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

http://www.stat.ucla.edu/~frederic/13/W23 .

0. Remember, **no lecture Fri Mar10.** Extra-credit.
1. Multiple testing and publication bias.
2. Two variables and correlation.

Read ch7 and 10.
Hw4 is due Fri Mar10 at 2pm by email to statgrader or statgrader2.
10.1.8, 10.3.14, 10.3.21, and 10.4.11.
http://www.stat.ucla.edu/~frederic/13/W23 .

0. Remember, **no lecture Fri Mar10.** Extra-credit.

If you can find a website or research article with detailed spatial-temporal information on each patient in some region with some contagious disease, and if you are the only student in the class identifying this particular website or article, then you get 5% bonus on your overall course grade.
It must have a specific location and time for each patient.
Not just a total number of patients on each day in each city.
If you find one but other students find the same one too, you get a 2% bonus.
If you find an article with a plot of the spatial or spatial-temporal points, where each subject is a point, but the exact coordinates are not published, you get a 3% bonus.
It must be a contagious disease, not cholera or cancer.
If you find one, email me at frederic@stat.UCLA.edu by Mar20. I will not accept emails after Mar20. I will give no partial credit for trying unsuccessfully.

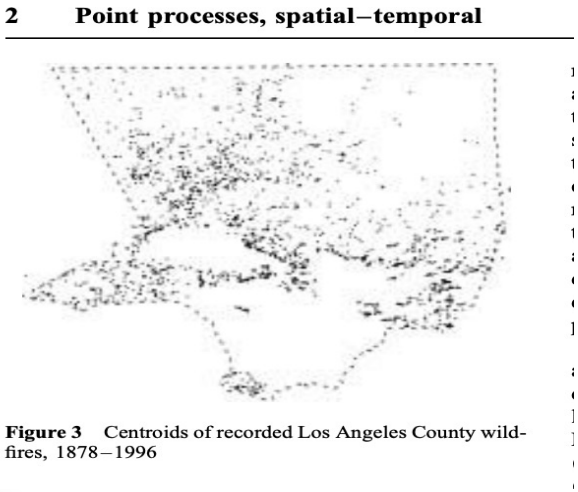


Figure 3 Centroids of recorded Los Angeles County wildfires, 1878–1996

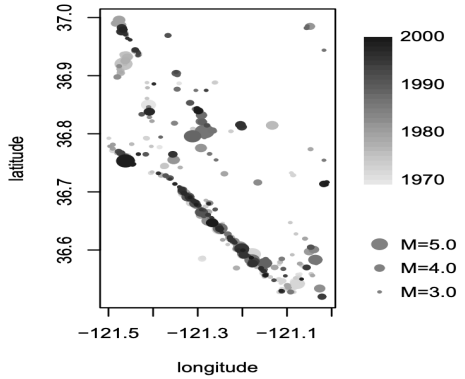


Figure 1: Locations, times and magnitudes of moderate-sized ($M \geq 3.5$) earthquakes in Bear Valley, CA, between 1970 and 2000.

