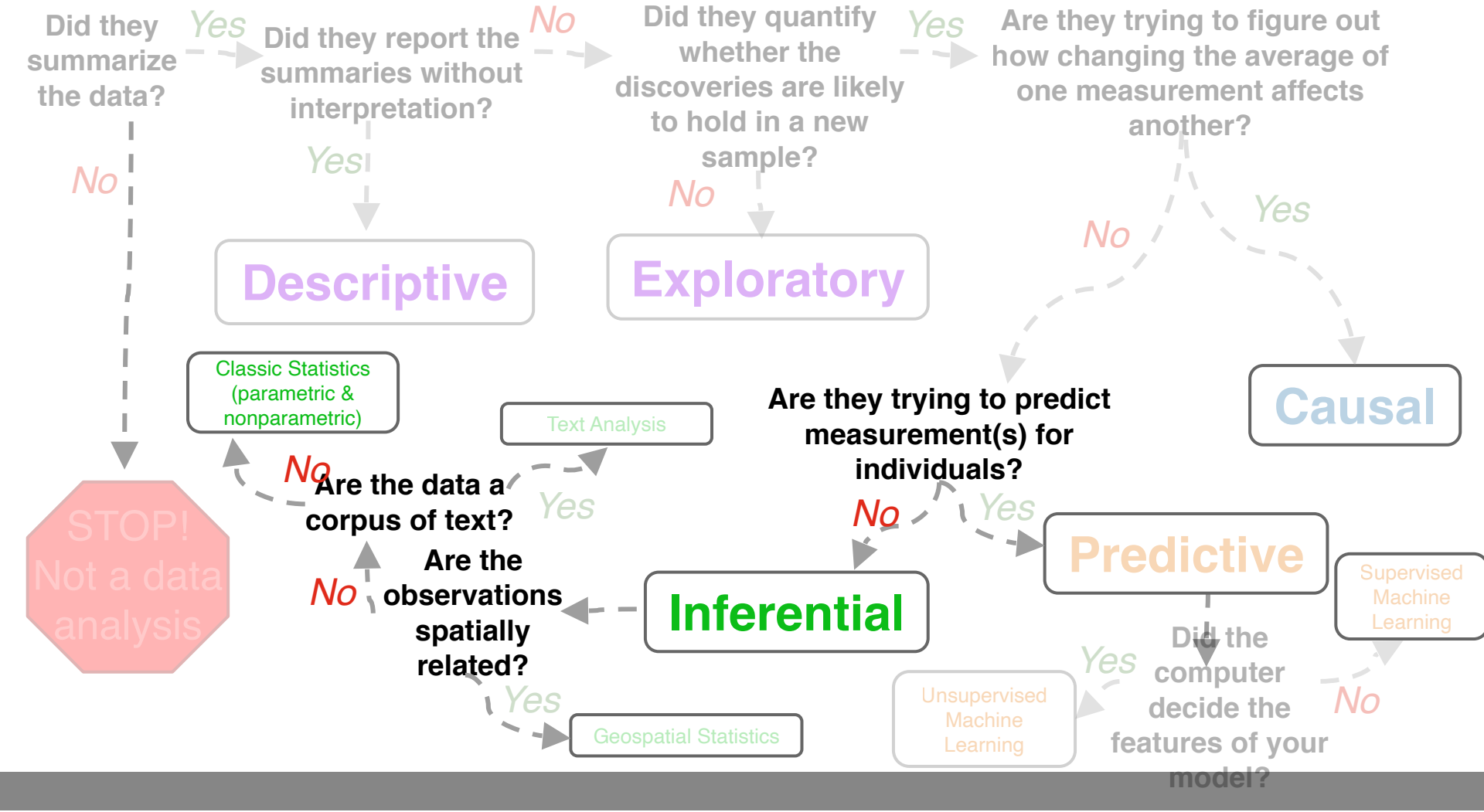
Inferential analysis

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego
jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com


Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
Yes
No
No
Descriptive
Exploratory
Causal
Are they trying to predict measurement(s) for individuals?
No
Yes
Predictive
Supervised Machine Learning
Did the computer decide the features of your model?
Yes
No
Unsupervised Machine Learning
Inferential
Are the observations spatially related?
No
Yes
Geospatial Statistics
Are the data a corpus of text?
No
Yes
Text Analysis
Classic Statistics (parametric & nonparametric)
STOP! Not a data analysis

Population

We want to learn something about this...

Sampling

Inference

Sample

....but we can only *actually* collect data from this

1million comments from 2020

Published in final edited form as:
Epidemiology. 2013 January ; 24(1): 23–31. doi:10.1097/EDE.0b013e3182770237.

The Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 US counties for the period 2000 to 2007

Andrew W. Correia,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

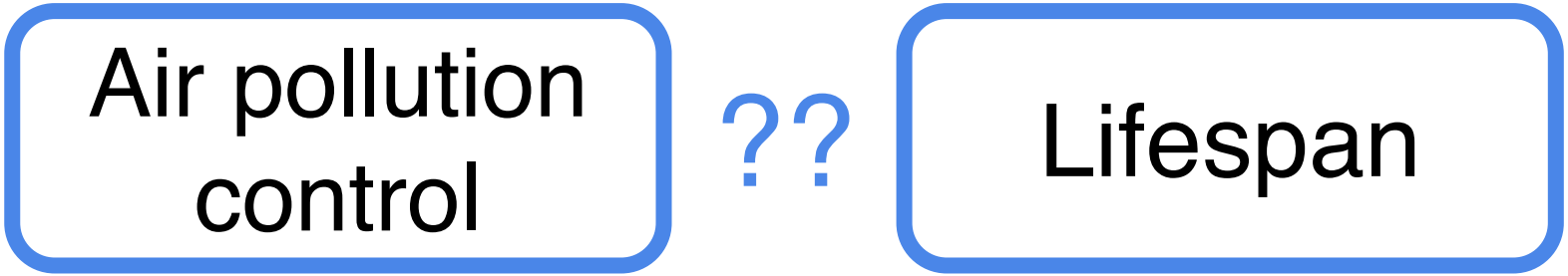
C. Arden Pope III,
Department of Economics, Brigham Young University, 142 Faculty Office Building, Provo, UT 84602

Douglas W. Dockery,
Departments of Environmental Health and Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 1, 1301B, Boston, MA 02115

Yun Wang,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

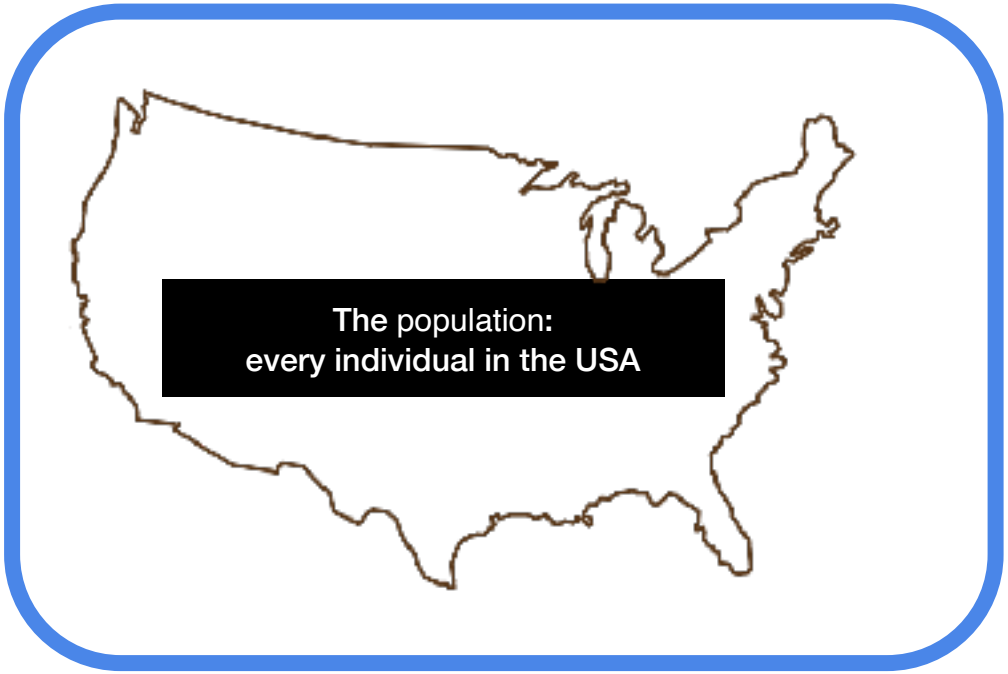
Majid Ezzati, and
MRC-HPA Centre for Environment and Health and Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place, St Mary's Campus, London W2 1PG

Francesca Dominici
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115, fdominic@hsph.harvard.edu, P: (617) 432-1056; F: (617)-739-1781



Is there a relationship between air pollution control and lifespan?

What if we want to know the effect of air pollution on everyone in the United States?



The population:
every individual in the USA

The sample:
545 US counties

Approaches to Inference

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

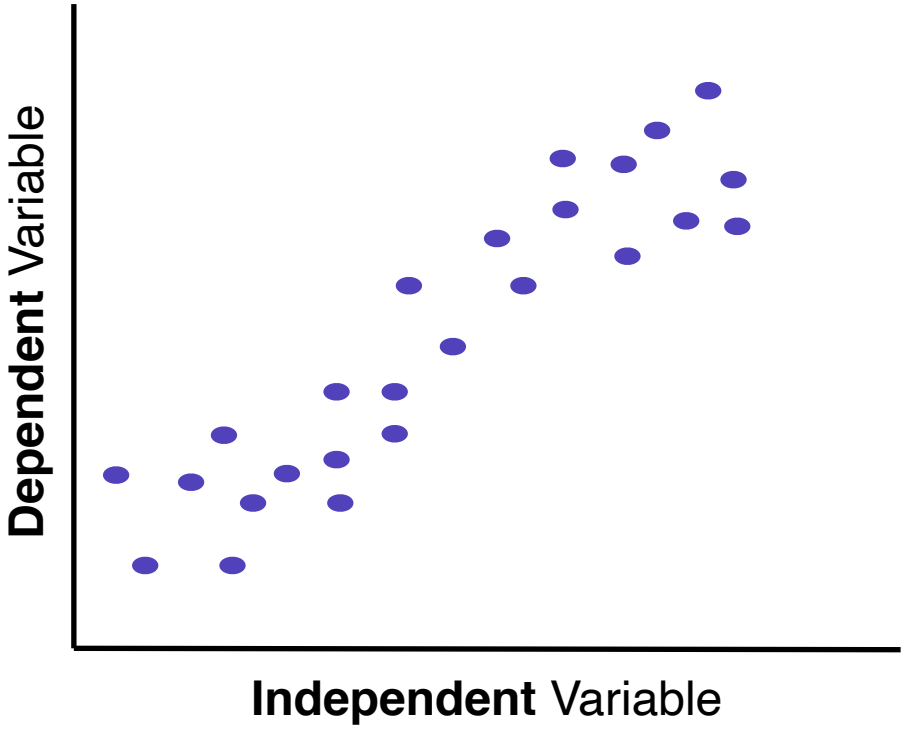
DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

