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

1. Paired data, simulation approach and baseball example.
2. Paired data, theory based approach and bowlsize example.
3. Multiple testing, publication bias, and Reboxitine example.
4. Two variables and correlation.

HW3, due Fri, May22, 1159pm. 4.CE.10, 5.3.28, 6.1.17, and 6.3.14.
In 5.3.28d, use the theory-based formula. You do not need to use an applet.
Read ch7 and 10.
The course website is http://www.stat.ucla.edu/~frederic/13/S26 .

1. Simulation based Approach for Analyzing Paired Data, and rounding first base example.

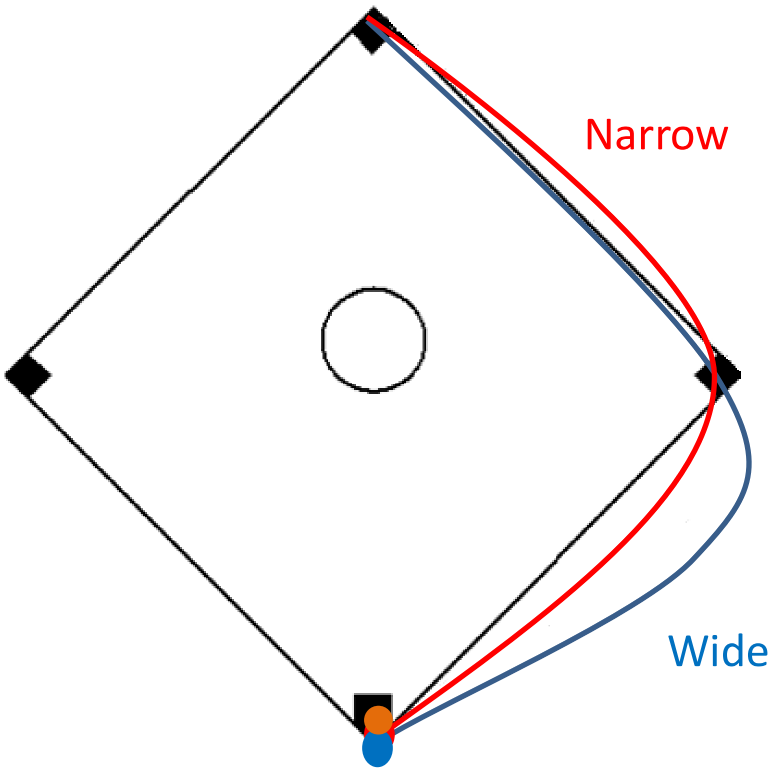
Section 7.2

Rounding First Base

Example 7.2

Rounding First Base

- Imagine you've hit a line drive and are trying to reach second base.
- Does the path that you take to round first base make much of a difference?
  - Narrow angle
  - Wide angle



Rounding First Base

- Woodward (1970) investigated these base running strategies.
- He timed 22 different runners from a spot 35 feet past home to a spot 15 feet before second.
- Each runner used each strategy (paired design), with a rest in between.
- He used random assignment to decide which path each runner should do first.
- **This paired design controls for the runner-to-runner variability.**

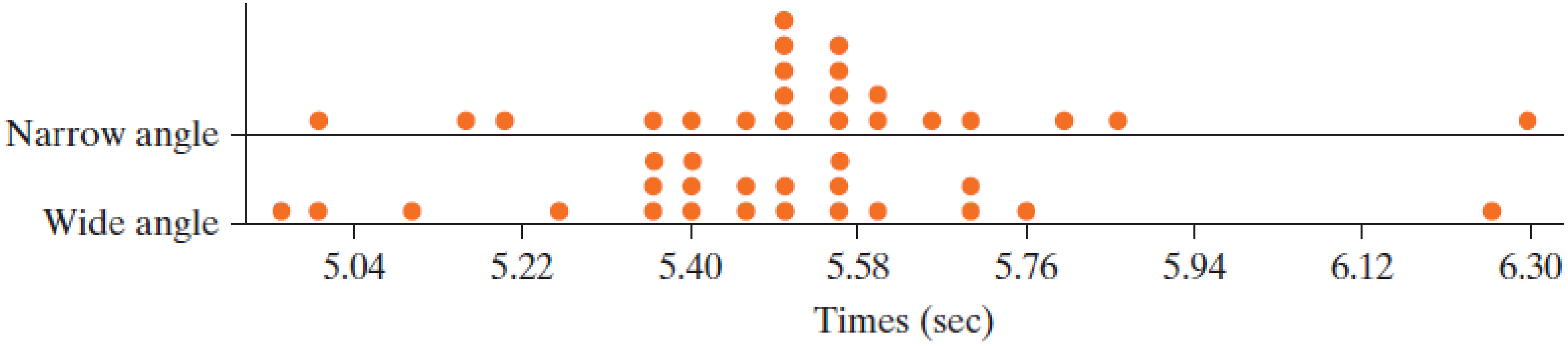
First Base

- What are the observational units in this study?
  - The runners (22 total)
- What variables are recorded? What are their types and roles?
  - Explanatory variable: base running method: wide or narrow angle (categorical)
  - Response variable: time from home plate to second base (quantitative)
- Is this an observational study or an experiment?
  - Randomized experiment.

The results

TABLE 7.1 The running times (seconds) for the first 10 of the 22 subjects

Subject	1	2	3	4	5	6	7	8	9	10	
Narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	...
Wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	...



The Statistics

- There is a lot of overlap in the distributions and substantial variability.

	Mean	SD
Narrow	5.534	0.260
Wide	5.459	0.273

- It is difficult to detect a difference between the methods when these is so much variation.

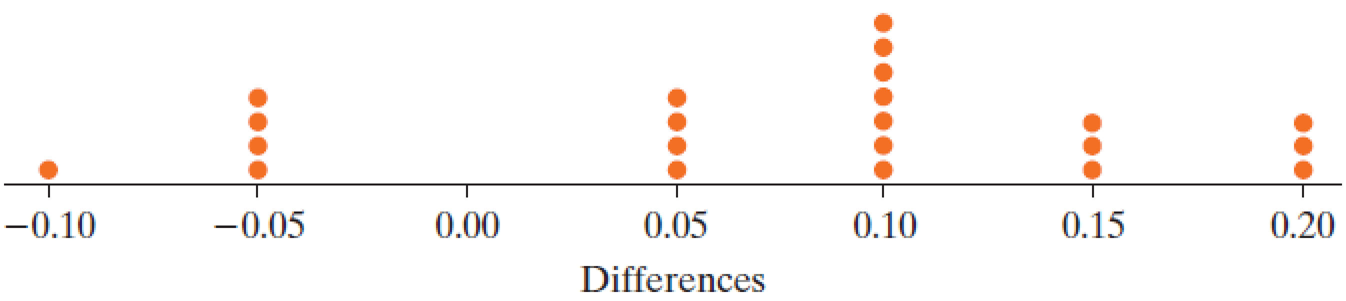
Rounding First Base

- However, these data are clearly paired.
- The paired response variable is time difference in running between the two methods and we can use this in analyzing the data.