Table 1. Shallow Shocks (M ≥ 6.0) in OFF Tohoku Area for 1885–1980

NO	YEAR	MO	DY	HR	MN	MAG	C	NO	YEAR	MO	DY	HR	MN	MAG	C	NO	YEAR	MO	DY	HR	MN	MAG	C
1	1885	2	9	2	0	6.0	0	84	1908	1	15	21	56	6.9	0	167	1923	5	31	14	55	6.2	1
2	1885	6	11	9	20	6.9	0	85	1908	1	18	1	5	6.0	0	168	1923	6	2	2	24	7.3	0
3	1885	7	29	5	30	6.0	0	86	1908	2	5	21	7	6.0	0	169	1923	6	2	5	14	7.1	2
4	1885	10	30	20	30	6.2	0	87	1908	6	27	23	21	6.1	0	170	1923	6	7	2	36	6.2	2
5	1885	12	7	13	2	6.3	0	88	1908	11	22	16	15	6.4	0	171	1923	9	2	18	49	6.3	2
6	1885	12	19	18	26	6.0	2	89	1909	9	17	4	39	6.8	0	172	1923	11	18	5	40	6.3	1
7	1886	4	13	5	44	6.3	0	90	1910	1	22	8	25	6.0	0	173	1923	12	27	23	39	6.4	0
8	1886	7	2	12	33	6.3	2	91	1910	5	9	18	53	6.0	1	174	1924	2	3	7	25	6.3	0
9	1887	5	29	0	50	6.4	0	92	1910	5	10	22	56	6.1	0	175	1924	5	31	21	2	6.3	1
10	1887	5	29	1	10	6.2	2	93	1910	5	12	12	22	6.0	2	176	1924	5	31	21	4	6.4	1
11	1888	2	5	0	50	7.1	0	94	1910	10	13	23	56	6.3	0	177	1924	8	6	23	22	6.3	0
12	1888	11	24	2	3	6.5	0	95	1912	1	4	4	4	6.1	0	178	1924	8	15	3	2	7.1	0
13	1889	3	31	6	42	6.6	0	96	1912	1	9	6	21	6.1	0	179	1924	8	15	8	27	6.7	2
14	1890	11	17	9	31	6.3	0	97	1912	6	8	13	41	6.6	0	180	1924	8	17	10	45	6.3	2
15	1891	4	7	9	49	6.7	0	98	1912	12	9	8	50	6.6	0	181	1924	8	17	11	10	6.6	2
16	1891	5	5	8	16	6.2	0	99	1913	2	20	17	58	6.9	0	182	1924	8	25	23	31	6.7	2
17	1891	7	21	20	19	7.0	0	100	1913	5	22	5	36	6.1	1	183	1925	2	7	2	11	6.0	0
18	1892	10	22	19	9	6.0	0	101	1913	5	29	19	14	6.4	0	184	1925	4	20	5	24	6.3	0
19	1894	2	25	4	18	6.8	0	102	1913	10	3	9	17	6.1	1	185	1925	6	2	14	18	6.4	0
20	1894	3	14	18	15	6.0	2	103	1913	10	11	18	10	6.9	0	186	1925	11	10	23	44	6.0	0
21	1894	8	29	19	55	6.6	0	104	1913	10	13	2	5	6.6	2	187	1926	4	7	4	33	6.3	0
22	1894	11	28	1	5	7.1	0	105	1914	2	7	15	50	6.8	0	188	1926	5	27	4	45	6.4	0
23	1894	12	1	18	37	6.3	0	106	1914	12	26	3	18	6.1	0	189	1926	9	5	0	37	6.8	0
24	1896	1	9	22	17	7.5	0	107	1915	3	9	0	29	6.8	0	190	1926	10	3	17	25	6.4	0
25	1896	1	10	5	52	6.0	2	108	1915	4	6	5	25	6.0	1	191	1926	10	19	9	29	6.2	0
26	1896	1	10	11	25	6.3	0	109	1915	4	6	14	32	6.2	0	192	1926	11	11	12	1	6.1	2
27	1896	2	23	19	42	6.1	2	110	1915	4	25	2	9	6.4	0	193	1927	1	18	6	58	6.4	0
28	1896	3	6	23	52	6.0	2	111	1915	5	28	2	26	6.0	2	194	1927	3	16	15	52	6.4	2
29	1896	4	11	23	0	6.0	2	112	1915	6	5	6	59	6.7	0	195	1927	7	30	23	18	6.4	0
30	1896	6	15	19	32	8.5	0	113	1915	7	9	7	21	6.4	0	196	1927	8	6	6	12	6.7	0
31	1896	6	16	4	16	7.5	2	114	1915	10	13	6	30	6.8	0	197	1927	9	30	16	38	6.3	0
32	1896	6	16	8	1	7.5	2	115	1915	10	14	4	43	6.2	2	198	1928	5	27	18	50	7.0	0
33	1896	7	29	17	44	6.1	2	116	1915	10	15	1	28	6.1	2	199	1928	5	29	0	35	6.7	2
34	1896	8	1	11	49	6.5	0	117	1915	10	15	3	40	6.3	2	200	1928	6	1	22	12	6.5	2
35	1896	9	5	23	7	6.5	2	118	1915	10	16	1	55	6.0	2	201	1928	6	2	7	6	6.0	2
36	1897	2	20	5	50	7.4	0	119	1915	10	17	0	21	6.1	2	202	1928	8	1	4	28	6.1	2
37	1897	2	20	8	47	7.0	2	120	1915	11	1	16	24	7.5	0	203	1929	3	15	10	57	6.0	2
38	1897	3	27	19	49	6.3	2	121	1915	11	1	16	50	6.7	2	204	1929	4	1	5	17	6.3	0
39	1897	5	23	21	22	6.9	2	122	1915	11	1	18	1	7.0	2	205	1929	4	16	9	53	6.3	0
40	1897	7	22	18	31	6.8	0	123	1915	11	2	0	43	6.2	2	206	1929	5	31	9	10	6.1	0
41	1897	7	29	22	45	6.0	2	124	1915	11	4	12	13	6.4	2	207	1929	6	27	1	49	6.1	0
42	1897	8	5	9	10	7.7	0	125	1915	11	18	13	4	7.0	2	208	1929	8	29	3	51	6.3	0

1. 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.