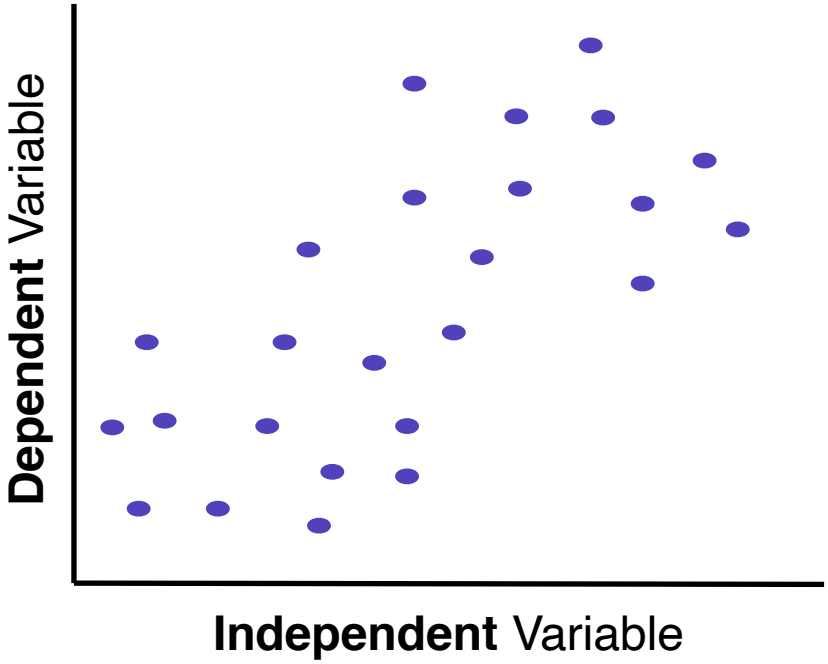
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

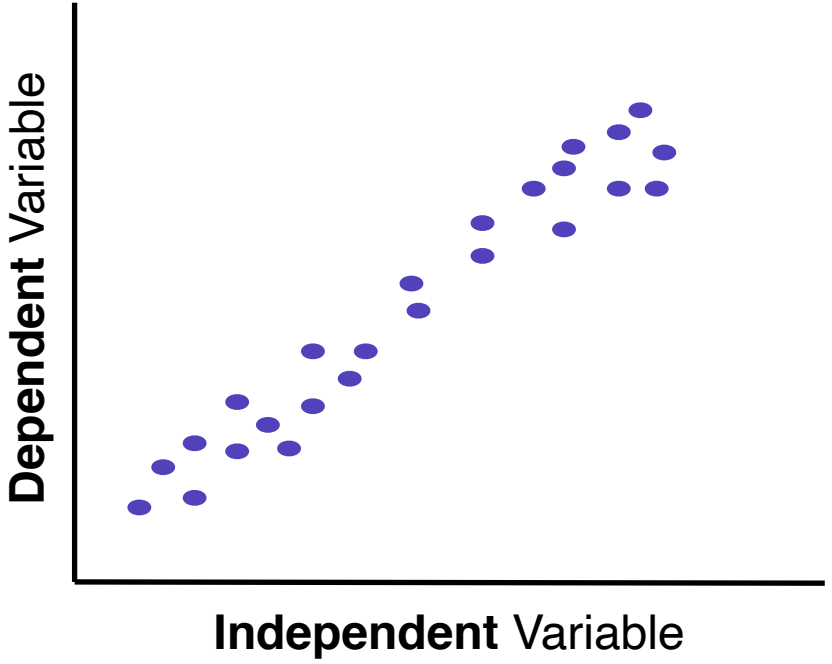
i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test