The Differences in Times

TABLE 7.2 Last row is difference in times for each of the first 10 runners (narrow − wide)

Subject	1	2	3	4	5	6	7	8	9	10	
Narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	...
Wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	...
Difference	−0.05	−0.05	0.10	0.10	0.15	−0.05	0.05	0.15	0.15	0.10	...

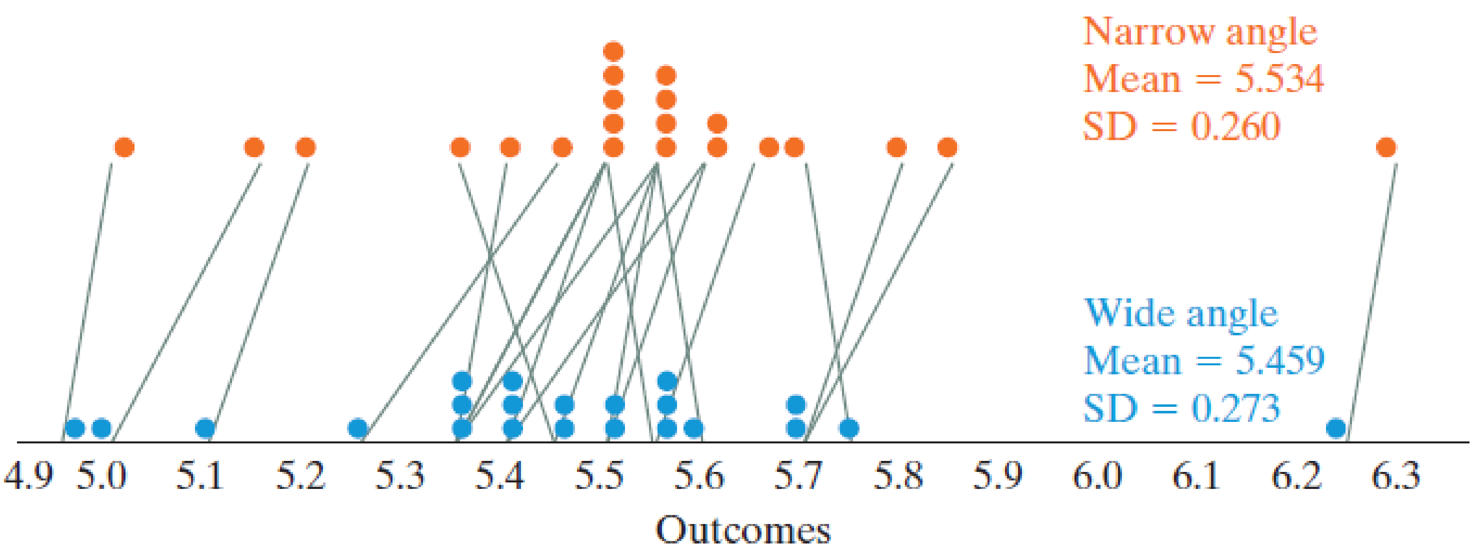


The Differences in Times

- Mean difference is $\bar{x}_{\mathrm{d}} = 0.075$ seconds
- Standard deviation of the differences is $SD_{\mathrm{d}} = 0.0883$ sec.
- This standard deviation of 0.0883 is smaller than the original standard deviations of the running times, which were 0.260 and 0.273.

Rounding First Base

- Below are the original dotplots with each observation paired between the base running strategies.
- What do you notice?



Rounding First Base

- Is the average difference of $\bar{x}_d = 0.075$ seconds significantly different from 0?
- The parameter of interest, μ_d, is the long run mean difference in running times for runners using the narrow angled path instead of the wide angled path. (narrow – wide)

Rounding First Base

The hypotheses:

- H_0: $\mu_d = 0$
  - The long run mean difference in running times is 0.
- H_a: $\mu_d \neq 0$
  - The long run mean difference in running times is not 0.

- The statistic $\bar{x}_d$ = 0.075 is above zero.
- *How likely is it to see an average difference in running times this big or bigger by chance alone, even if the base running strategy has no genuine effect on the times?*

Rounding First Base

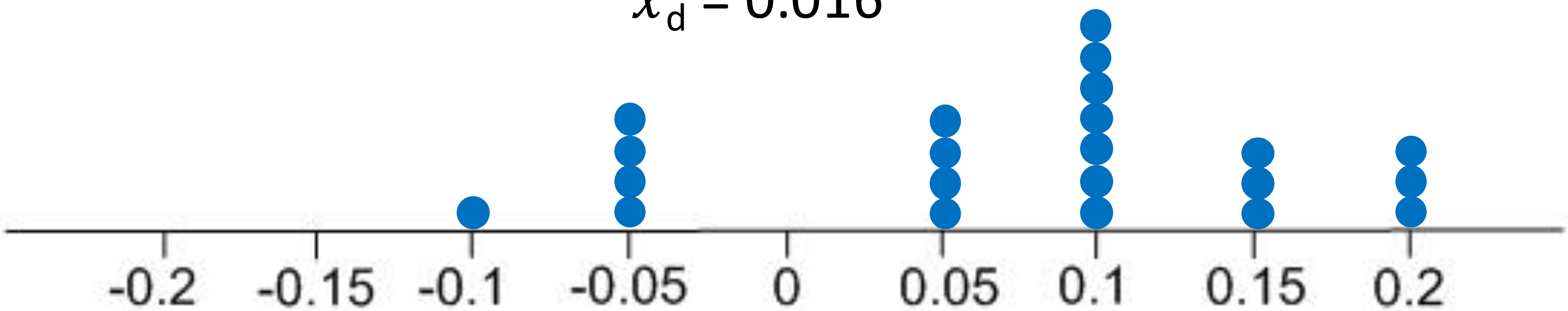
How can we use simulation-based methods to find an approximate p-value?

- The null hypothesis says the running path does not matter.
- So we can use our same data set and, for each runner, randomly decide which time goes with the narrow path and which time goes with the wide path and then compute the difference. (Notice we do not break our pairs.)
- After we do this for each runner, we then compute a mean difference.
- We will then repeat this process many times to develop a null distribution.

