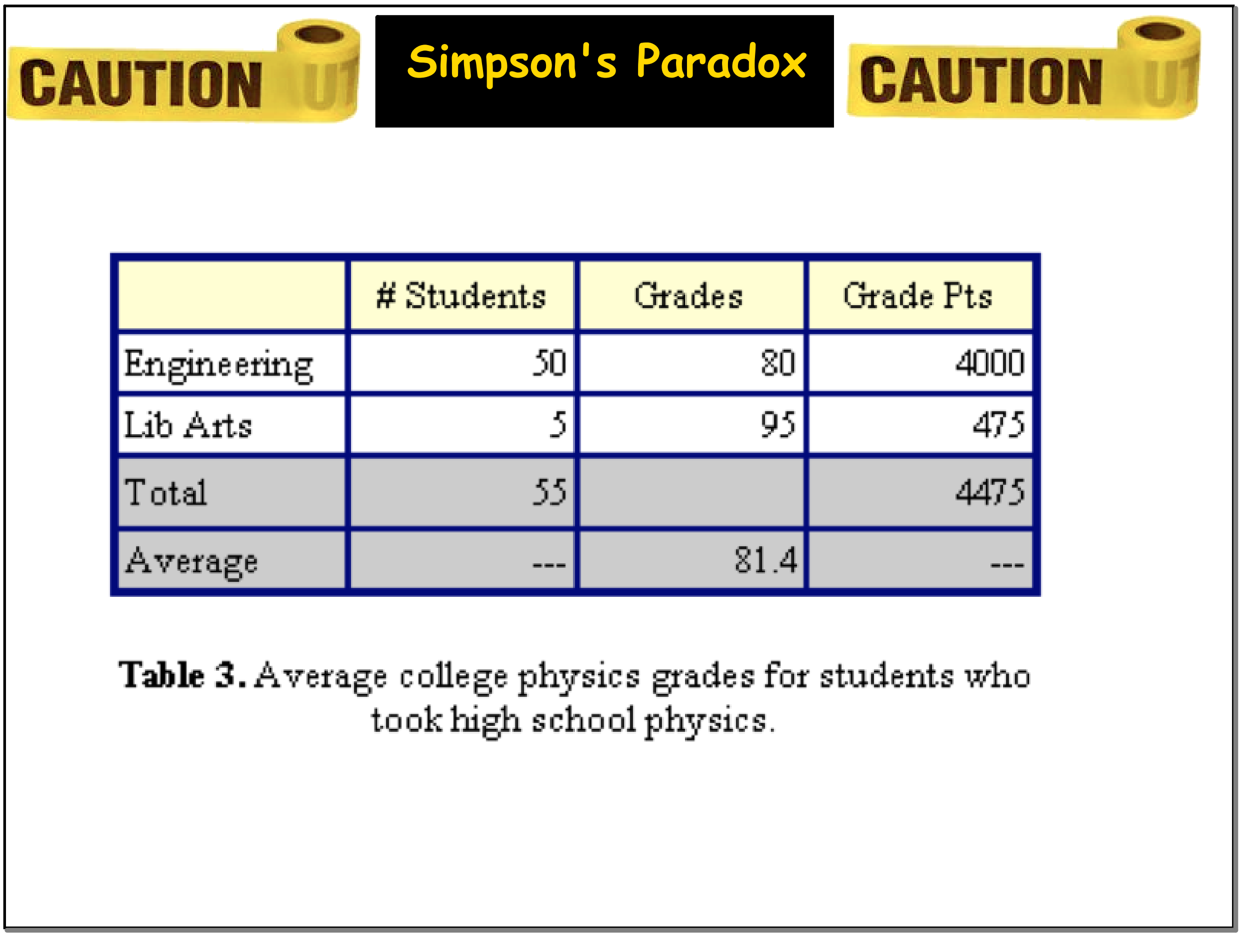

CAUTION
CAUTION

Simpson's Paradox

	HS Physics	None	Improvement
Student	5	50	---
Ave Grade	95	85	10

Table 2. Average college physics grades for students in a liberal arts program.