Dependent Variable
Independent Variable

weaker relationship

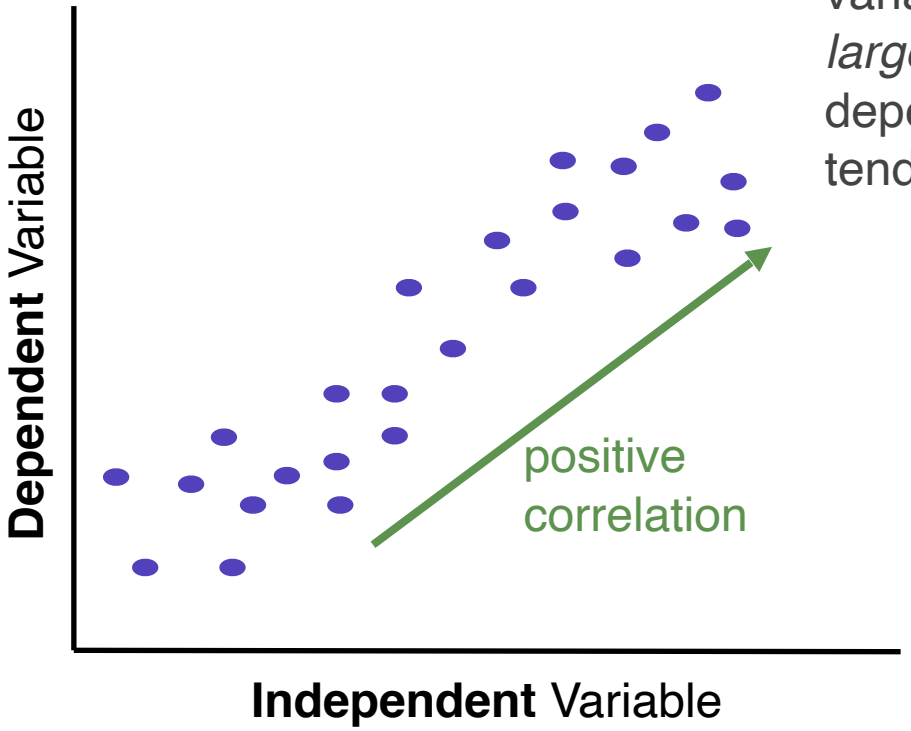


stronger relationship



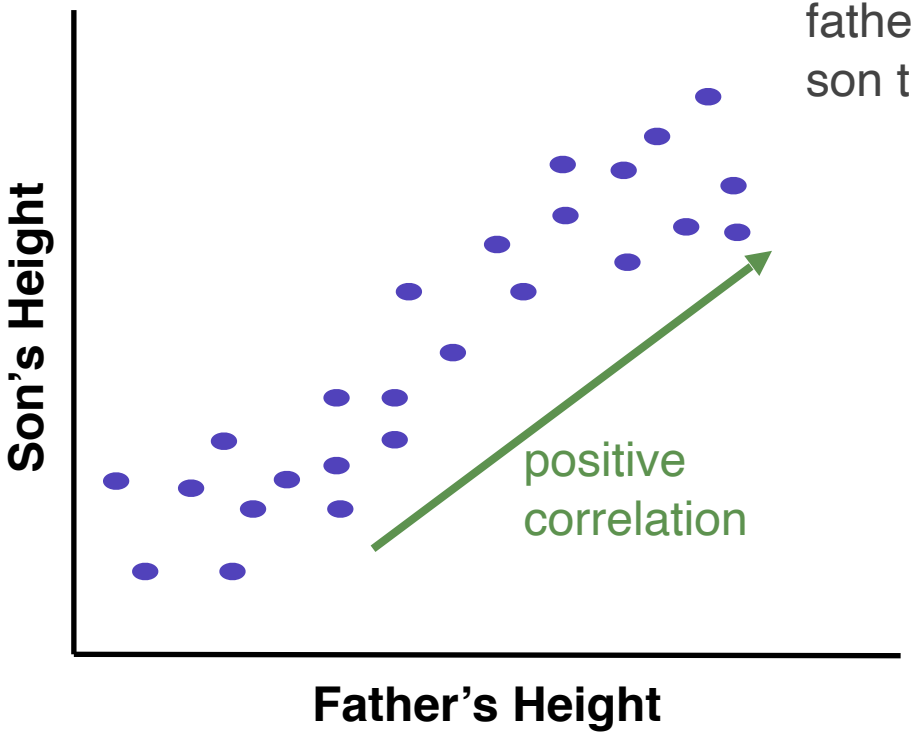
stronger relationship = higher correlation

The *larger* the independent variable value, the *larger* the dependent variable tends to be



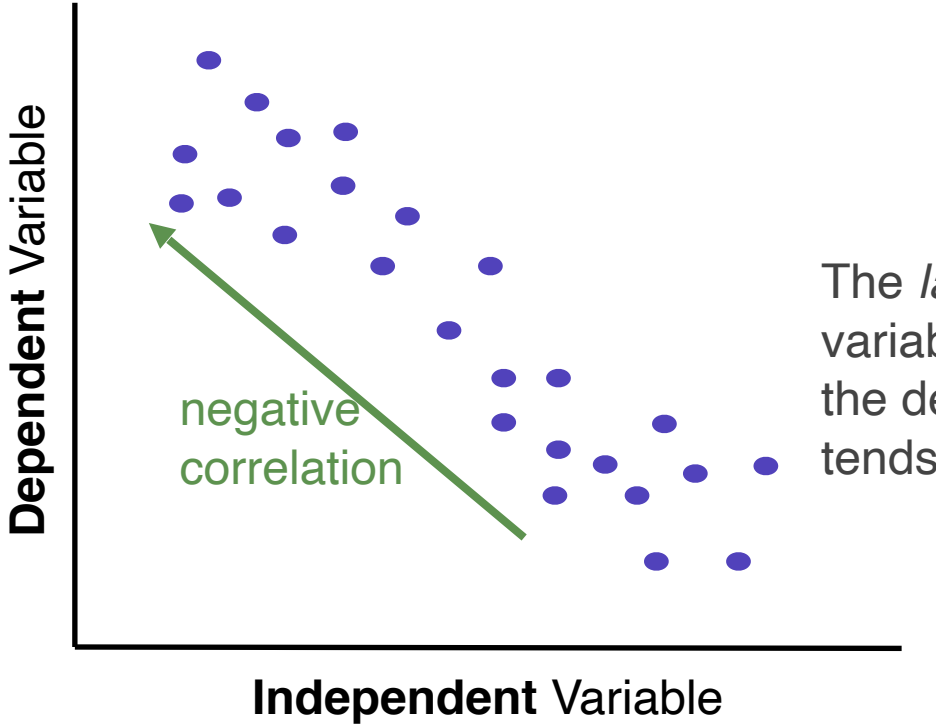
The *smaller* the independent variable value, the *smaller* the dependent variable tends to be

