

Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Theory-based approach for paired data, and M&M example.
2. Multiple testing and publication bias.
3. Two variables and correlation.
4. Linear regression.
5. Calculating correlation.
6. Slope of the regression line.
7. Goodness of fit.

Read ch7 and 10.

Hw4 is due Sep11, 10pm, again by email to statgrader or statgrader2, and is Problems 10.1.8, 10.3.14, 10.3.21, and 10.4.11.

<http://www.stat.ucla.edu/~frederic/13/sum24> .

1. Theory based approach for Analyzing Data from Paired Samples, and M&Ms.

Section 7.3

How Many M&Ms Would You Like?

Example 7.3

How Many M&Ms Would You Like?

- Does your bowl size affect how much you eat?
- Brian Wansink studied this question with college students over several days.
- At one session, the 17 participants were assigned to receive either a small bowl or a large bowl and were allowed to take as many M&Ms as they would like.
- At the following session, the bowl sizes were switched for each participant.

How Many M&Ms Would You Like?

- What are the observational units?
- What is the explanatory variable?
- What is the response variable?
- Is this an experiment or an observational study?
- Will the resulting data be paired?

How Many M&Ms Would You Like?

The hypotheses:

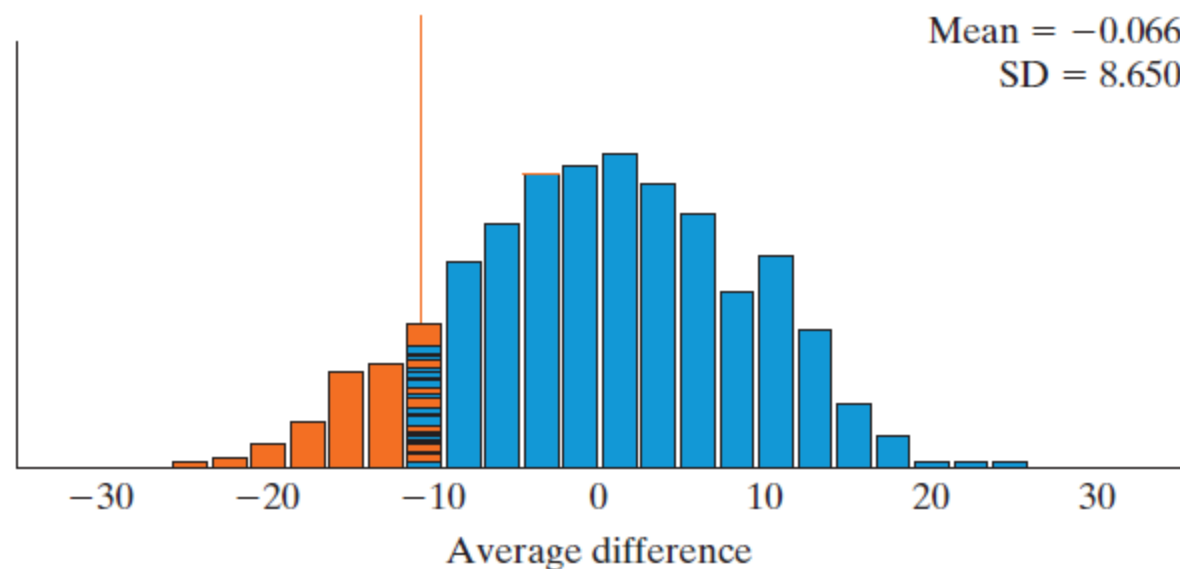
- $H_0: \mu_d = 0$
 - The long-run mean difference in number of M&Ms taken (small – large) is 0.
- $H_a: \mu_d < 0$
 - The long-run mean difference in number of M&Ms taken (small – large) is less than 0.

TABLE 7.5 Summary statistics, including the difference (small – large) in the number of M&Ms taken between the two bowl sizes

Bowl size	Sample size, n	Sample mean	Sample SD
Small	17	$\bar{x}_s = 38.59$	$s_s = 16.90$
Large	17	$\bar{x}_l = 49.47$	$s_l = 27.21$
Difference = small – large	17	$\bar{x}_d = -10.88$	$s_d = 36.30$

How Many M&Ms Would You Like?

- Here are the results of a simulation-based test.
- The p-value is quite large at 0.1220.

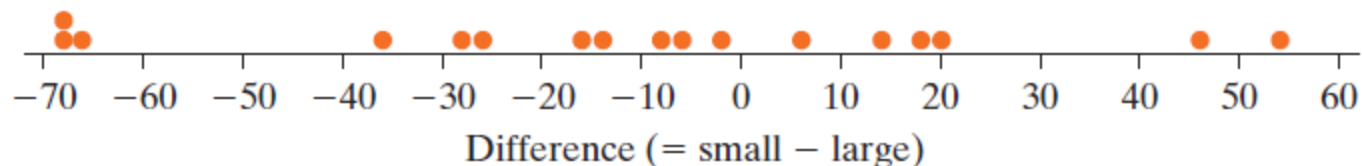


Count samples:

Count = 122/1000 (0.1220)

How Many M&Ms Would You Like?

- Our null distribution was centered at zero and fairly bell-shaped.
- Theory-based methods using the t distribution should be valid if σ is unknown and the population distribution of differences is normal (we can guess at this by looking at the sample distribution of differences). Alternatively, we can use the normal distribution if our sample size is at least 30.
- Our sample size was only 17, but this distribution of differences looks pretty normal, so we will proceed with a t-test.



Theory-based test

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n}}$$

- This kind of test is called a paired t -test.

Theory-based results

Scenario:

☐ Paste data

n:

mean, \bar{x} :

sample sd, s:

☒ Confidence interval

confidence level %

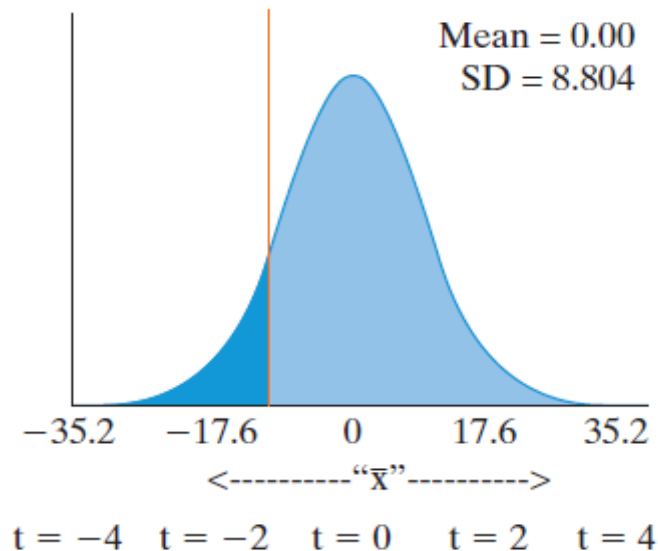
(-29.5435, 7.7835)

Theory-based inference

☒ Test of significance

$H_0: \mu =$

$H_a: \mu <$



Standardized statistic df = 16

p-value

Conclusion

- The theory-based test gives slightly different results than simulation, 11.7% instead of 12.2% for the p-value, but we come to the same conclusion. We do not have strong evidence that the bowl size affects the number of M&Ms taken.
- We can see this in the large p-value (0.1172) and the confidence interval that included zero (-29.5, 7.8).
- The confidence interval tells us that we are 95% confident that when given a small bowl, people will take somewhere between 29.5 fewer M&Ms to 7.8 more M&Ms on average than when given a large bowl.

Why wasn't the difference statistically significant?

- There could be a number of reasons we didn't get significant results.
 - Maybe bowl size doesn't matter.
 - Maybe bowl size does matter and the difference was too small to detect with our small sample size.
 - Maybe bowl size does matter with some foods, like pasta or cereal, but not with a snack food like M&Ms.

Strength of Evidence

- We will have stronger evidence against the null (smaller p-value) when:
 - The sample size is increased.
 - The variability of the data is reduced.
 - The effect size, or mean difference, is farther from 0.
- We will get a narrower confidence interval when:
 - The sample size is increased.
 - The variability of the data is reduced.
 - The confidence level is decreased.

2. Multiple testing and publication bias.

A p-value is the probability, assuming the null hypothesis of no relationship is true, that you will see a difference as extreme as, or more extreme than, you observed.

So, when you are looking at unrelated things, 5% of the time you will find a statistically significant relationship.