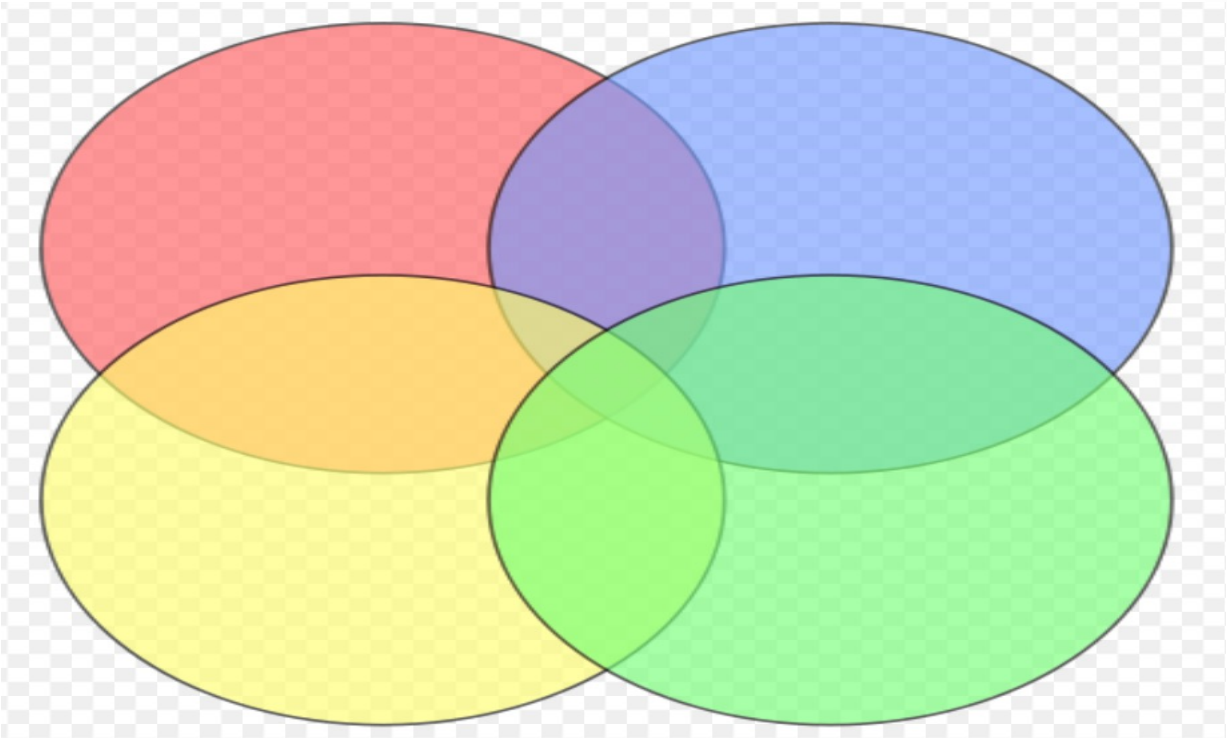
Multiple testing and publication bias.

* 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 ≤ 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.

2. Two quantitative variables.

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

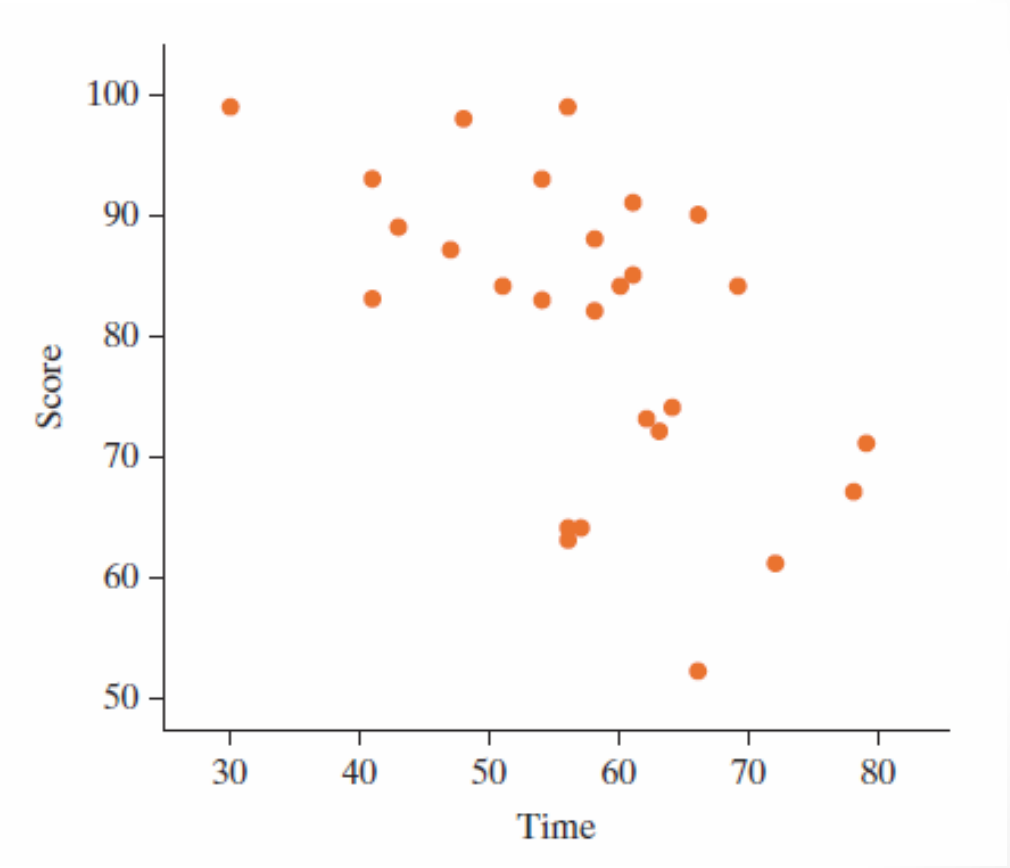
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

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.