Random Swapping

Subject	1	2	3	4	5	6	7	8	9	10	
narrow angle	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
wide angle	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
diff	0.05	-0.05	-0.10	0.10	0.15	0.05	0.05	0.15	0.15	-0.10	…

$\bar{x}_{\text{d}} = 0.016$



More Simulations

0.002 -0.002 0.030 -0.011 -0.007

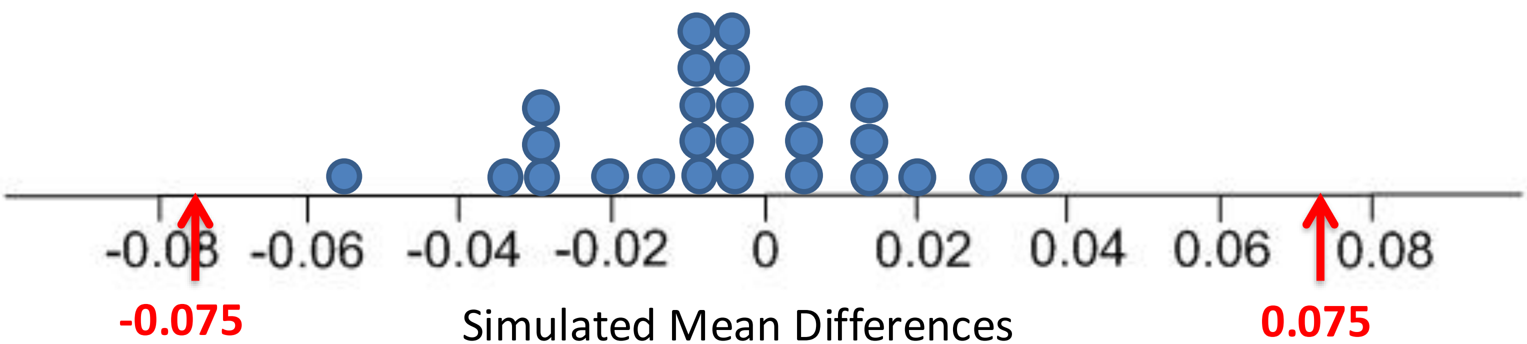
-0.002 -0.016 0.016 -0.007

-0.067 0.002 0.020 -0.007 -0.002

-0.034 -0.030 -0.016 0.002 0.020 0.066

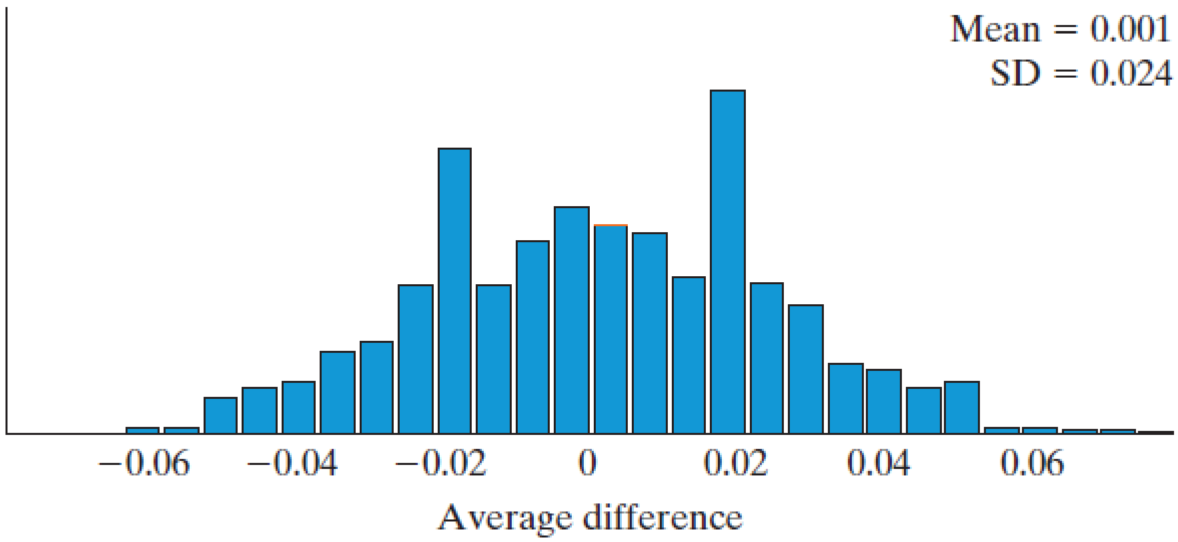
-0.002 -0.002 -0.025

With 26 repetitions of creating simulated mean differences, we did not get any that were as extreme as 0.075.



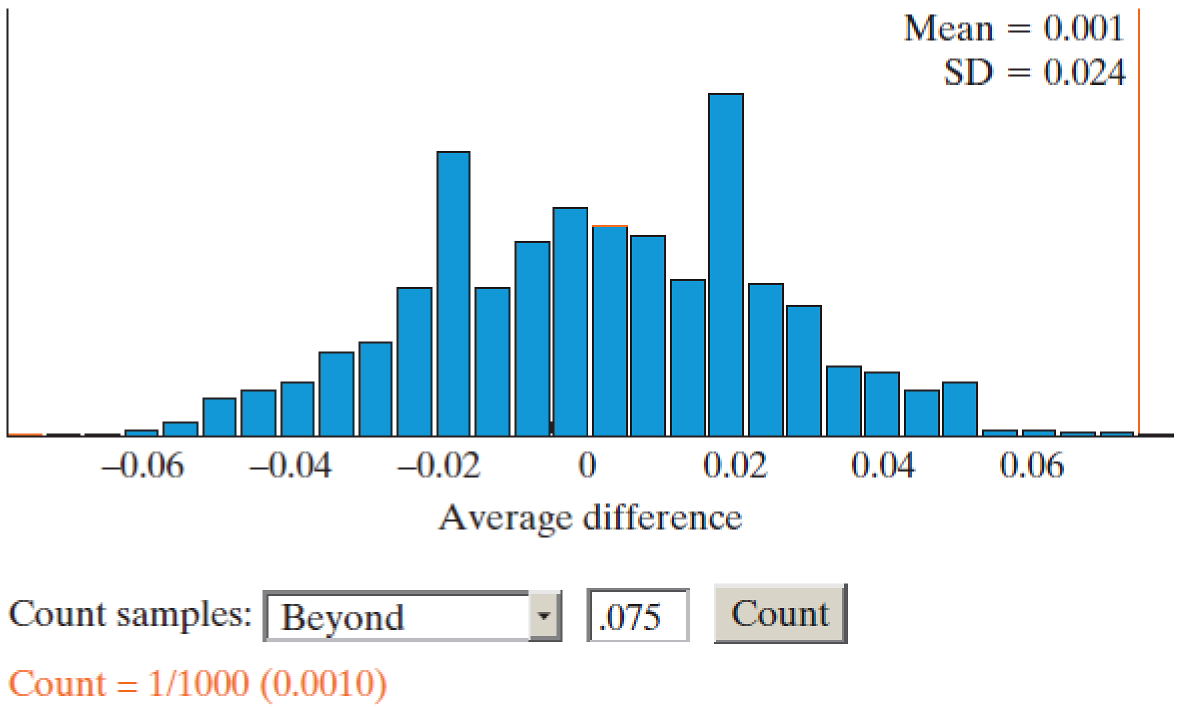
First Base

- Here is a null distribution of 1000 simulated mean differences.
- Notice it is centered at zero, which makes sense in agreement with the null hypothesis.
- Notice also the SD of these MEAN DIFFERENCES is 0.024. This is the SE.
- SD of time differences was 0.0883. SE = SD of mean time diff.s = .024.
- Where is our observed statistic of 0.075?