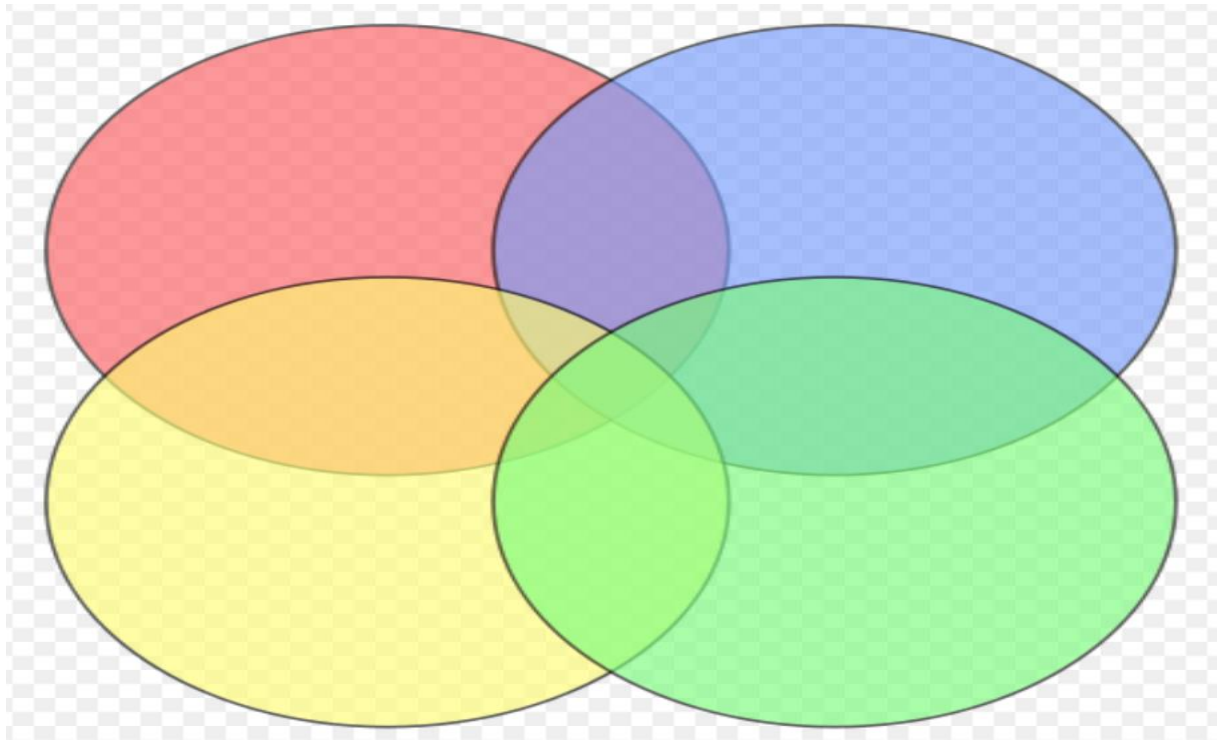
This underscores the need for followup confirmation studies. If testing many explanatory variables simultaneously, it can become very likely to find something significant even if nothing is actually related to the response variable.

Multiple testing and publication bias.

- * For example, if the significance level is 5%, then for 100 tests where all null hypotheses are true, the expected number of incorrect rejections (Type I errors) is 5. If the tests are independent, the probability of at least one Type I error would be 99.4%. $P(\text{no Type I errors}) = .95^{100} = 0.6\%$.

- * To address this problem, scientists sometimes change the significance level so that, under the null hypothesis that none of the explanatory variables is related to the response variable, the probability of rejecting at least one of them is 5%.

- * One way is to use Bonferroni's correction: with m explanatory variables, use significance level $5\%/m$. $P(\text{at least 1 Type I error}) \text{ will be } \leq m (5\%/m) = 5\%$.



$P(\text{Type I error on explanatory 1}) = 5\%/m.$

$P(\text{Type I error on explanatory 2}) = 5\%/m.$

$P(\text{Type 1 error on at least one explanatory}) \leq$

$P(\text{error on 1}) + P(\text{error on 2}) + \dots + P(\text{error on } m) = m \times 5\%/m.$

Multiple testing and publication bias.

Imagine a scenario where a drug is tested many times to see if it reduces the incidence of some response variable. If the drug is tested 100 times by 100 different researchers, the results will be stat. sig. about 5 times.

If only the stat. sig. results are published, then the published record will be very misleading.

Multiple testing and publication bias.

A drug called Reboxetine made by Pfizer was approved as a treatment for depression in Europe and the UK in 2001, based on positive trials.

A meta-analysis in 2010 found that it was not only ineffective but also potentially harmful. The report found that 74% of the data on patients who took part in the trials of Reboxetine were not published because the findings were negative. Published data about reboxetine overestimated its benefits and underestimated its harm.

A subsequent 2011 analysis indicated Reboxetine might be effective for severe depression though.

3. Two quantitative variables.

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

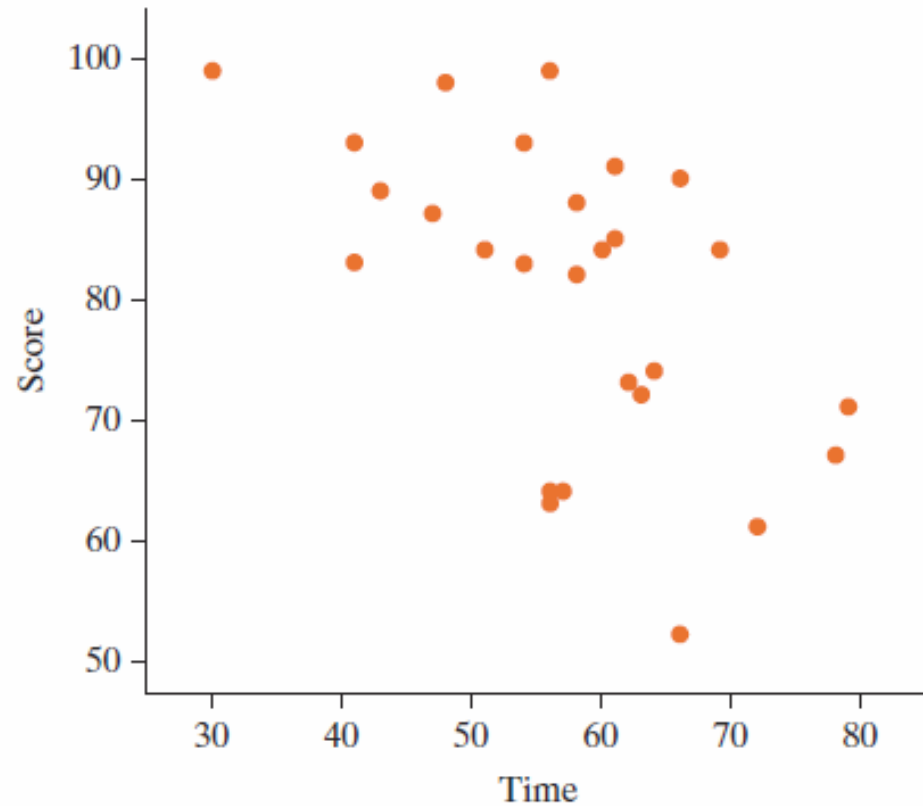
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

Time	30	41	41	43	47	48	51	54	54	56	56	56	57	58
Score	100	84	94	90	88	99	85	84	94	100	65	64	65	89
Time	58	60	61	61	62	63	64	66	66	69	72	78	79	
Score	83	85	86	92	74	73	75	53	91	85	62	68	72	

Scatterplot

Put explanatory variable on the horizontal axis.

Put response variable on the vertical axis.



Describing Scatterplots

- When we describe data in a scatterplot, we describe the
 - Direction (positive or negative)
 - Form (linear or not)
 - Strength (strong-moderate-weak, we will let correlation help us decide)
 - Unusual Observations
- How would you describe the time and test scatterplot?

Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
- Correlation is a number between -1 and 1.
- With positive correlation one variable increases, on average, as the other increases.
- With negative correlation one variable decreases, on average, as the other increases.
- The closer it is to either -1 or 1 the closer the points fit to a line.
- The correlation for the test data is -0.56.