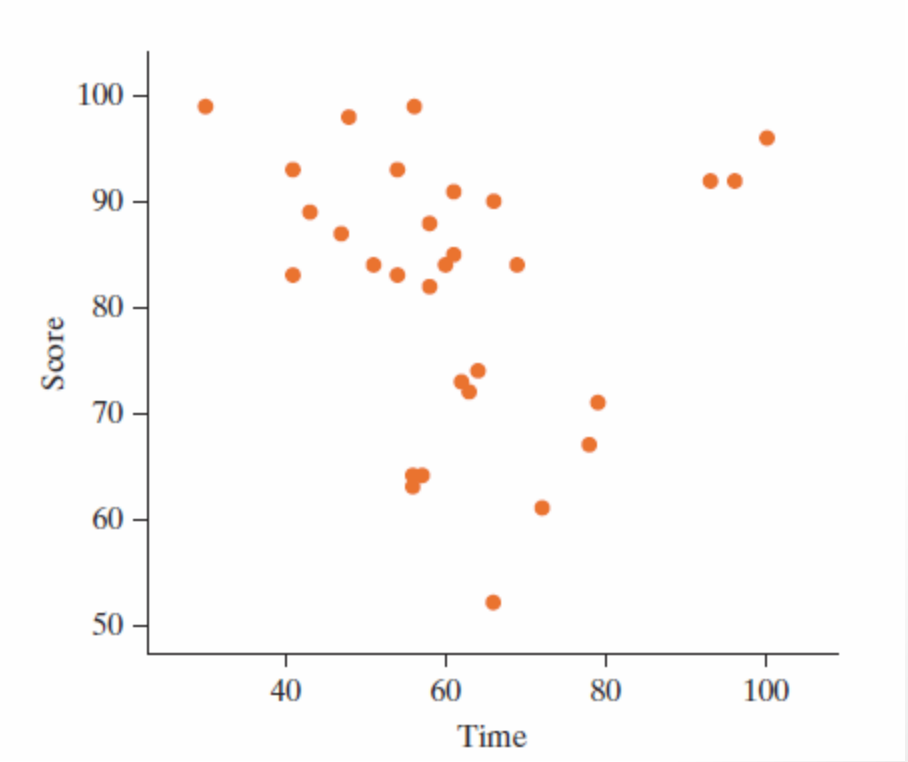
Correlation Guidelines

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

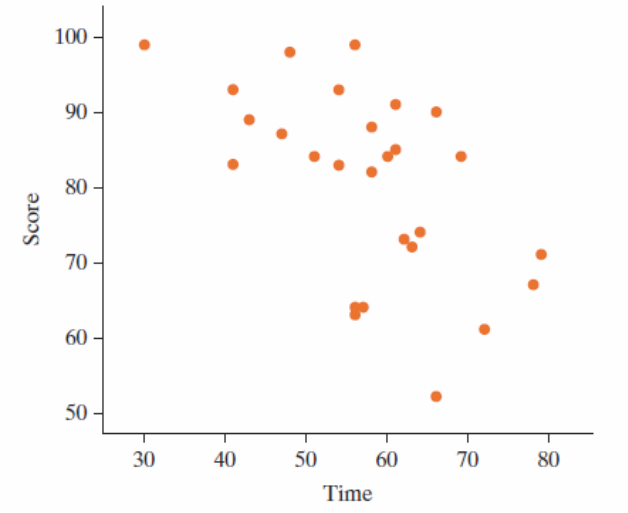
Actually the last three people to finish the test had scores of 93, 93, and 97.

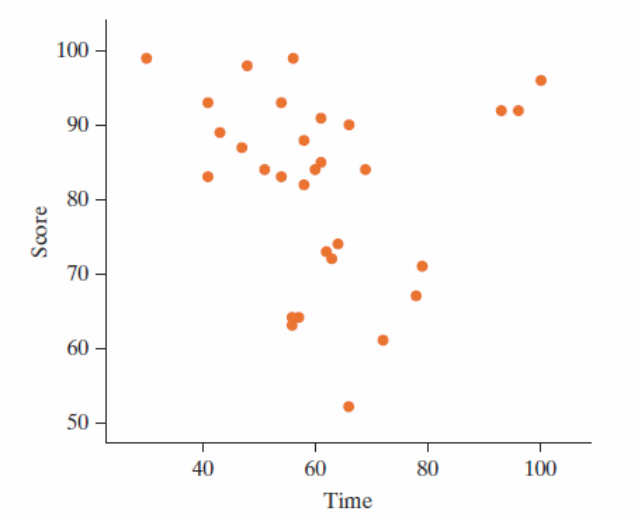
What does this do to the correlation?