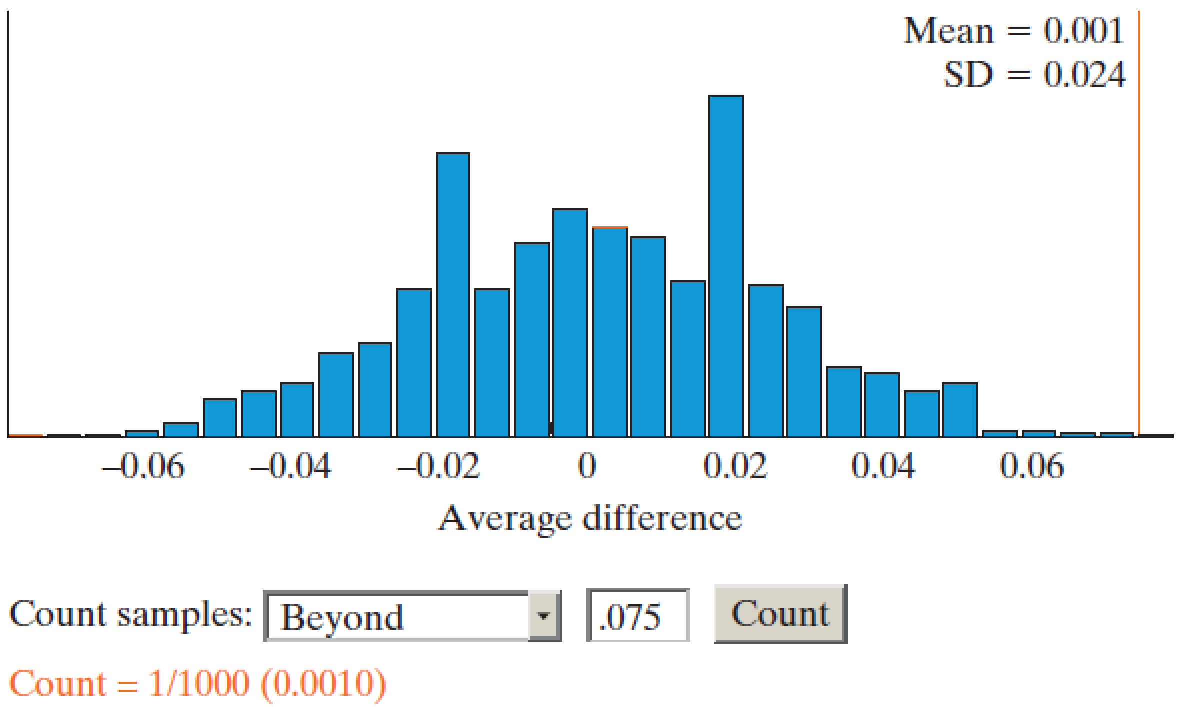
First Base

- Only 1 of the 1000 repetitions of random swappings gave a $\bar{x}_d$ value at least as extreme as



First Base

- We can also standardize 0.075 by dividing by the SE of 0.024 to see our standardized statistic = $\frac{0.075}{\,} = 3.125$



Rounding First Base

- With a p-value of 0.1%, we have very strong evidence against the null hypothesis. The running path makes a statistically significant difference with the wide-angle path being faster on average.
- We can draw a cause-and-effect conclusion since the researcher used random assignment of the two base running methods for each runner.
- There was not much information about how these 22 runners were selected though so it is unclear if we can generalize to a larger population.

3S Strategy

- **Statistic:** Compute the statistic in the sample. In this case, the statistic we looked at was the observed mean difference in running times.
- **Simulate:** Identify a chance model that reflects the null hypothesis. We tossed a coin for each runner, and if it landed heads we swapped the two running times for that runner. If the coin landed tails, we did not swap the times. We then computed the mean difference for the 22 runners and repeated this process many times.
- **Strength of evidence:** We found that only 1 out of 1000 of our simulated mean differences was at least as extreme as the observed difference of 0.075 seconds.

First Base

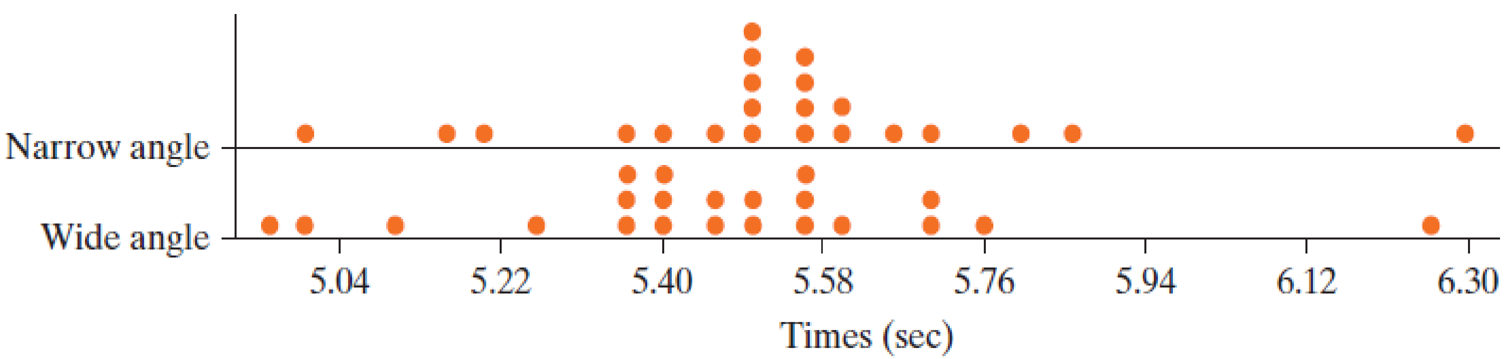
- Approximate a 95% confidence interval for μ_d:
  - $0.075 \pm 1.96(0.024)$ seconds.
  - (0.028, 0.122) seconds.
- What does this mean?
  - We are 95% confident that, if we were to keep testing this indefinitely, the narrow angle route would take somewhere between 0.028 to 0.122 seconds longer on average than the wide angle route.

Since n = 22 here, the sample size is pretty small and the multiplier of 1.96 is not quite correct. If we assume the population of differences is normal, we should use a t multiplier, which here would be 2.08, so the 95% CI would be (.025, .125).