Correlation Guidelines

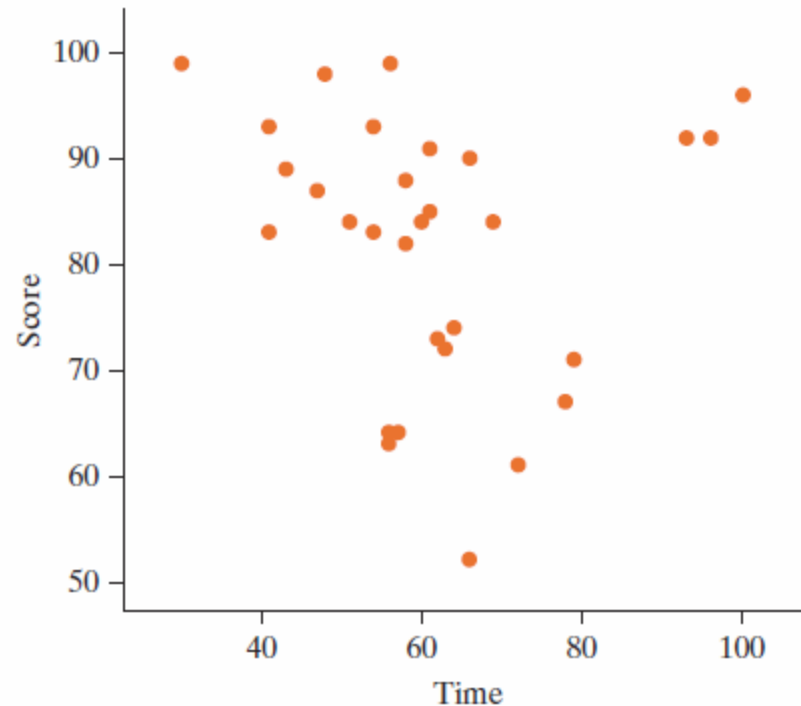
Correlation Value	Strength of Association	What this means
0.7 to 1.0	Strong	The points will appear to be nearly a straight line
0.3 to 0.7	Moderate	When looking at the graph the increasing/decreasing pattern will be clear, but there is considerable scatter.
0.1 to 0.3	Weak	With some effort you will be able to see a slightly increasing/decreasing pattern
0 to 0.1	None	No discernible increasing/decreasing pattern

Same Strength Results with Negative Correlations

Back to the test data

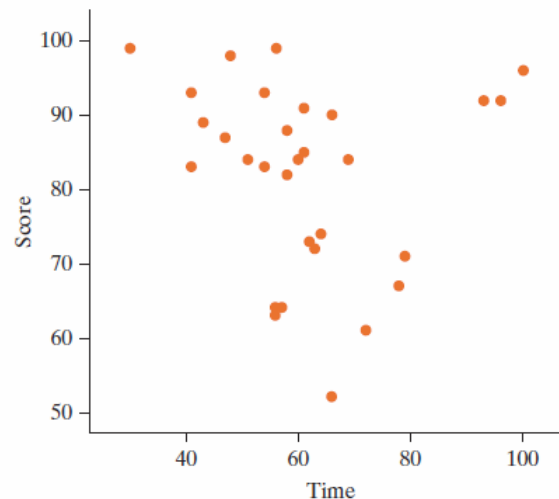
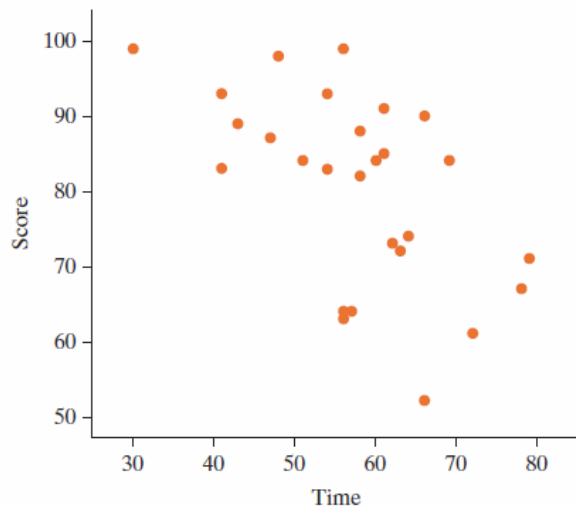
Actually the last three people to finish the test had scores of 93, 93, and 97.

What does this do
to the correlation?



Influential Observations

- The correlation changed from -0.56 (a fairly moderate negative correlation) to -0.12 (a weak negative correlation).
- Points that are far to the left or right and not in the overall direction of the scatterplot can greatly change the correlation. (influential observations)



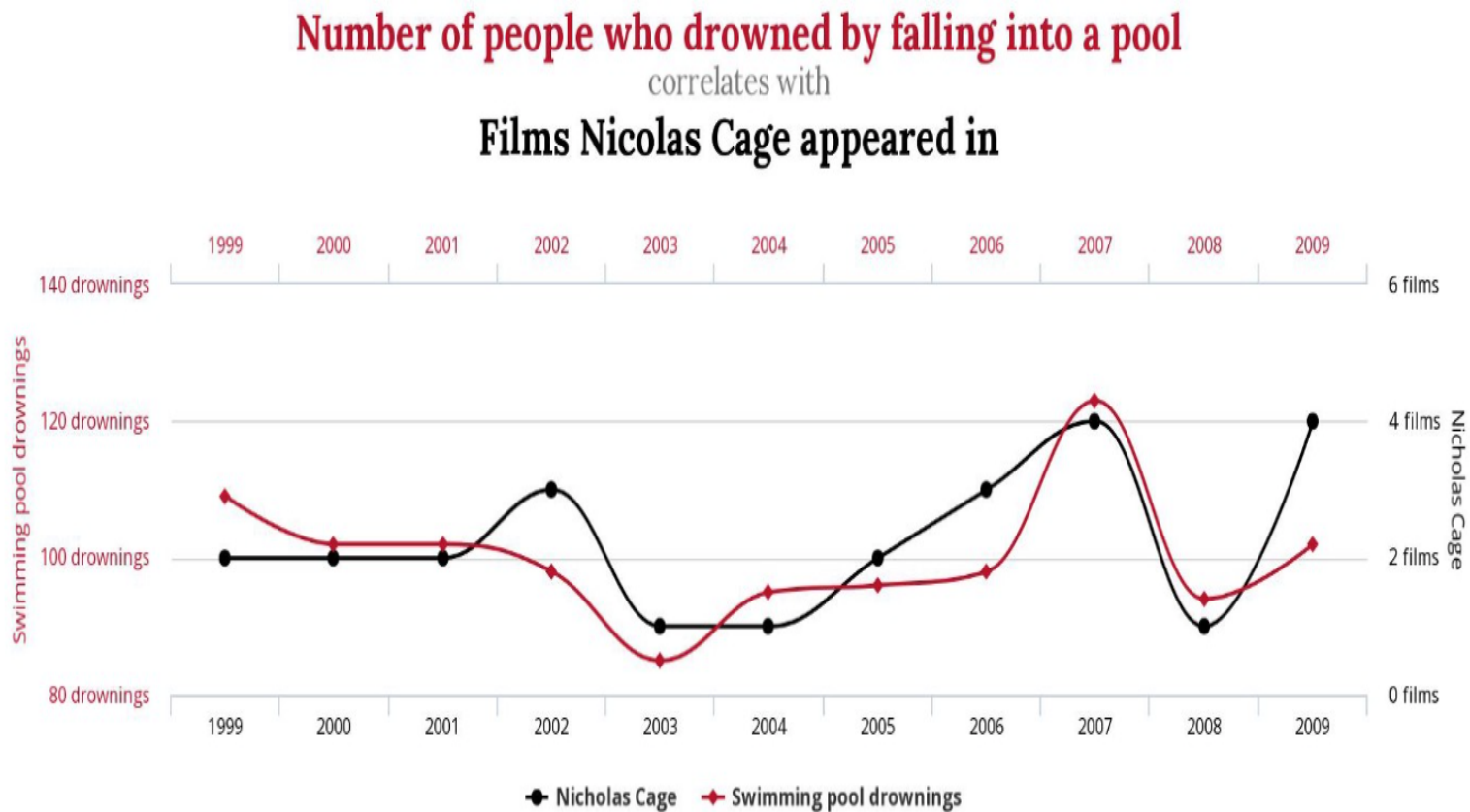
Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
 - $-1 \leq r \leq 1$
 - Correlation makes no distinction between explanatory and response variables.
 - Correlation has no units.
 - Correlation is not resistant to outliers. It is sensitive.