The *taller* the father, the taller his son tends to be



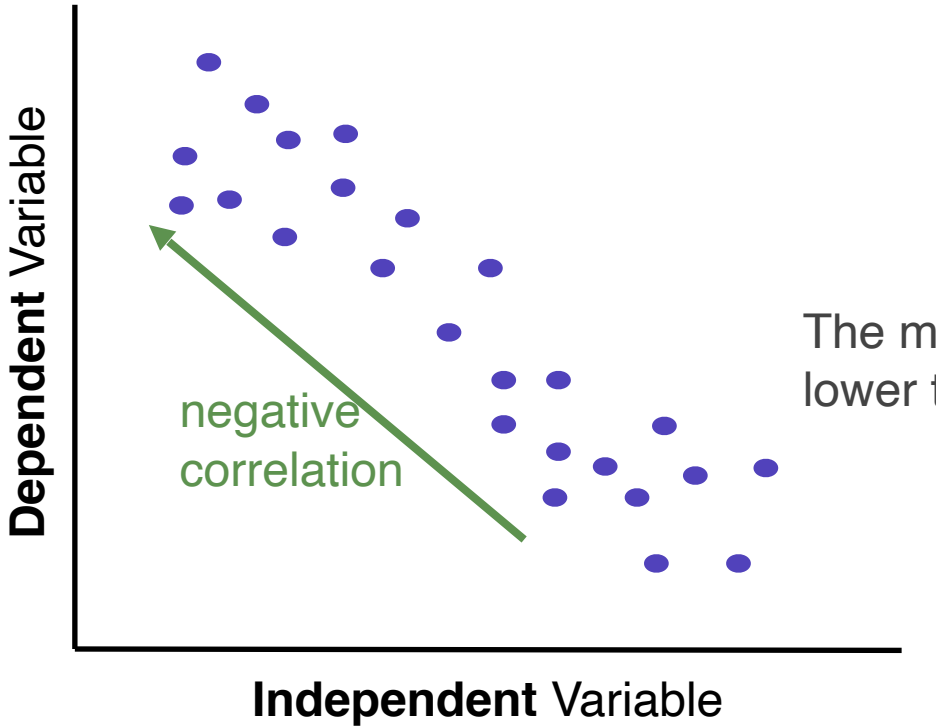
The *shorter* the father, the shorter his son tends to be

The *smaller* the independent variable value, the *larger* the dependent variable tends to be



The *larger* the independent variable value, the *smaller* the dependent variable tends to be

The *lower* the number of absences, the *higher* the students' grades tend to be



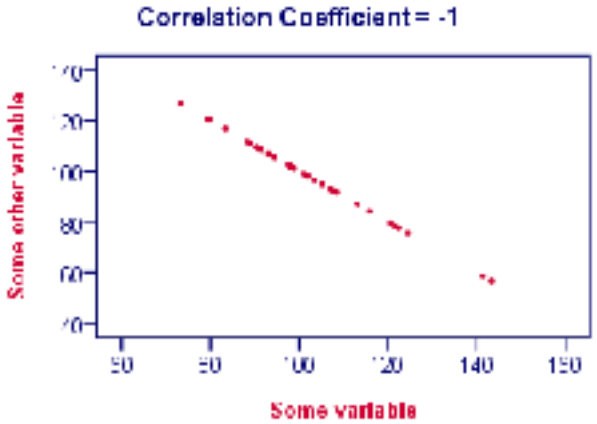
The more absences, the lower the students' grades

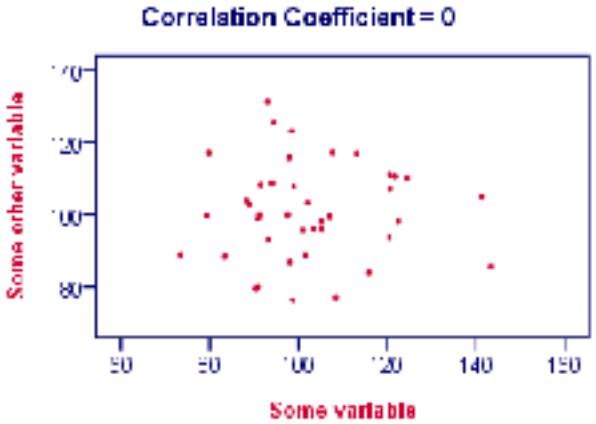
Pearson's *r* :

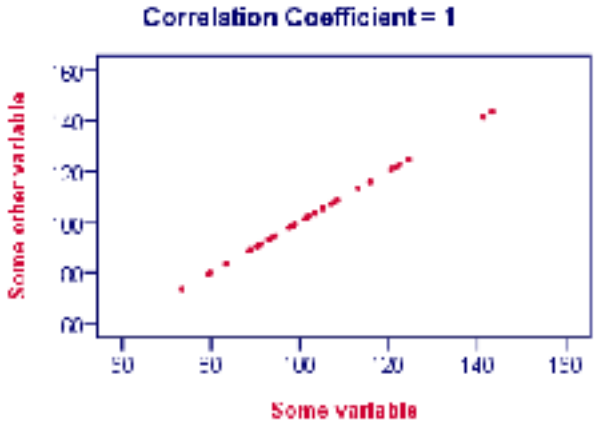
linear correlation between two variables

takes values [-1,1]