First Base

Alternative Analysis

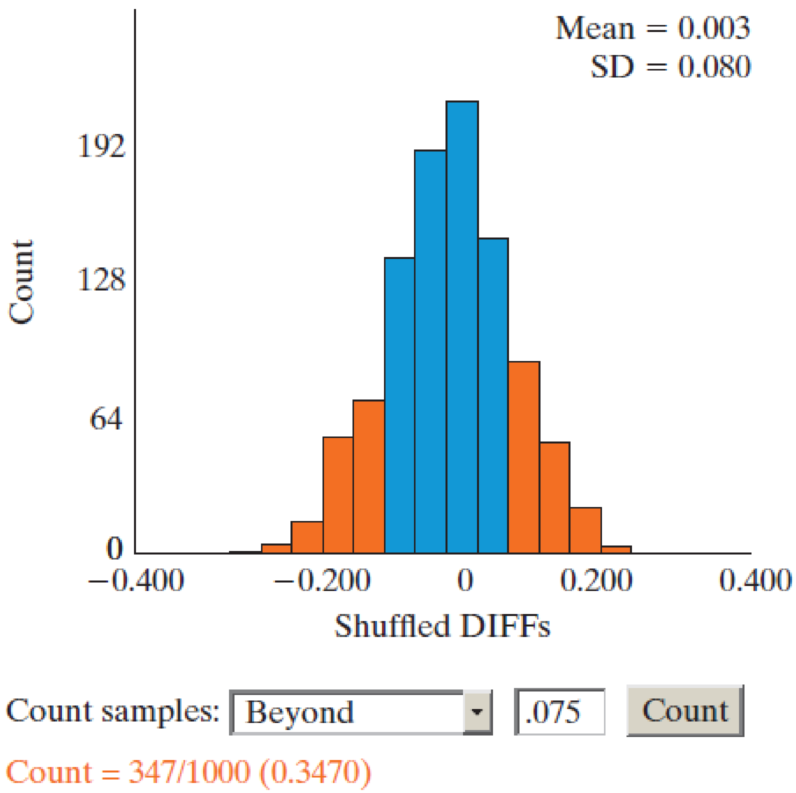
- What do you think would happen if we wrongly analyzed the data using a 2 independent samples procedure? (i.e. The researcher selected 22 runners to use the wide method and an independent sample of 22 other runners to use the narrow method, obtaining the same 44 times as in the actual study.



First Base

Ignoring the fact that it is paired data, we get a p-value of 0.3470.

Does it make sense that this p-value is larger than the one we obtained earlier?



2. 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:

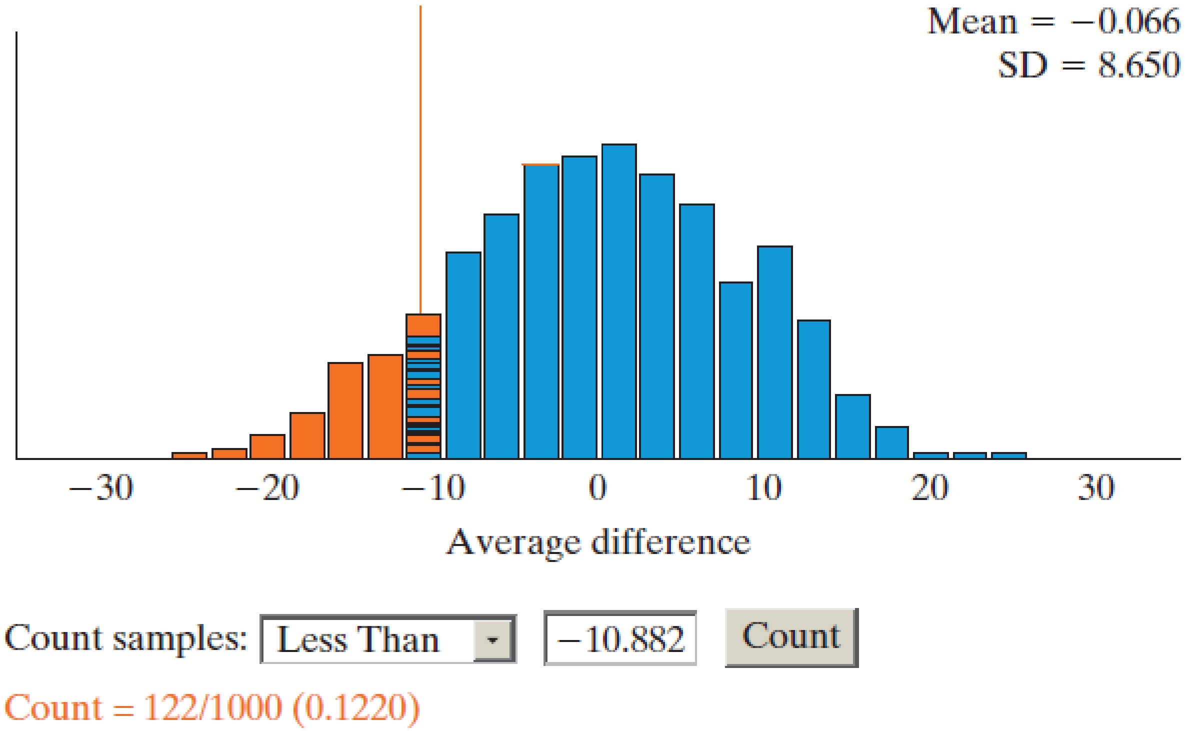
- H_0: $\mu_d = 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is 0.
- H_a: $\mu_d < 0$
  - The long-run mean difference in number of M&Ms taken (small – large) is less than 0.

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	$\bar{x}_d = -10.88$	$s_d = 36.30$

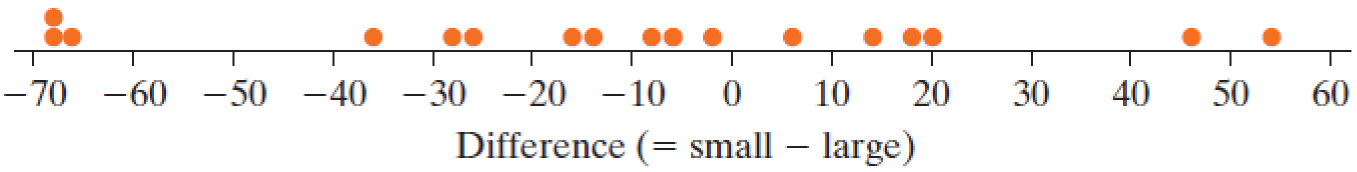
How Many M&Ms Would You Like?