Learning Objectives for Section 10.1

- Summarize the characteristics of a scatterplot by describing its direction, form, strength and whether there are any unusual observations.
- Recognize that the correlation coefficient is appropriate only for summarizing the strength and direction of a scatterplot that has linear form.
- Recognize that a scatterplot is the appropriate graph for displaying the relationship between two quantitative variables and create a scatterplot from raw data.
- Recognize that a correlation coefficient of 0 means there is no linear association between the two variables and that a correlation coefficient of -1 or 1 means that the scatterplot is exactly a straight line.
- Understand that the correlation coefficient is influenced by extreme observations.

Note that correlation \neq causation.



tylervigen.com

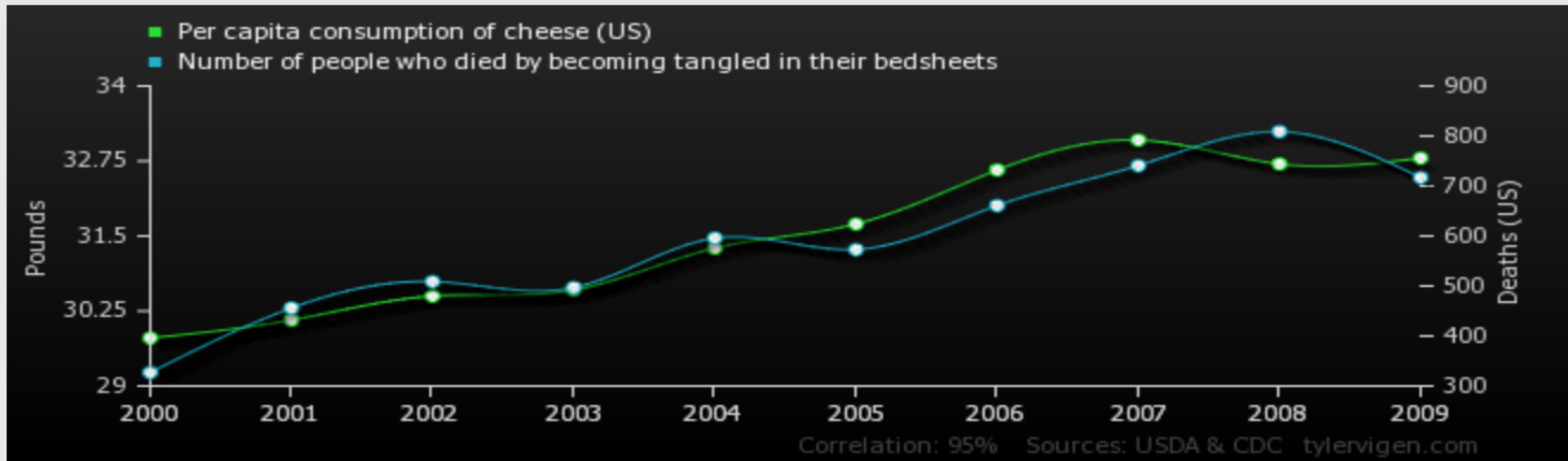
from: <http://tylervigen.com>

Note that correlation \neq causation.

Per capita consumption of cheese (US)

correlates with

Number of people who died by becoming tangled in their bedsheets

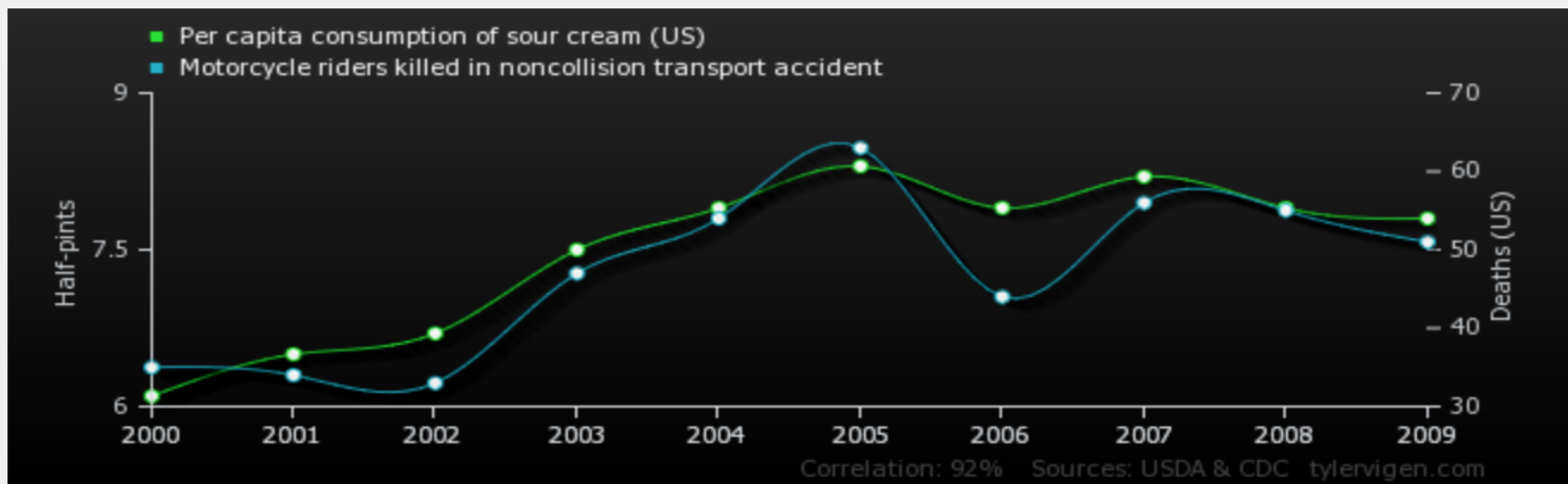


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

Note that correlation \neq causation.

Per capita consumption of sour cream (US)
correlates with
Motorcycle riders killed in noncollision transport accident



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of sour cream (US) Half-pints (USDA)	6.1	6.5	6.7	7.5	7.9	8.3	7.9	8.2	7.9	7.8
Motorcycle riders killed in noncollision transport accident Deaths (US) (CDC)	35	34	33	47	54	63	44	56	55	51

Correlation: 0.916391

Inference for the Correlation Coefficient: Simulation-Based Approach

Section 10.2

We will look at a small sample example to see if body temperature is associated with heart rate.

Temperature and Heart Rate

Hypotheses

- Null: There is no association between heart rate and body temperature. ($\rho = 0$)
- Alternative: There is a positive linear association between heart rate and body temperature. ($\rho > 0$)

$\rho = \text{rho}$

Inference for Correlation with Simulation

(Section 10.2)

1. Compute the observed statistic. (Correlation)
2. Scramble the response variable, compute the simulated statistic, and repeat this process many times.
3. Reject the null hypothesis if the observed statistic is in the tail of the null distribution.