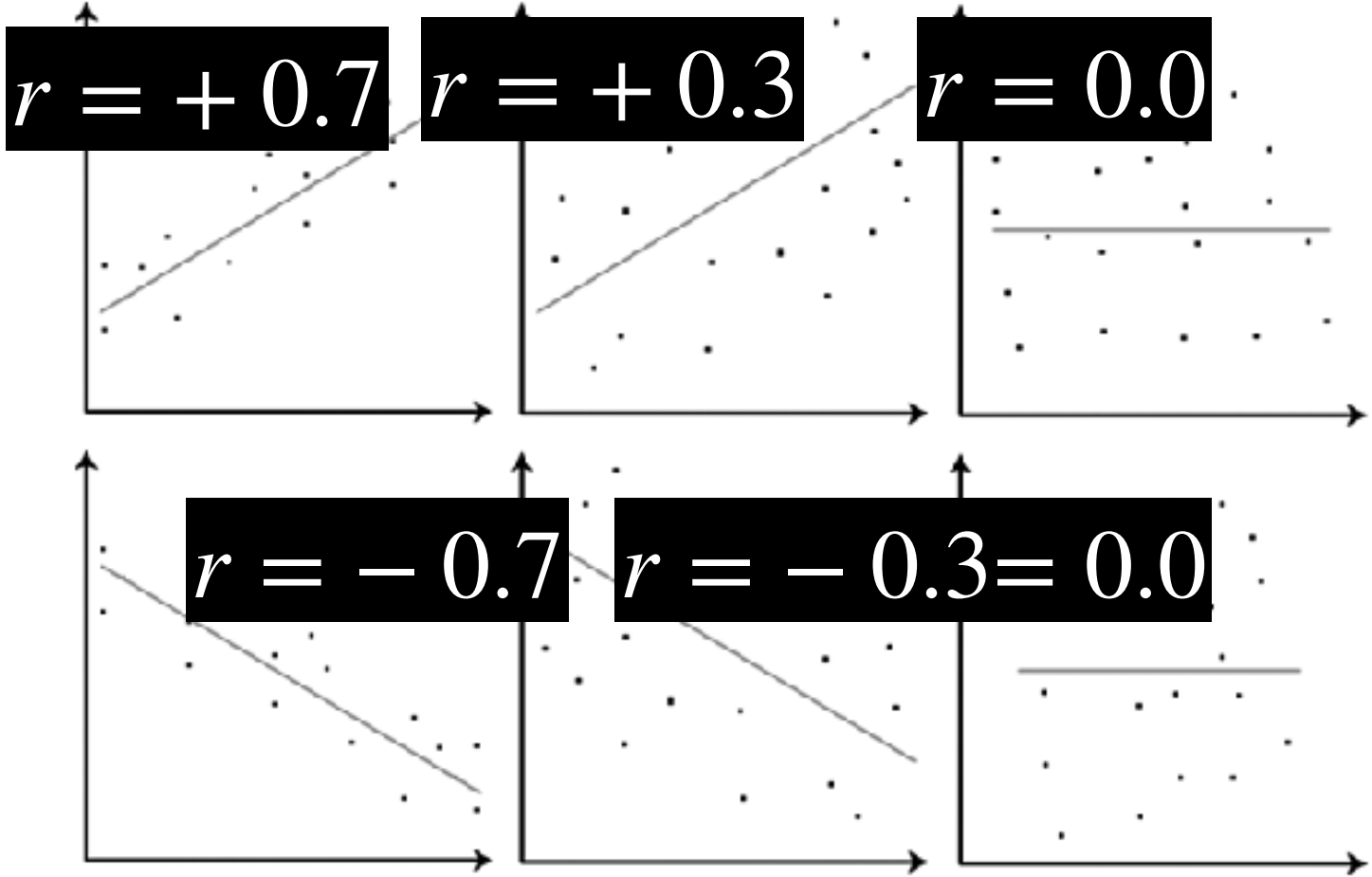
Correlation is how close the data are to being in a line…
BUT IT HAS NOTHING TO DO WITH THE SLOPE





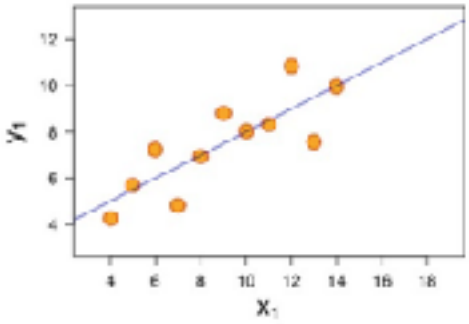


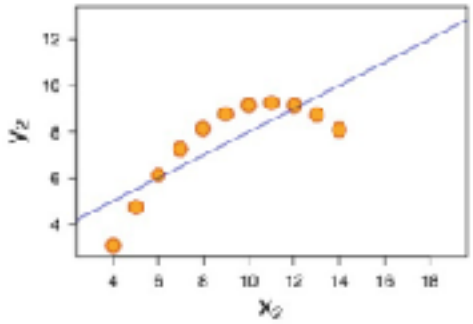
https://www.spss-tutorials.com/pearson-correlation-coefficient/