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)





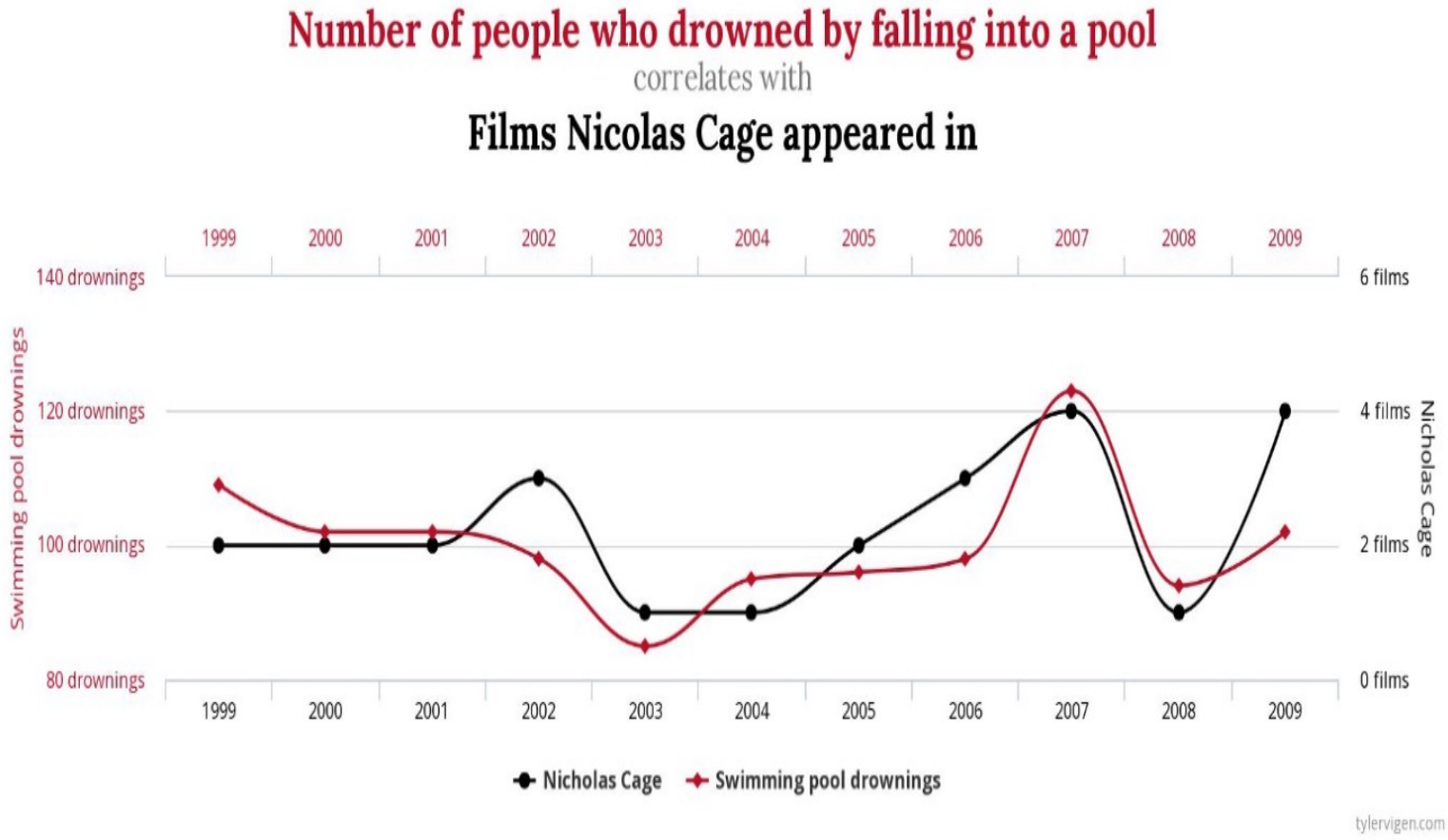
Correlation

- **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 ≠ causation.