Temperature and Heart Rate

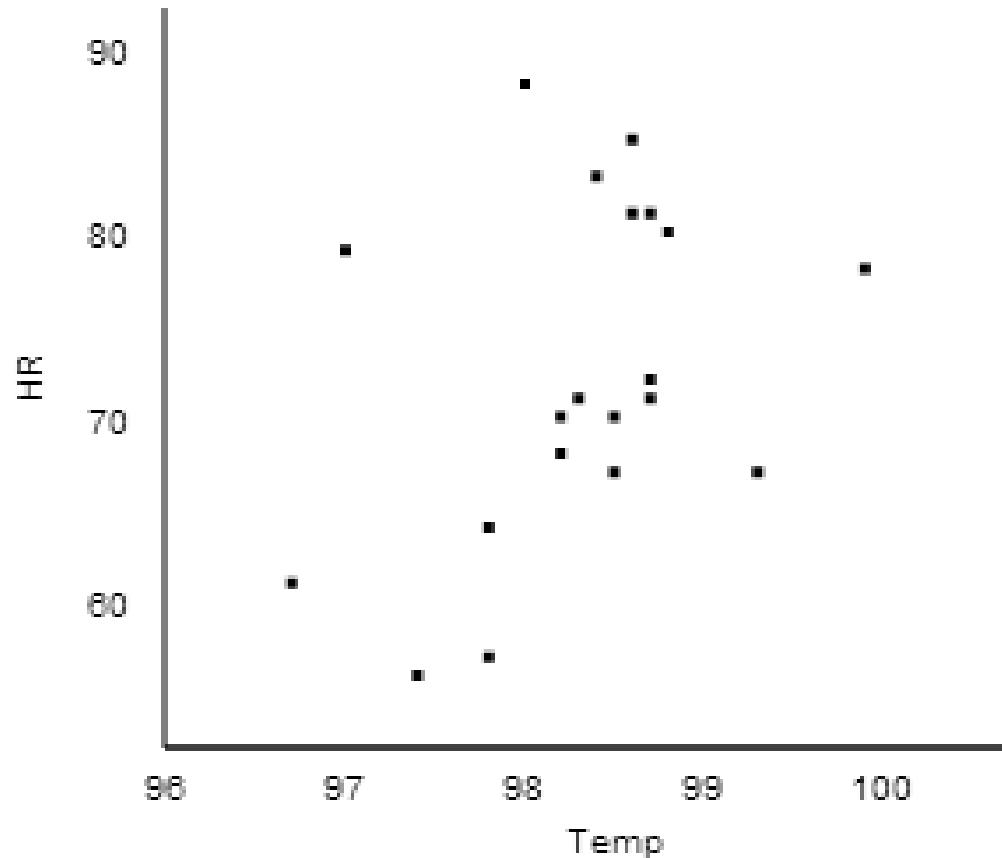
Collect the Data

Tmp	98.3	98.2	98.7	98.5	97.0	98.8	98.5	98.7	99.3	97.8
HR	72	69	72	71	80	81	68	82	68	65
Tmp	98.2	99.9	98.6	98.6	97.8	98.4	98.7	97.4	96.7	98.0
HR	71	79	86	82	58	84	73	57	62	89

Temperature and Heart Rate

Explore the Data

$r = 0.378$



Temperature and Heart Rate

- If there was no association between heart rate and body temperature, what is the probability we would get a correlation as high as 0.378 just by chance?
- If there is no association, we can break apart the temperatures and their corresponding heart rates. We will do this by shuffling one of the variables.

Shuffling Cards

- Let's remind ourselves what we did with cards to find our simulated statistics.
- With two proportions, we wrote the response on the cards, shuffled the cards and placed them into two piles corresponding to the two categories of the explanatory variable.
- With two means we did the same thing except this time the responses were numbers instead of words.

Dolphin Therapy

Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver

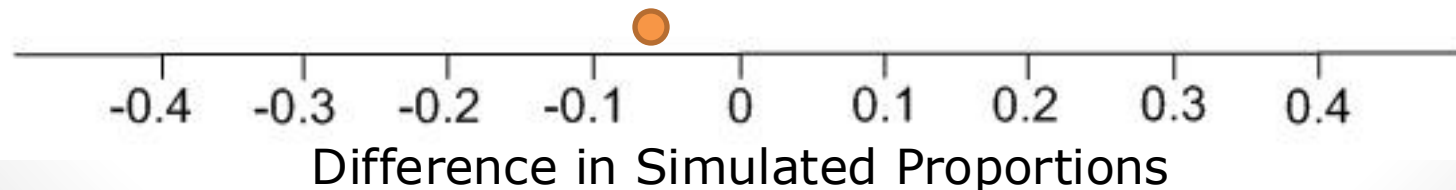
60.0%
Improvers

Control

Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

20.0%
Improvers

$$0.400 - 0.467 = -0.067$$



Music

No music

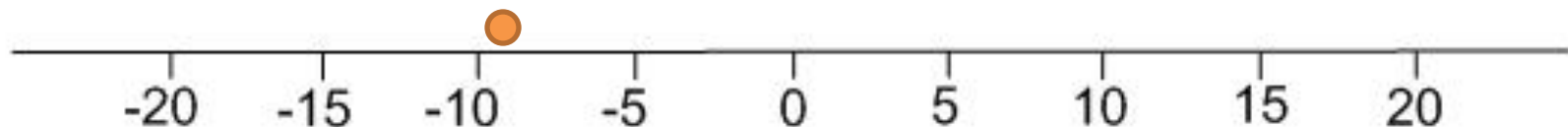
25.2	45.6
14.5	11.6
-7.0	18.6
12.6	12.1
34.5	30.5

mean = 6.38

-10.7	-10.7	10.0
4.5	9.6	
2.2	2.4	
21.3	21.8	
-14.7	7.2	

mean = 16.12

$$6.38 - 16.12 = -9.74$$



Difference in Simulated Means

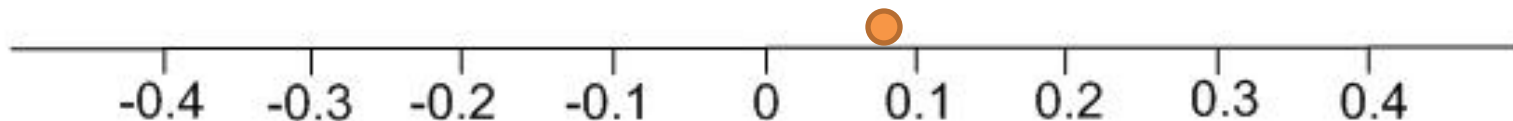
Shuffling Cards

- Now how will this shuffling be different when both the response and the explanatory variable are quantitative?
- We can't put things in two piles anymore.
- We still shuffle values of the response variable, but this time place them next to two values of the explanatory variable.

Body Temperature and Heart Rate

98.3° 72	98.2° 69	97.7° 72	98.5° 71	97.0° 80	98.8° 81	98.5° 68	98.7° 82	99.3° 68	97.8° 65
98.2° 71	99.9° 79	98.6° 86	98.6° 82	97.8° 58	98.4° 84	98.7° 73	97.4° 57	96.7° 62	98.0° 89

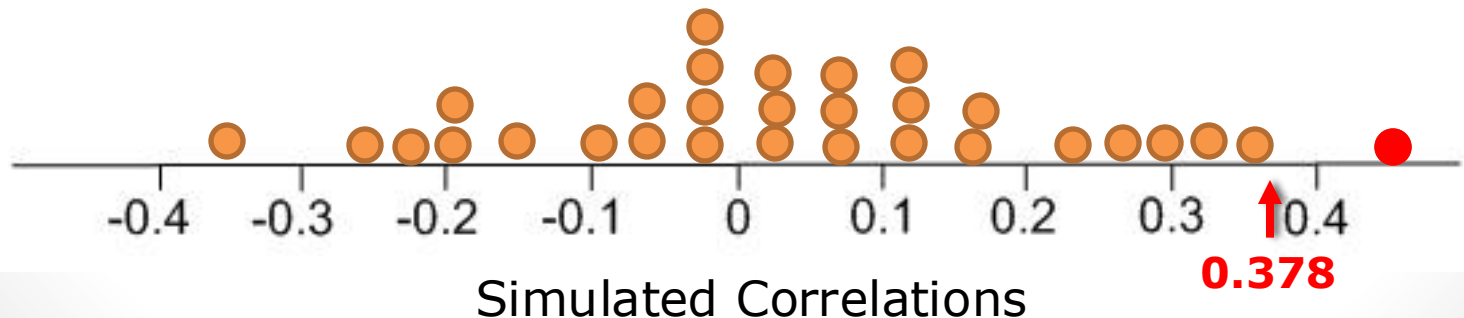
$$r = 0.078$$



Simulated Correlations

More Simulations

Only one simulated statistic out of 30 was as large or larger than our observed correlation of 0.378, hence our p-value for this null distribution is $1/30 \approx 0.03$.

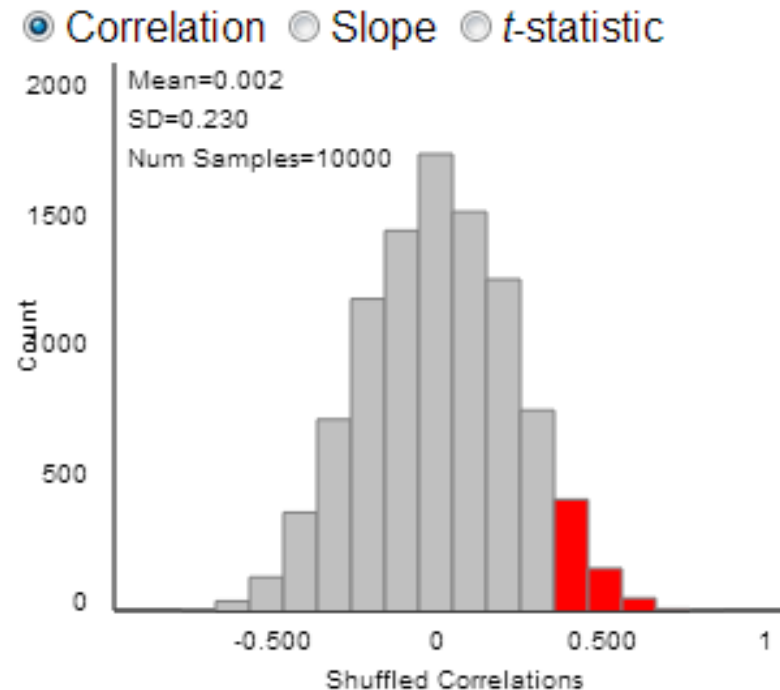


Temperature and Heart Rate

- We can look at the output of 1000 shuffles with a distribution of 1000 simulated correlations.