Sep 28-5:12 PM



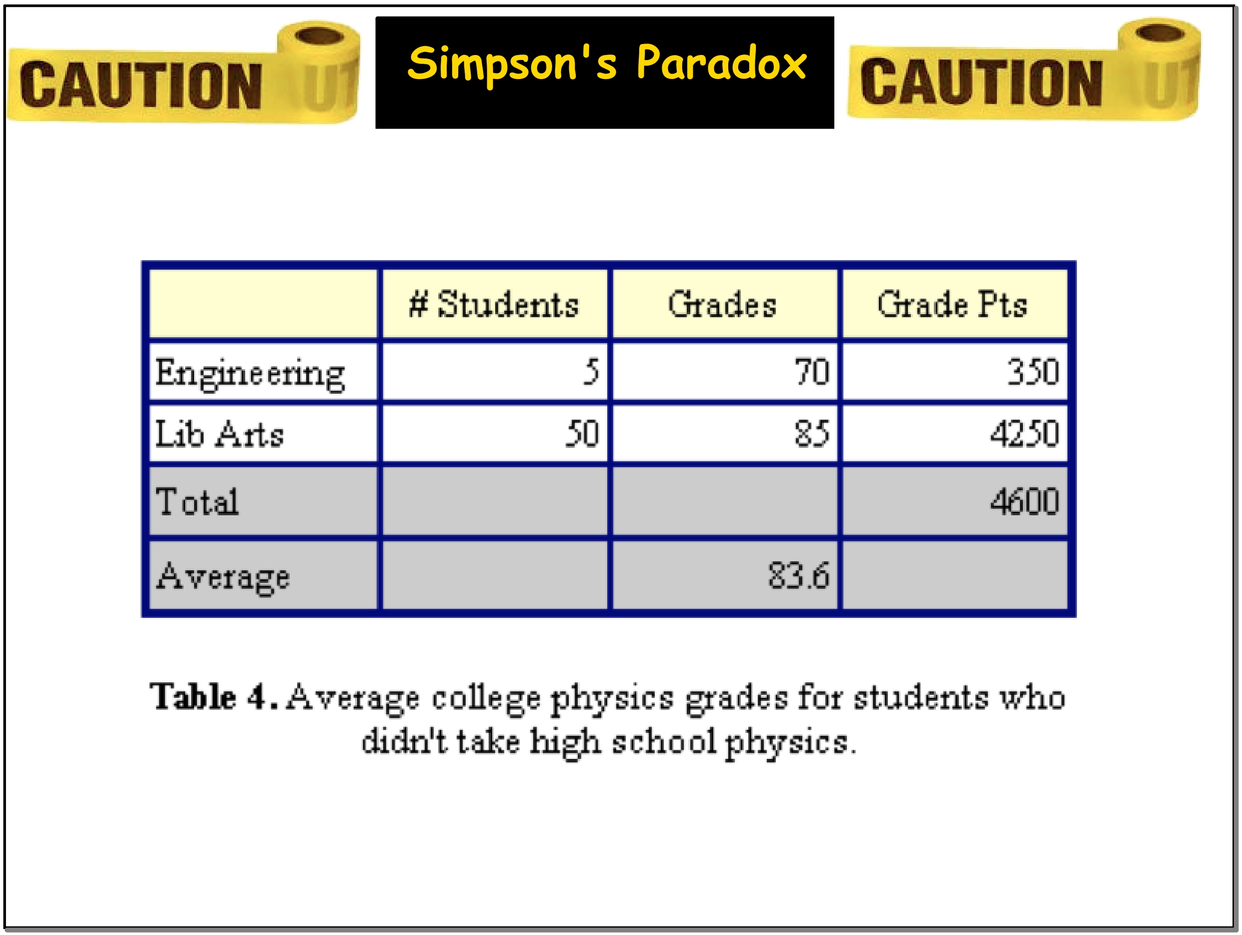
CAUTION
CAUTION

Simpson's Paradox

	# Students	Grades	Grade Pts
Engineering	50	80	4000
Lib Arts	5	95	475
Total	55		4475
Average	---	81.4	---

Table 3. Average college physics grades for students who took high school physics.

Sep 28-5:12 PM



CAUTION
CAUTION

Simpson's Paradox

	# Students	Grades	Grade Pts
Engineering	5	70	350
Lib Arts	50	85	4250
Total			4600
Average		83.6	

Table 4. Average college physics grades for students who didn't take high school physics.

Sep 28-5:12 PM

Simpson's Paradox

Simpson's Paradox is caused by a combination of a lurking variable and data from unequal sized groups being combined into a single data set. The unequal group sizes, in the presence of a lurking variable, can weight the results incorrectly. This can lead to seriously flawed conclusions.

The obvious way to prevent it is to not combine data sets of different sizes from diverse sources or sizes.

Sep 28-5:12 PM

Simpson's Paradox

Simpson's Paradox will generally not be a problem in a well designed experiment or survey if possible lurking variables are identified ahead of time and properly controlled. This includes eliminating them, holding them constant for all groups or making them part of the study.

Sep 28-5:12 PM

HOMEWORK

Page 141

#44, 45, 47, 53 - 56, and 57 c, d

- Collect that project data!!

Sep 28-5:12 PM