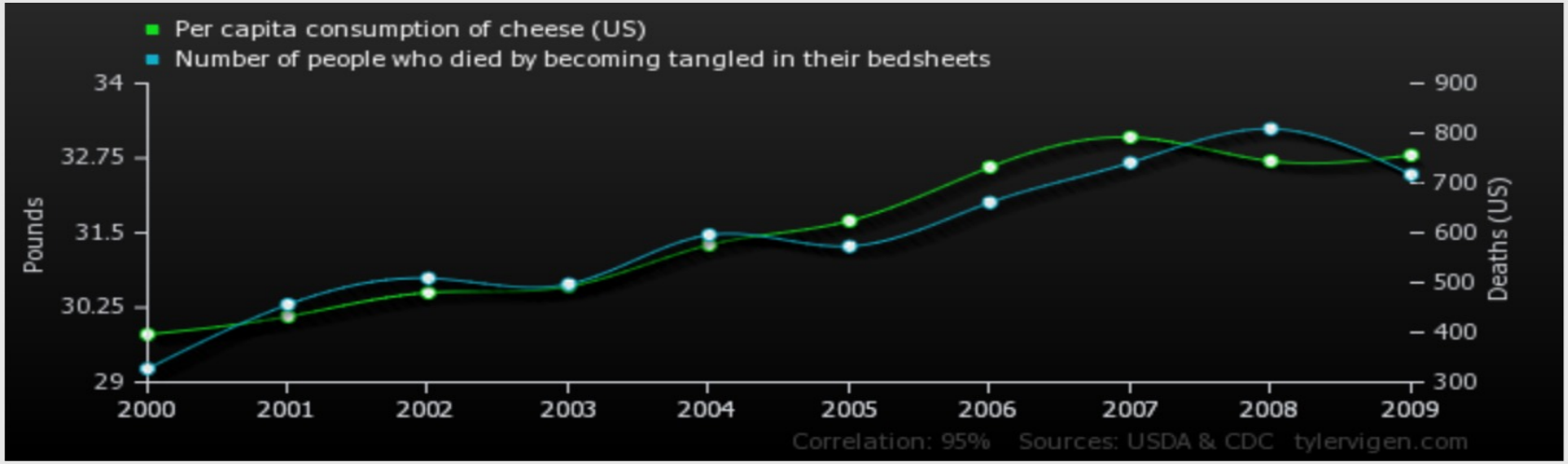
from: http://tylervigen.com

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

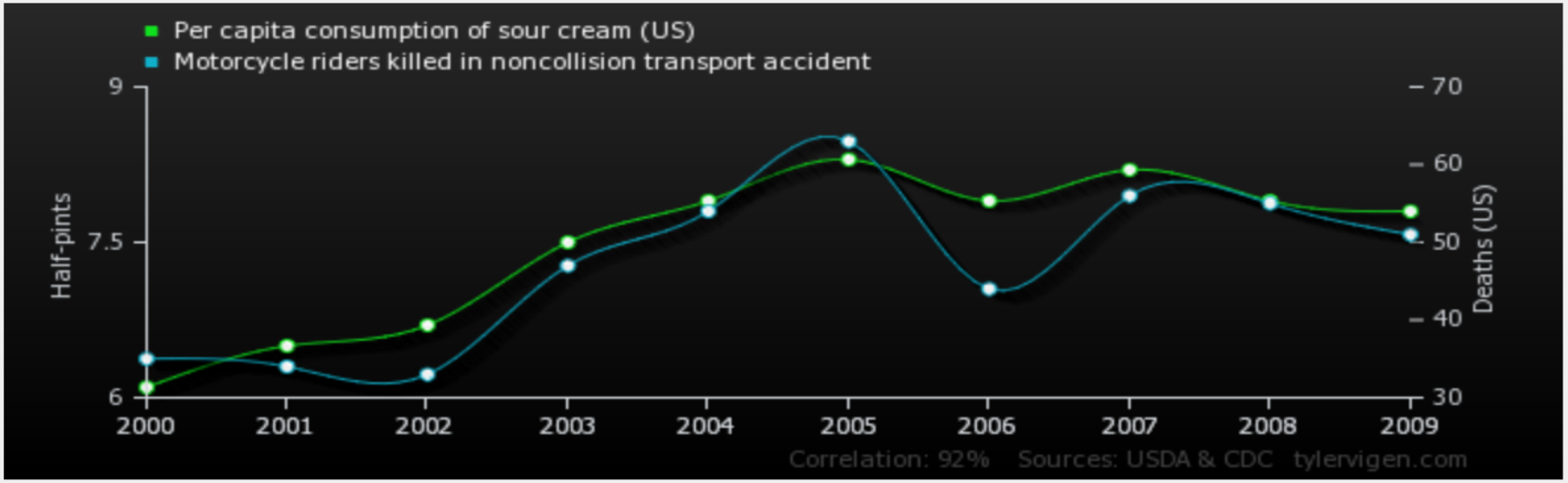
Correlation: 0.947091

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

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. ($\rho > 0$)