Temperature and Heart Rate

- Notice our null distribution is centered at 0 and somewhat symmetric.
- We found that 530/10000 times we had a simulated correlation greater than or equal to 0.378.



Count Samples
Count = 530/10000 (0.0530)

Temperature and Heart Rate

- With a p-value of $0.053 = 5.3\%$, we almost but do not quite have statistical significance. We observe a positive linear association between body temperature and heart rate but this association is not statistically significant. Perhaps a larger sample should be investigated to get a better idea if the two variables are related or not.

4. Linear Regression

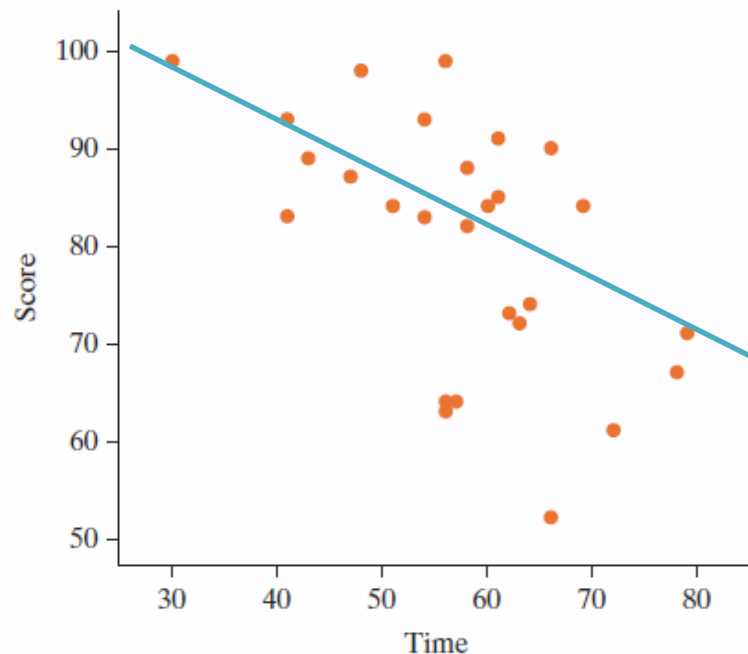
Section 10.3

Introduction

- If we decide an association is linear, it is helpful to develop a mathematical model of that association.
- Helps make predictions about the response variable.
- The *least-squares regression line* is the most common way of doing this.

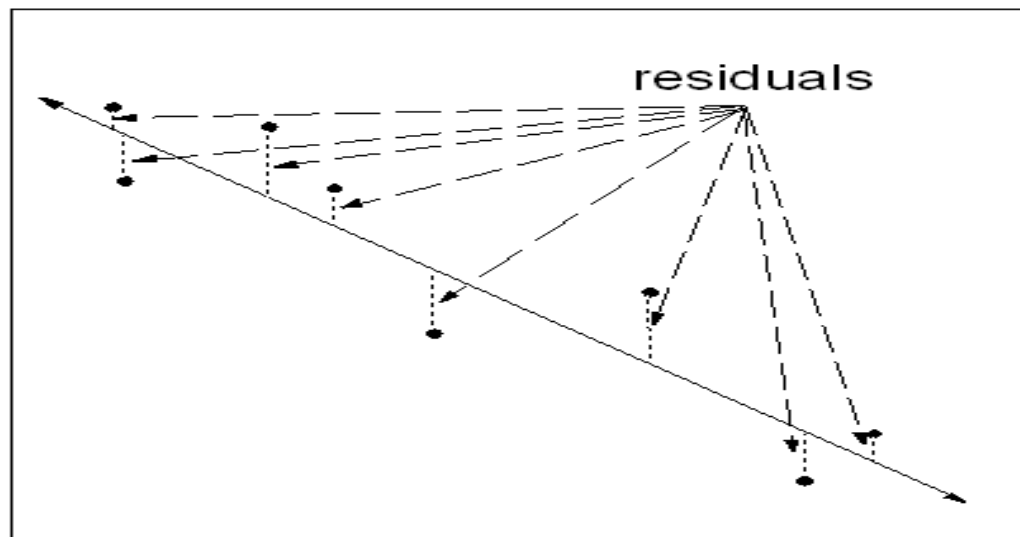
Introduction

- Unless the points are perfectly linearly aligned, there will not be a single line that goes through every point.



Introduction

- We want a line that minimizes the vertical distances between the line and the points
 - These distances are called **residuals**.
 - The line we will find actually minimizes the sum of the squares of the residuals.
 - This is called a **least-squares regression line**.

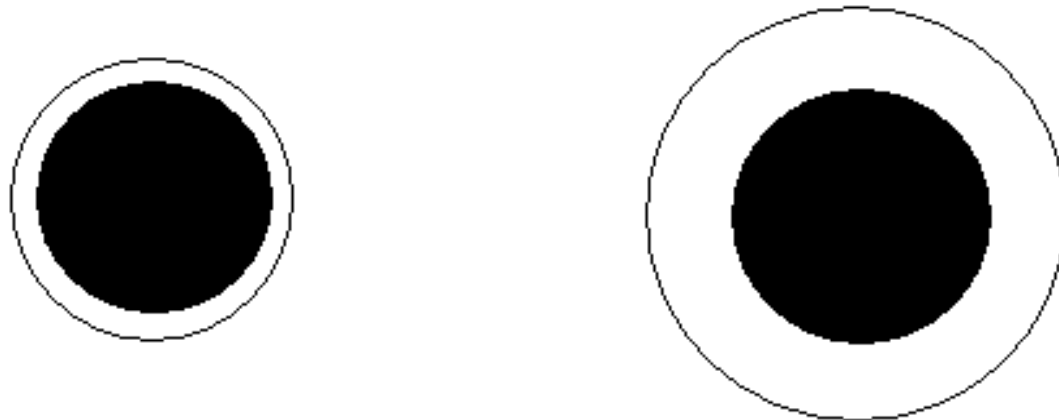


Are Dinner Plates Getting Larger?

Example 10.3

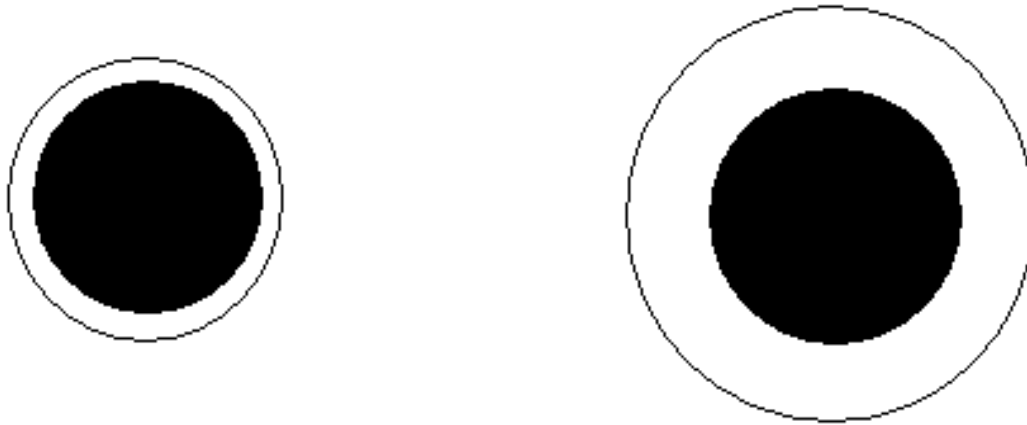
Growing Plates?