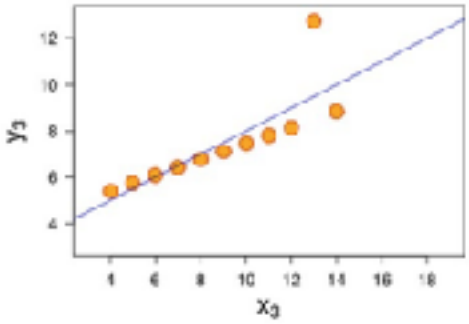

$r = +0.7$
$r = +0.3$
$r = 0.0$
$r = -0.7$
$r = -0.3 = 0.0$

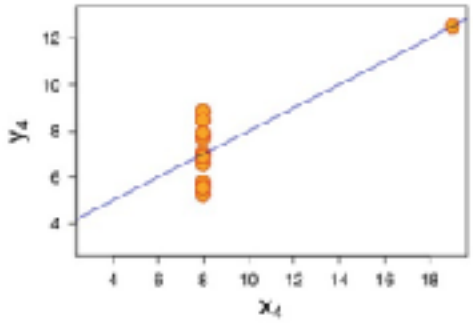
Anscombe's Quartet

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

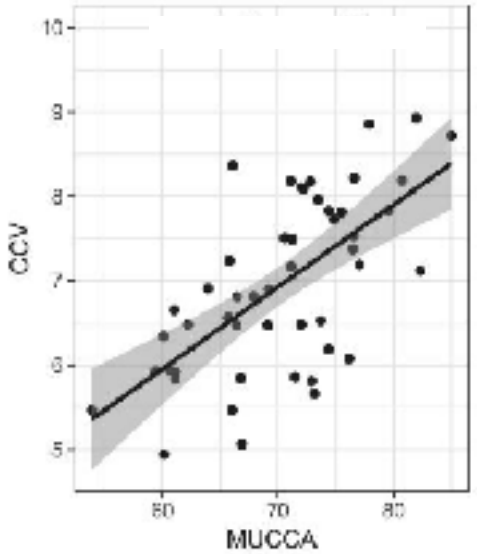




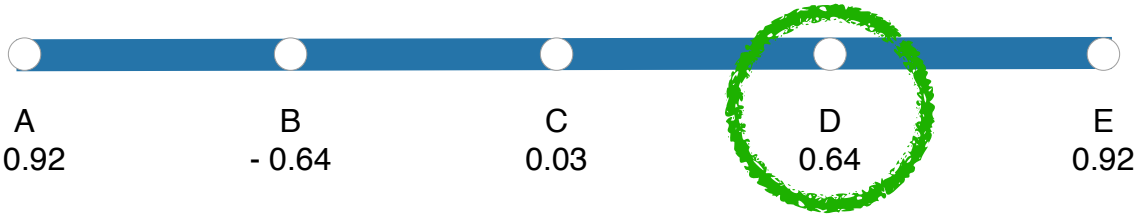




Correlation Champ

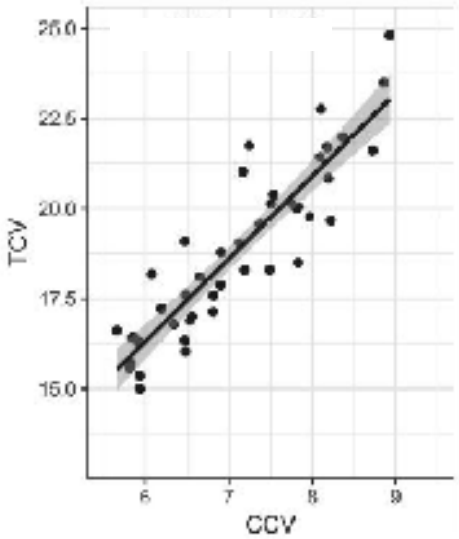


Which of the following is the Pearson correlation coefficient (*r*) for this relationship?