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



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: One mean

☐ Paste data

n: 17

mean, $\bar{x}$: −10.88

sample sd, s: 36.3

Calculate

☑ Confidence interval

confidence level 95 % Calculate CI

(−29.5435, 7.7835)

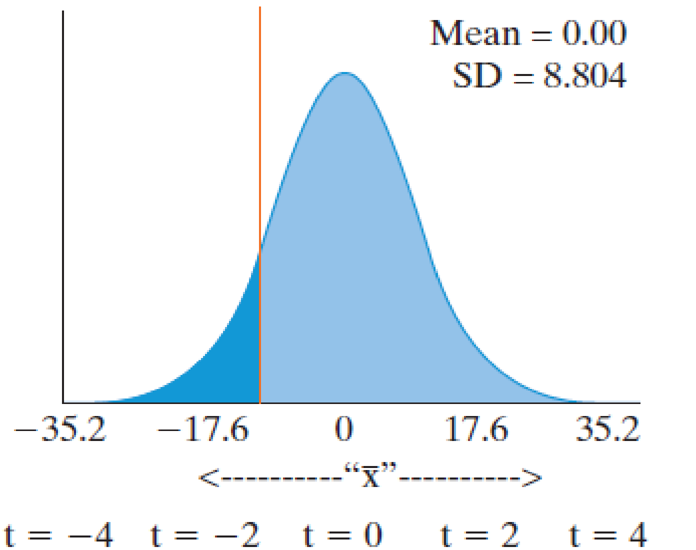
Theory-based inference

☑ Test of significance

$H_0: \mu = 0$

$H_a: \mu < 0$

Calculate



Standardized statistic $t = -1.24$ df = 16

p-value 0.1172

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.

3. 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.

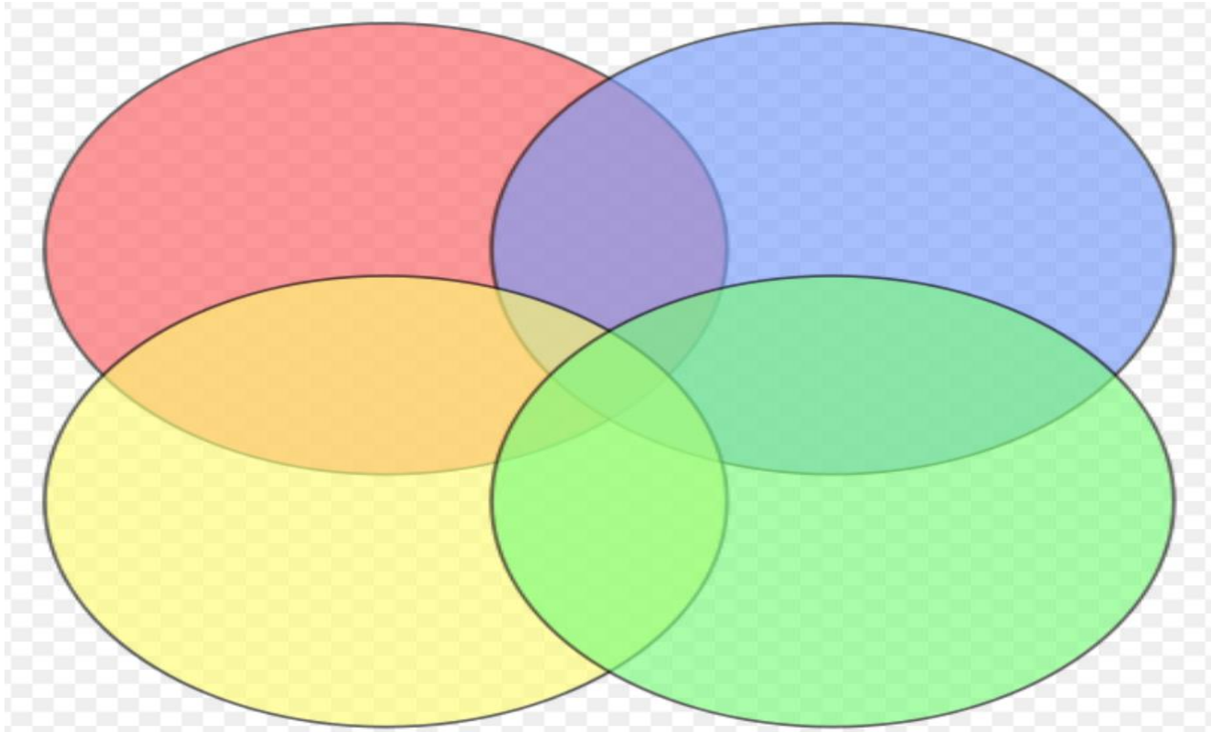
Multiple testing and publication bias.

* 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(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(at least 1 Type I error) will be $\le$ m (5%/m) = 5%.



P(Type I error on explanatory 1) = 5%/m.

P(Type I error on explanatory 2) = 5%/m.

P(Type 1 error on at least one explanatory) ≤

P(error on 1) + P(error on 2) + ... + P(error on m) = m x 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.

4. Two quantitative variables and correlation.

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

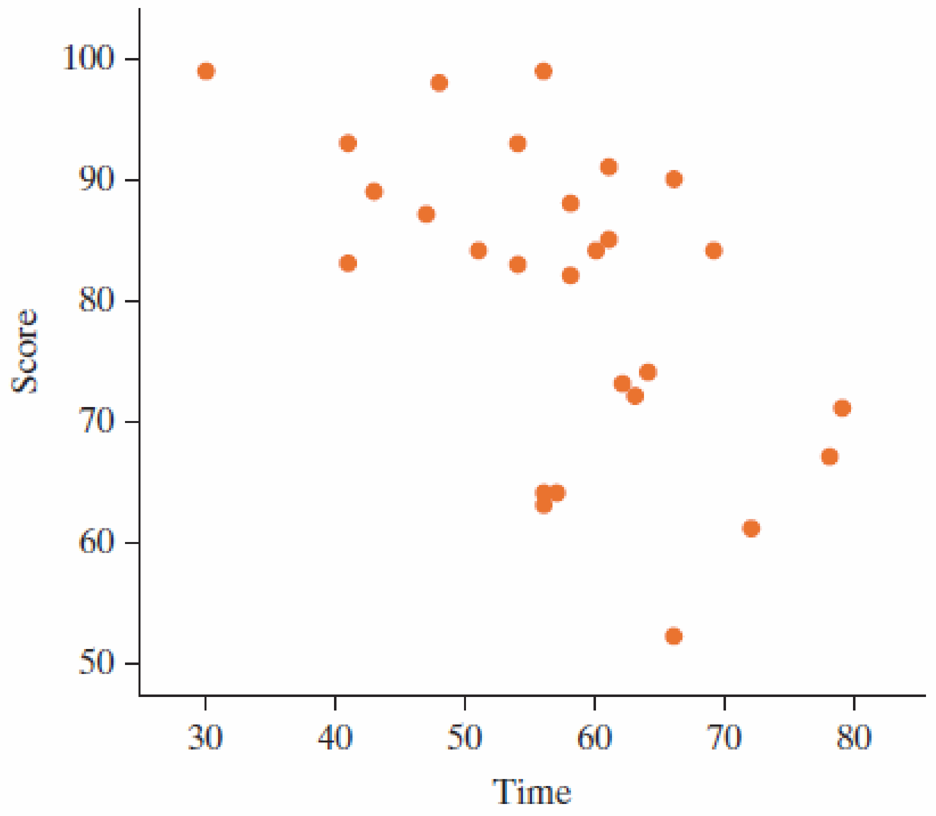
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

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.