- There are many recent articles and TV reports about the obesity problem.
- One reason some have given is that the size of dinner plates are increasing.
- Are these black circles the same size, or is one larger than the other?



Growing Plates?

- They appear to be the same size for many, but the one on the right is about 20% larger than the left.



- This suggests that people will put more food on larger dinner plates without knowing it.
- There is name for this phenomenon: *Delboeuf illusion*.

Growing Plates?

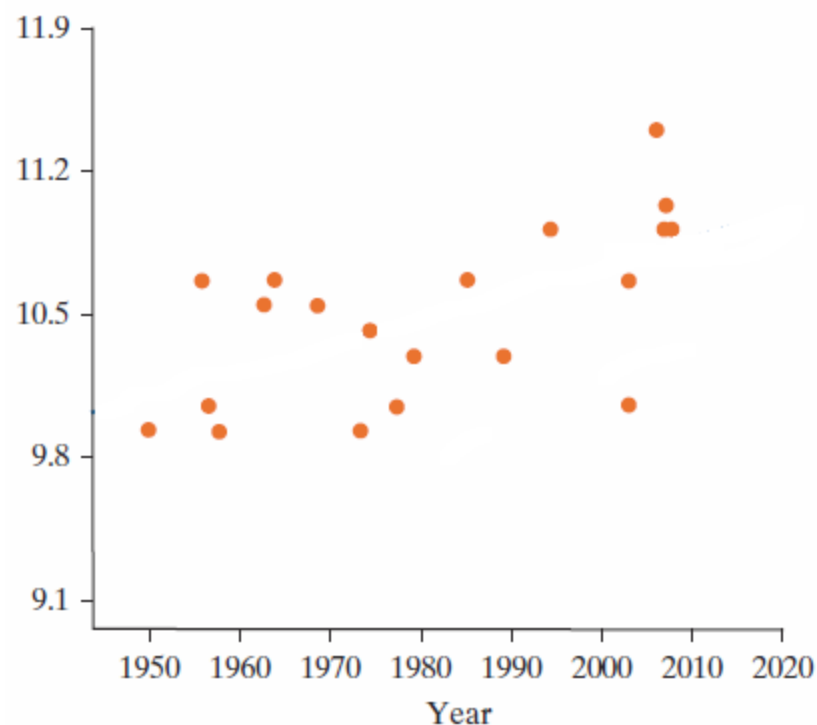
- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on ebay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

TABLE 10.1 Data for size (diameter, in inches) and year of manufacture for 20 American-made dinner plates

Year	1950	1956	1957	1958	1963	1964	1969	1974	1975	1978
Size	10	10.75	10.125	10	10.625	10.75	10.625	10	10.5	10.125
Year	1980	1986	1990	1995	2004	2004	2007	2008	2008	2009
Size	10.375	10.75	10.375	11	10.75	10.125	11.5	11	11.125	11

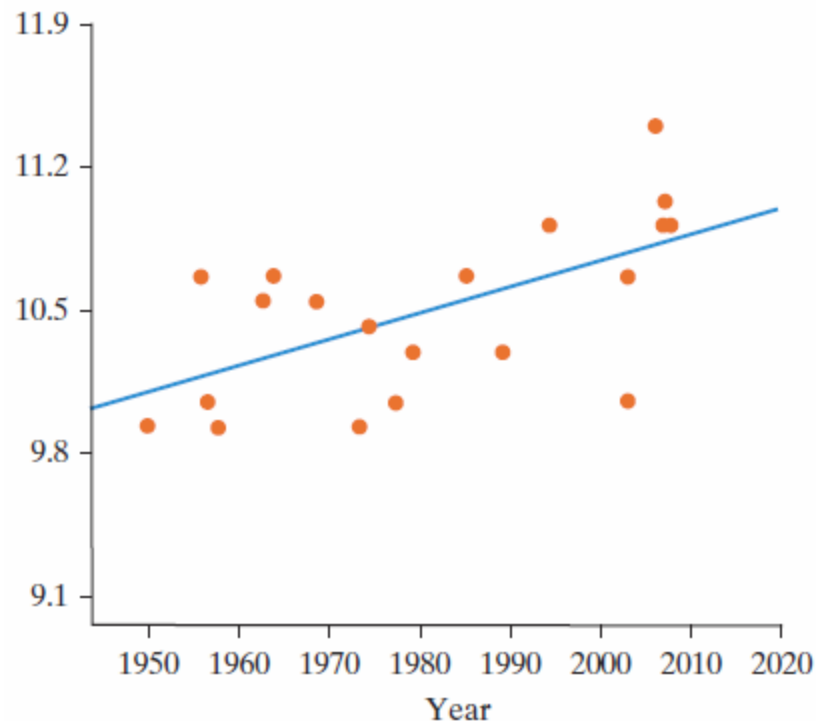
Growing Plates?

- Both year (explanatory variable) and diameter in inches (response variable) are quantitative.
- Each dot in this scatterplot represents one plate.



Growing Plates?

- The association appears to be roughly linear.
- The least squares regression line is added.
- The line slopes upward, but is the slope significant?



Regression Line

The regression equation is $\hat{y} = a + bx$:

- a is the y -intercept
- b is the slope
- x is a value of the explanatory variable
- \hat{y} is the predicted value for the response variable
- For a specific value of x , the corresponding distance $y - \hat{y}$ (or actual – predicted) is a residual

Regression Line

- The least squares line for the dinner plate data is $\hat{y} = -14.8 + 0.0128x$
- Or $\widehat{\text{diameter}} = -14.8 + 0.0128(\text{year})$
- This allows us to predict plate diameter for a particular year.

Slope

$$\hat{y} = -14.8 + 0.0128x$$

- What is the predicted diameter for a plate manufactured in 2000?
 - $-14.8 + 0.0128(2000) = 10.8$ in.
- What is the predicted diameter for a plate manufactured in 2001?
 - $-14.8 + 0.0128(2001) = 10.8128$ in.
- How does this compare to our prediction for the year 2000?
 - 0.0128 larger
- Slope $b = 0.0128$ means that diameters are predicted to increase by 0.0128 inches per year on average

Slope

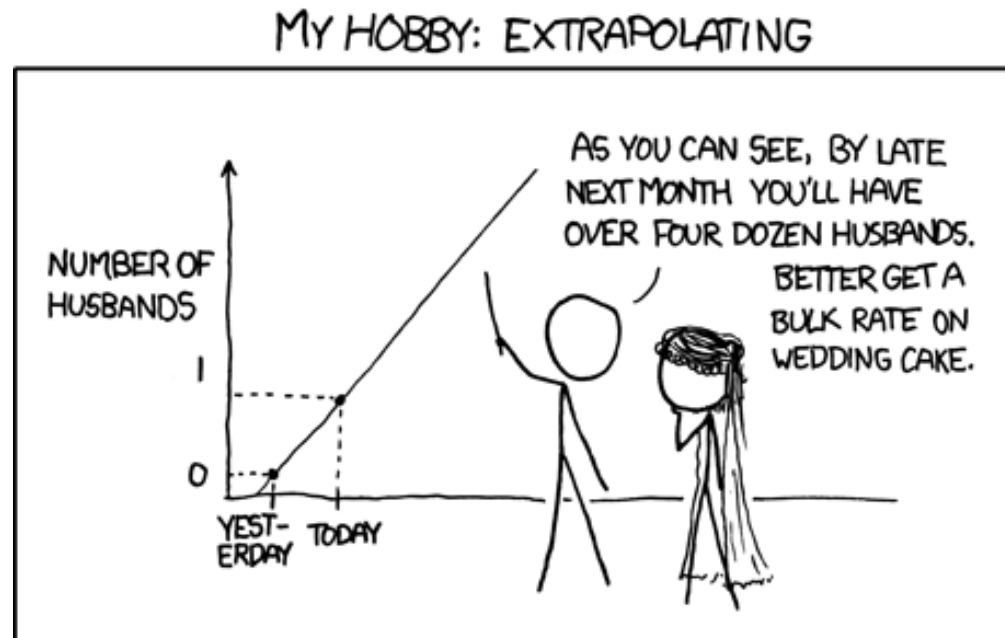
- Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
- Both the slope and the correlation coefficient for this study were positive.
 - The slope is 0.0128
 - The correlation is 0.604
- The slope and correlation coefficient will always have the same sign.

y-intercept

- The y-intercept is where the regression line crosses the y-axis. It is the predicted response when the explanatory variable equals 0.
- We had a y-intercept of -14.8 in the dinner plate equation. What does this tell us about our dinner plate example?
 - Dinner plates in year 0 would be predicted to be -14.8 inches???
- How can it be negative?
 - The equation works well within the range of values given for the explanatory variable, but fails outside that range.
- Our equation should only be used to predict the size of dinner plates from about 1950 to 2010.