ρ = 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

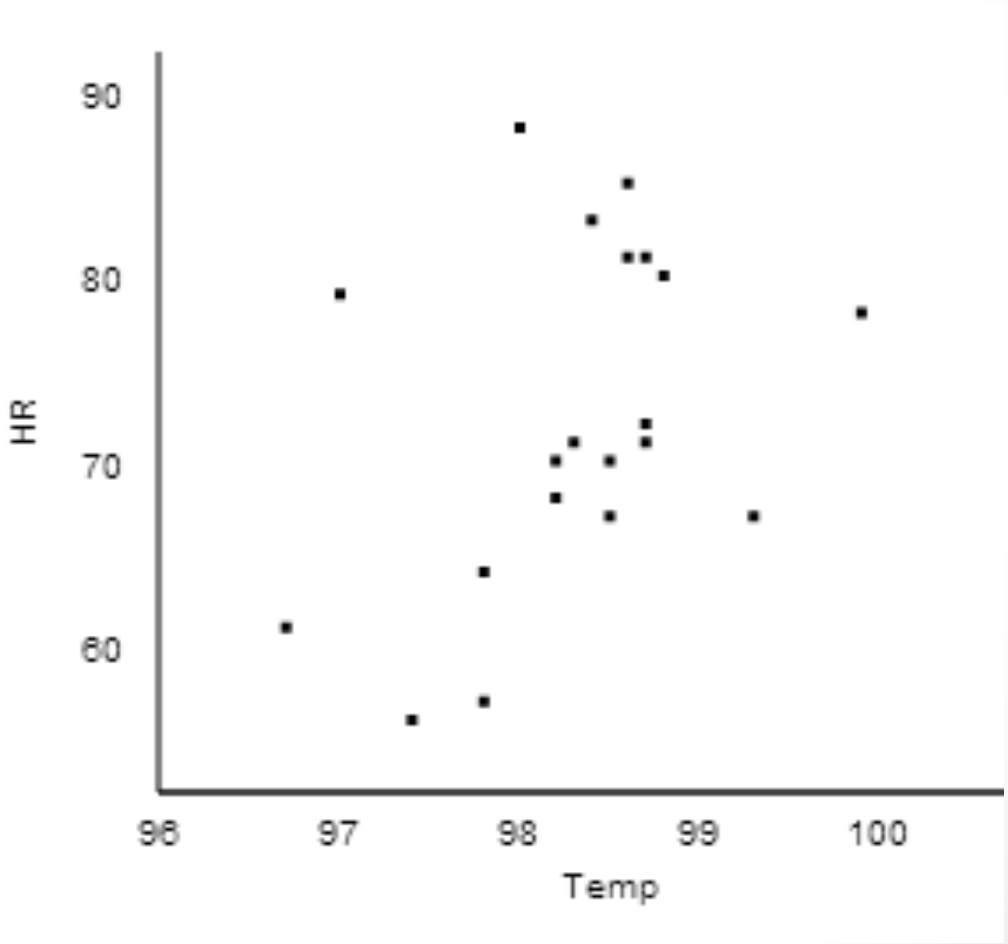
Collect the Data

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

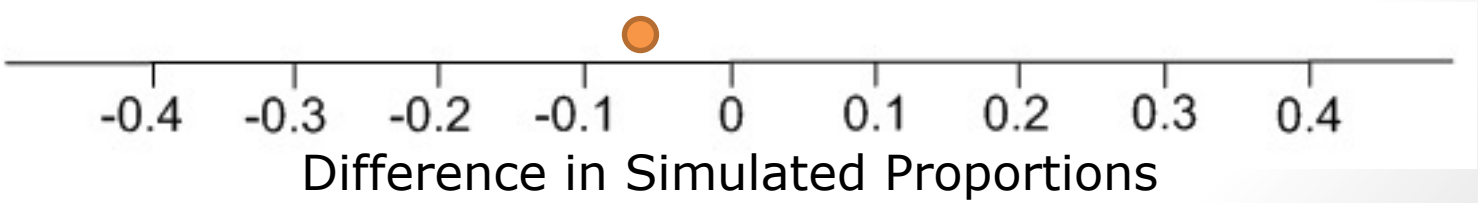
Control

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

40.0% Improvers

0.400 – 0.467 = -0.067

26.7% Improvers



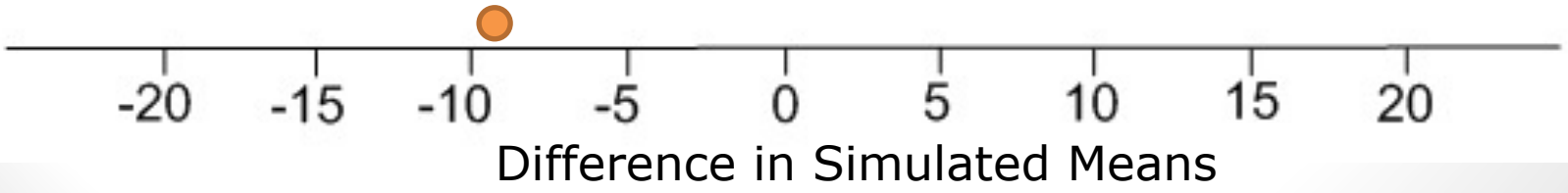
Music No music

25.2	45.6
14.5	11.6
-7.0	18.6
12.6	12.1
34.5	30.5

-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 **mean = 6.38**

6.38 – 16.12 = -9.74