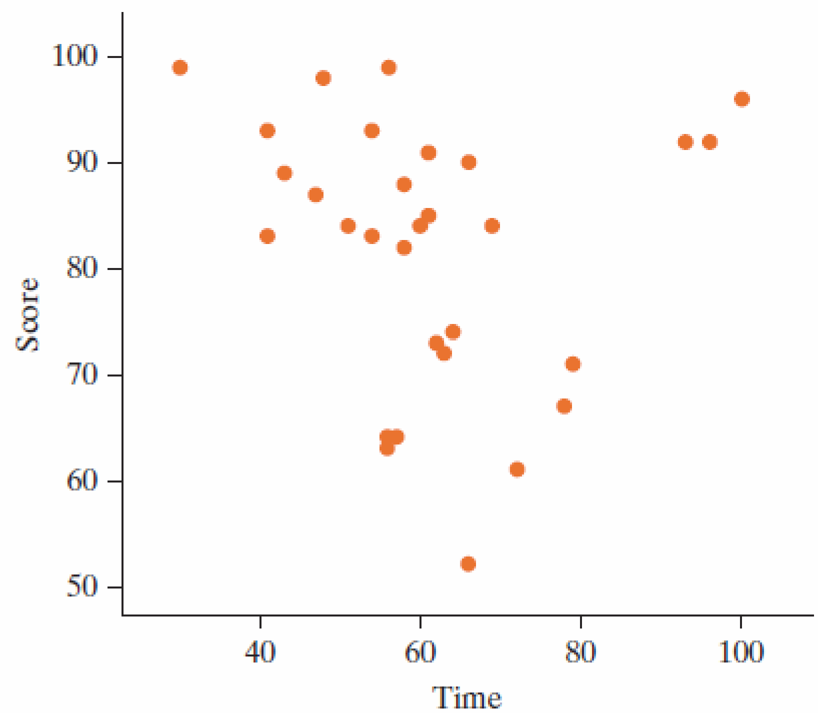
Correlation Guidelines

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

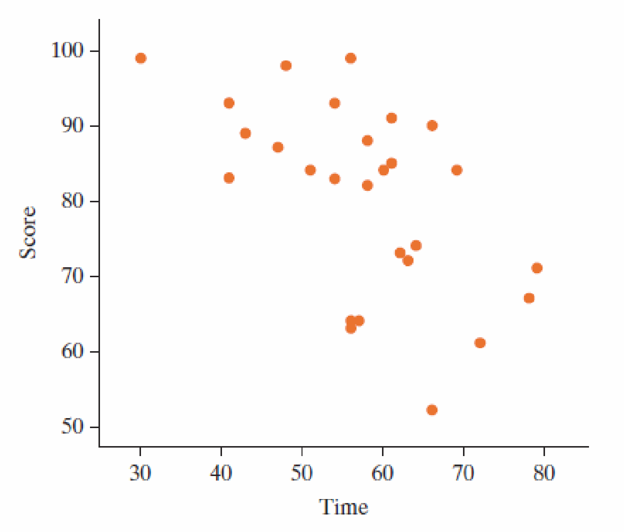
Actually the last three people to finish the test had scores of 93, 93, and 97.

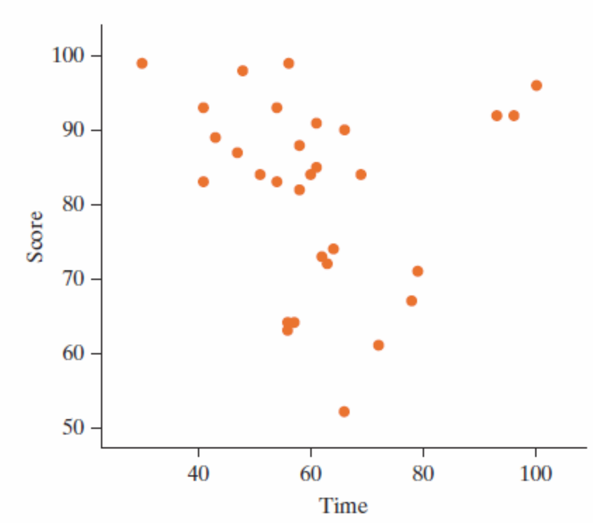
What does this do to the correlation?



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)