Extrapolation

- Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is called *extrapolation*.



Coefficient of Determination

- While the intercept and slope have meaning in the context of year and diameter, remember that the correlation does not. It is just 0.604.
- However, the square of the correlation (coefficient of determination or r^2) does have meaning.
- $r^2 = 0.604^2 = 0.365$ or 36.5%
- 36.5% of the variation in plate size (the response variable) can be explained by its linear association with the year (the explanatory variable).

Learning Objectives for Section 10.3

- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for a given value of the explanatory variable.
- Understand the concept of residual and find and interpret the residual for an observational unit given the raw data and the equation of the best fit (regression) line.
- Understand the relationship between residuals and strength of association and that the best-fit (regression) line this minimizes the sum of the squared residuals.

Learning Objectives for Section 10.3

- Find and interpret the coefficient of determination (r^2) as the squared correlation and as the percent of total variation in the response variable that is accounted for by the linear association with the explanatory variable.
- Understand that extrapolation is when a regression line is used to predict values outside of the range of observed values for the explanatory variable.
- Understand that when slope = 0 means no association, slope < 0 means negative association, slope > 0 means positive association, and that the sign of the slope will be the same as the sign of the correlation coefficient.
- Understand that influential points can substantially change the equation of the best-fit line.

5. Calculating correlation, r.

ρ = rho = correlation of the population.

Suppose there are N people in the population,

X = temperature, Y = heart rate,

the mean and sd of temp in the pop. are μ_x and σ_x ,

and the pop. mean and sd of heart rate are μ_y and σ_y .

$$\rho = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu_x}{\sigma_x} \right) \left(\frac{y_i - \mu_y}{\sigma_y} \right).$$

Given a sample of size n, we estimate ρ using

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

This is in Appendix A.

6. Slope of regression line.

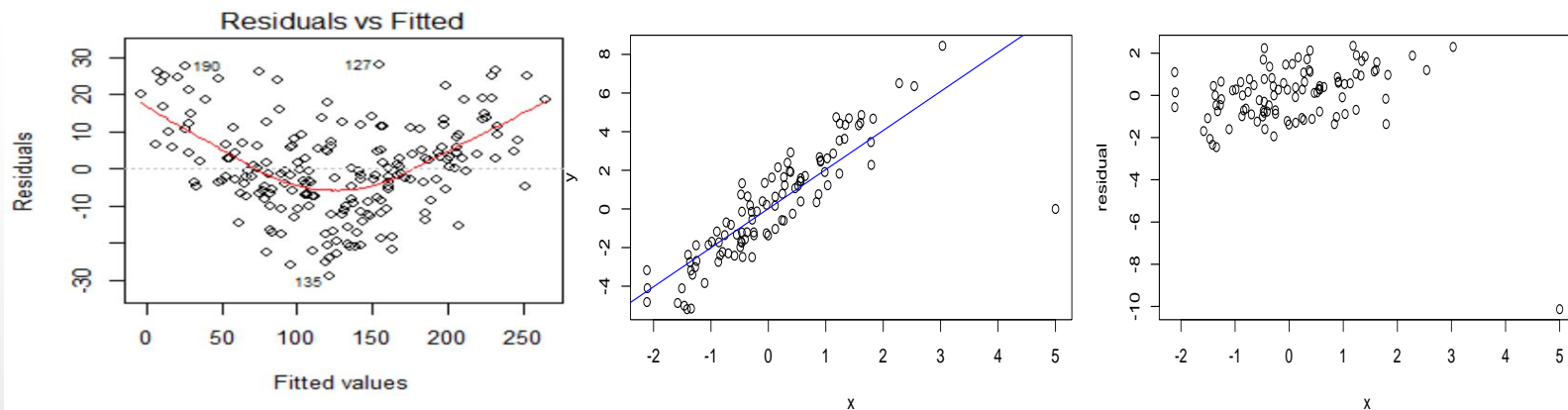
- Suppose $\hat{y} = a + bx$ is the regression line.
- The slope b of the regression line is $b = r \frac{s_y}{s_x}$.

This is usually the thing of primary interest to interpret, as the predicted increase in y for every unit increase in x .

- Beware of assuming causation though, esp. with observational studies. Be wary of extrapolation too.
- The intercept $a = \bar{y} - b \bar{x}$.
- The SD of the residuals is $\sqrt{1 - r^2} s_y$.
This is a good estimate of how much the regression predictions will typically be off by.

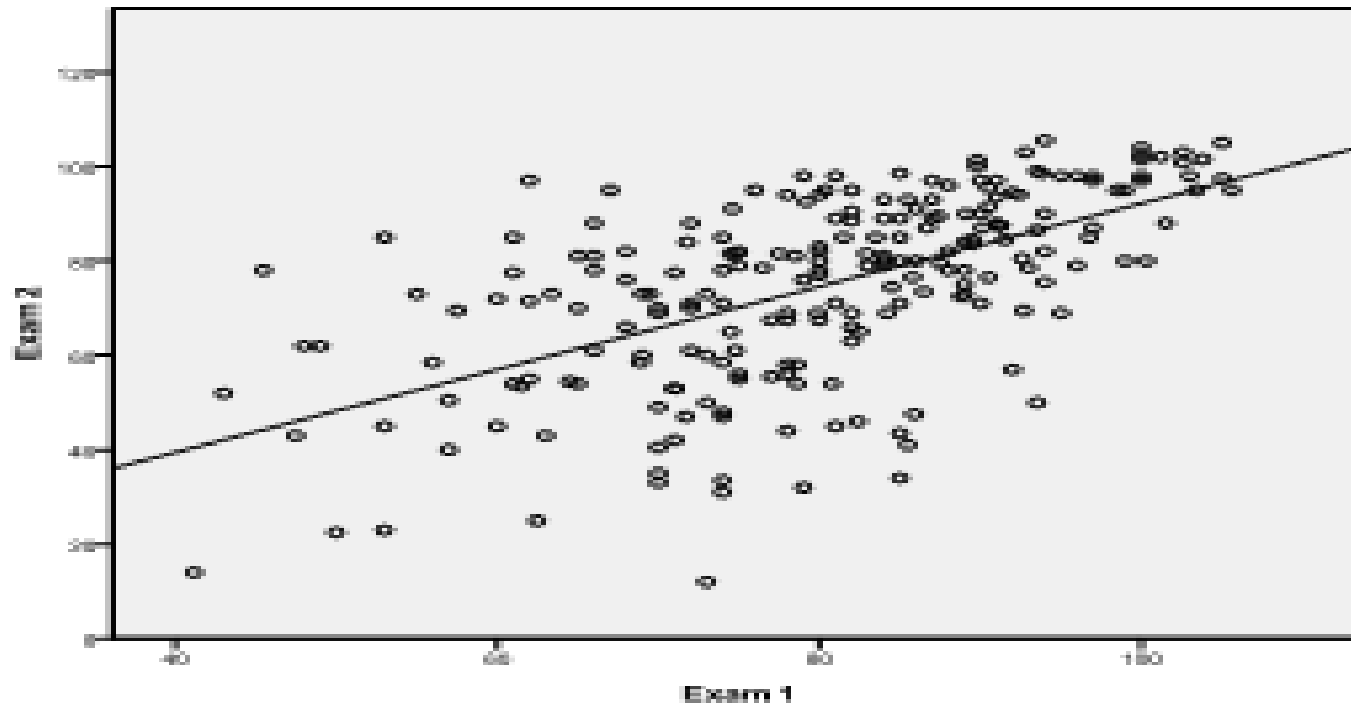
7. How well does the line fit?

- r^2 is a measure of fit. It indicates the amount of scatter around the best fitting line.
- $\sqrt{1 - r^2} s_y$ is useful as a measure of how far off predictions would have been on average.
- Residual plots can indicate curvature, outliers, or heteroskedasticity.



- Note that regression residuals have mean zero, whether the line fits well or poorly.

- Heteroskedasticity: when the variability in y is not constant as x varies.



(b)