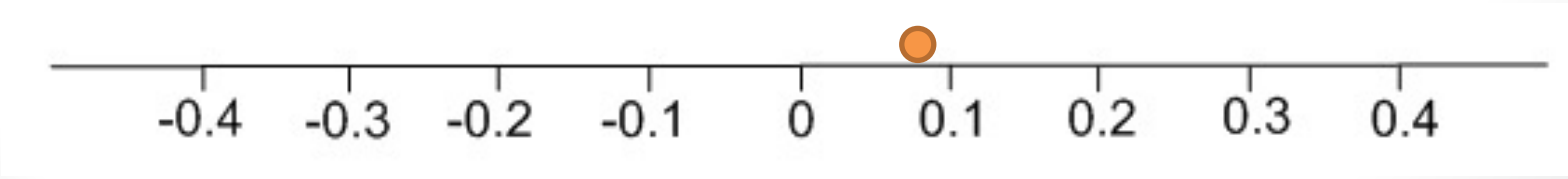
Shuffling Cards

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

r = 0.078

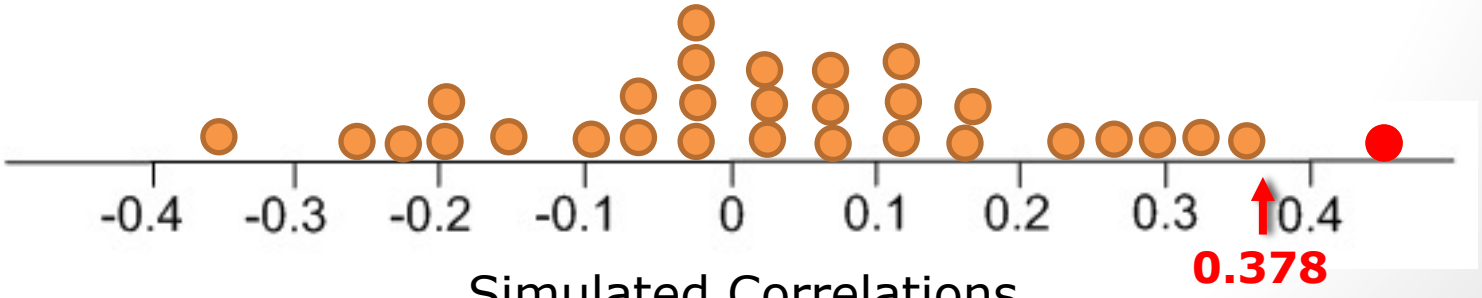


Simulated Correlations

More Simulations

0.097 0.054 0.062 0.259 0.034
-0.253 -0.345 0.339
0.314
0.059 0.447
0.167
0.067
0.232 -0.327 0.100 0.329

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 ≈ 0.03.

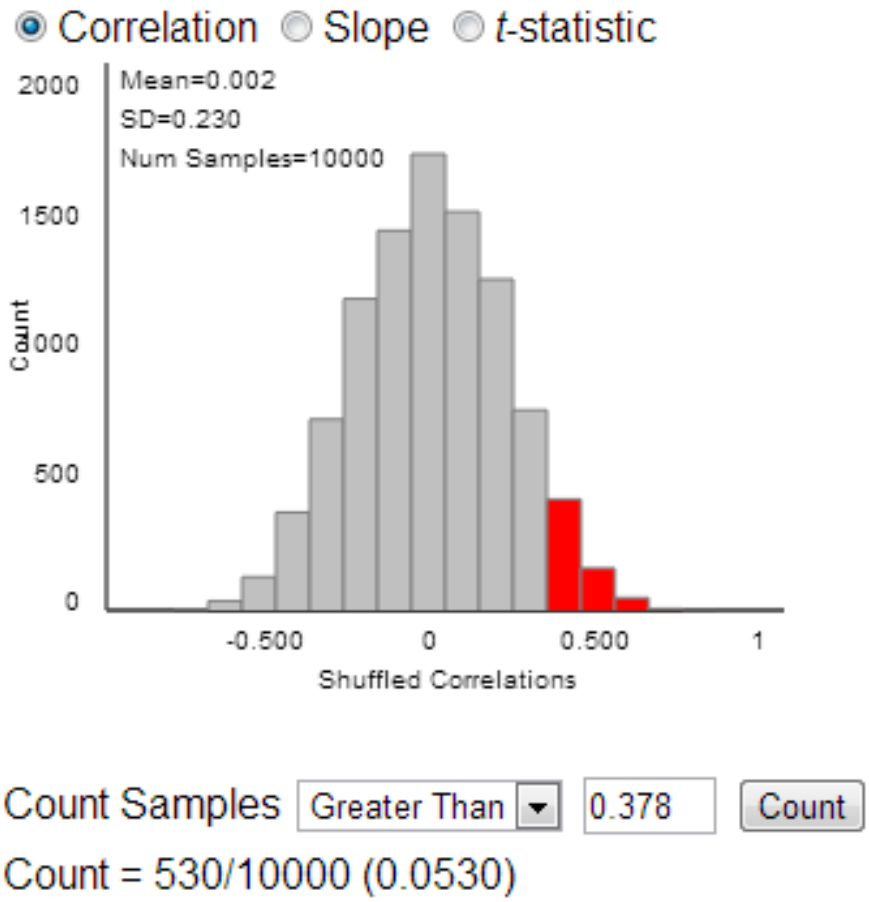


Temperature and Heart Rate

- 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.



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.