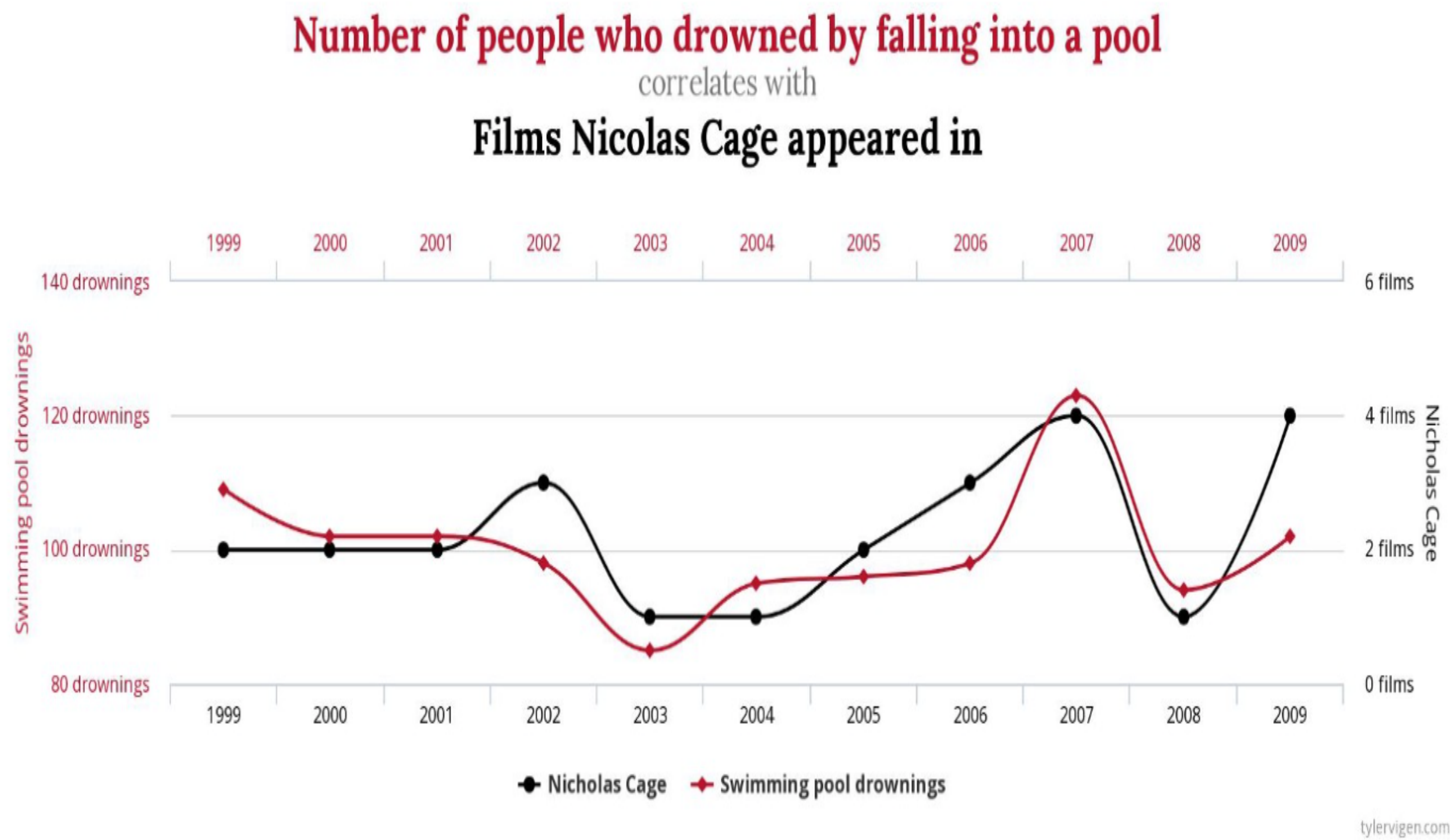
Correlation

- **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 ≠ causation.



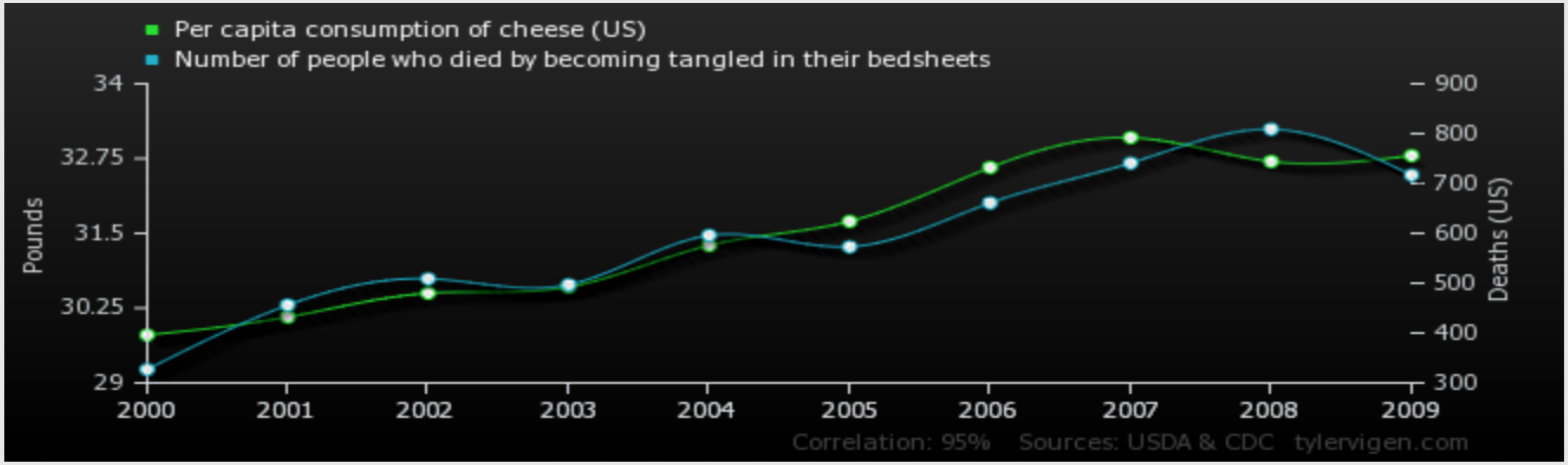
from: http://tylervigen.com

Note that correlation ≠ 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

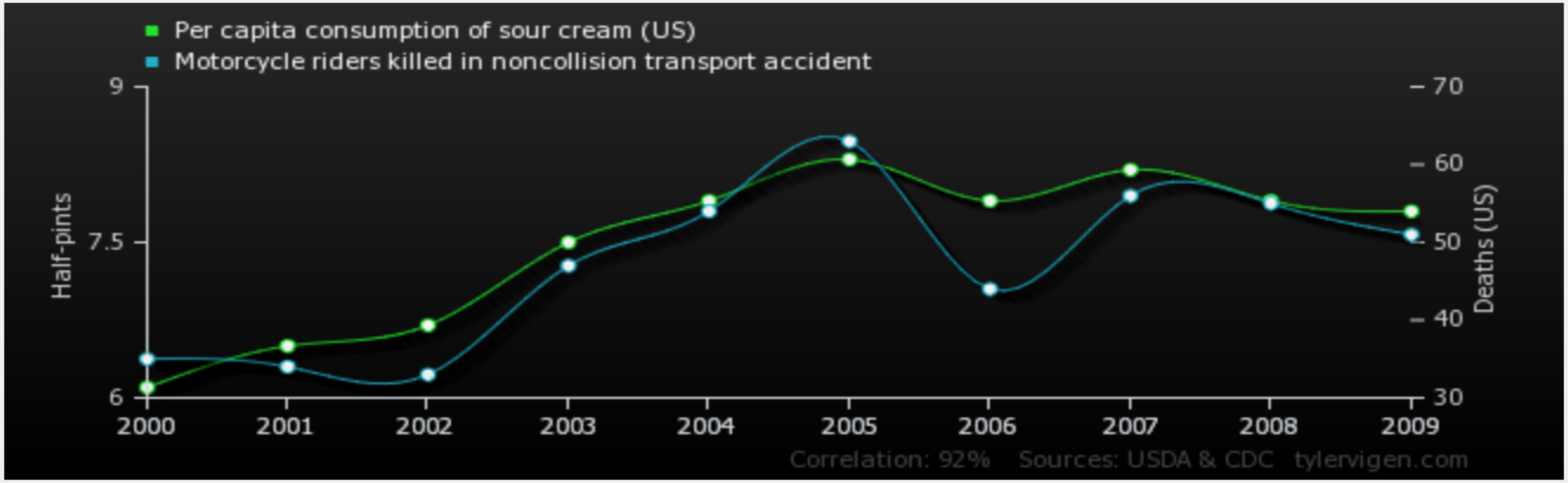
Correlation: 0.947091

Note that correlation ≠ 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