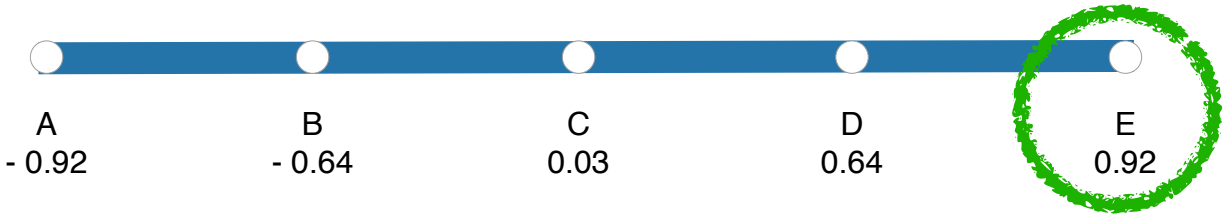


https://twitter.com/theajnr/status/1022082875343204353

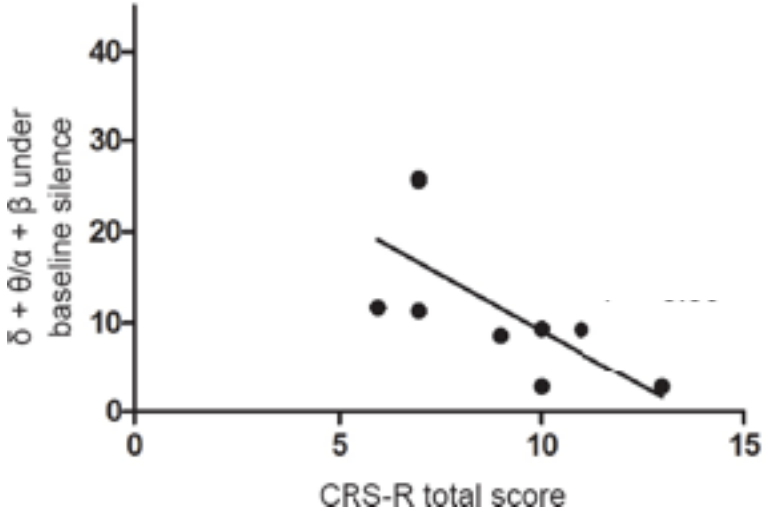
Correlation Champ



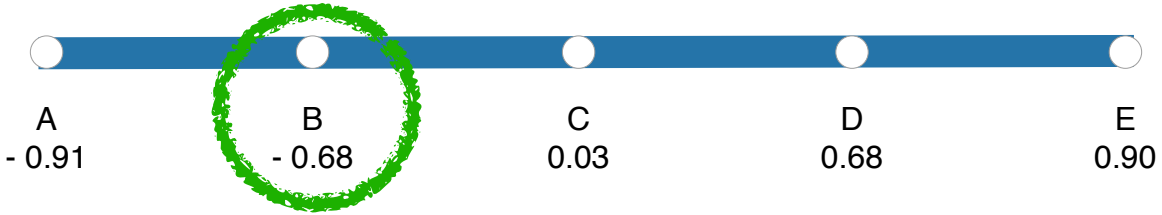
Which of the following is the Pearson correlation coefficient (*r*) for this relationship?



Correlation Champ



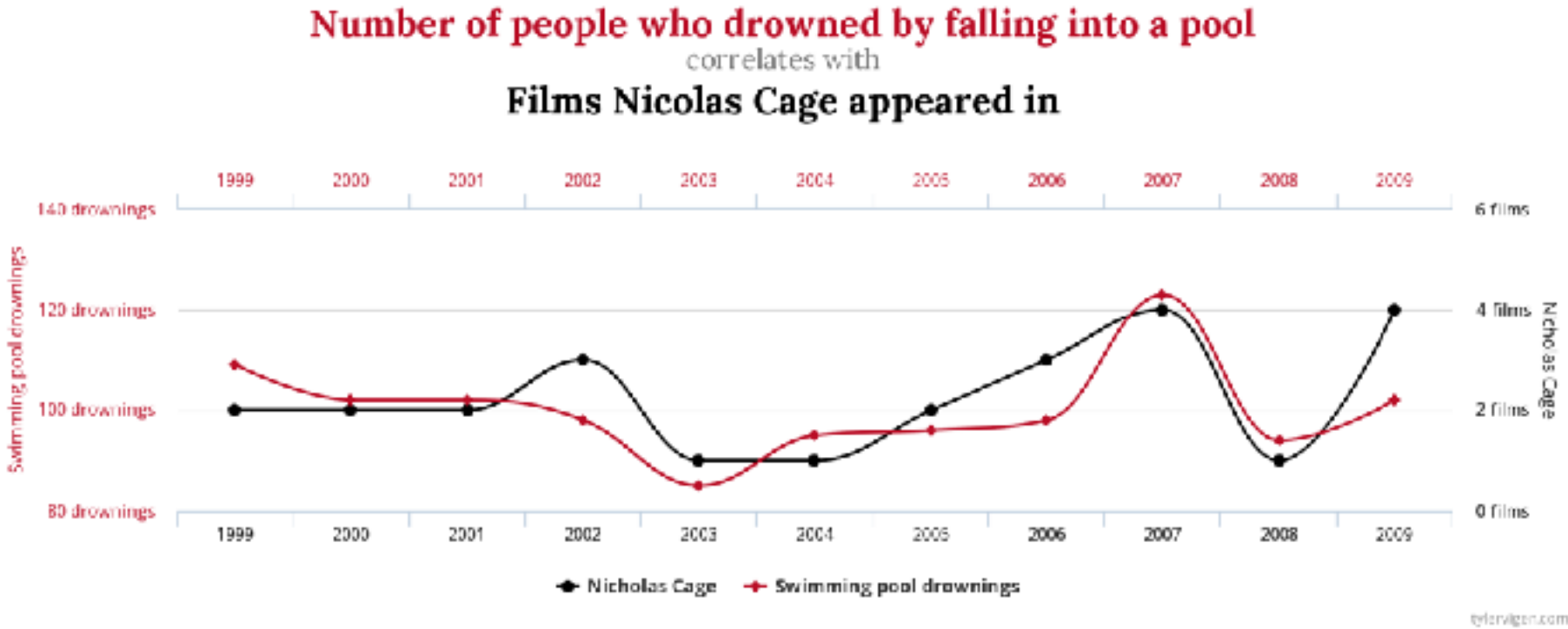
Which of the following is the Pearson correlation coefficient (*r*) for this relationship?



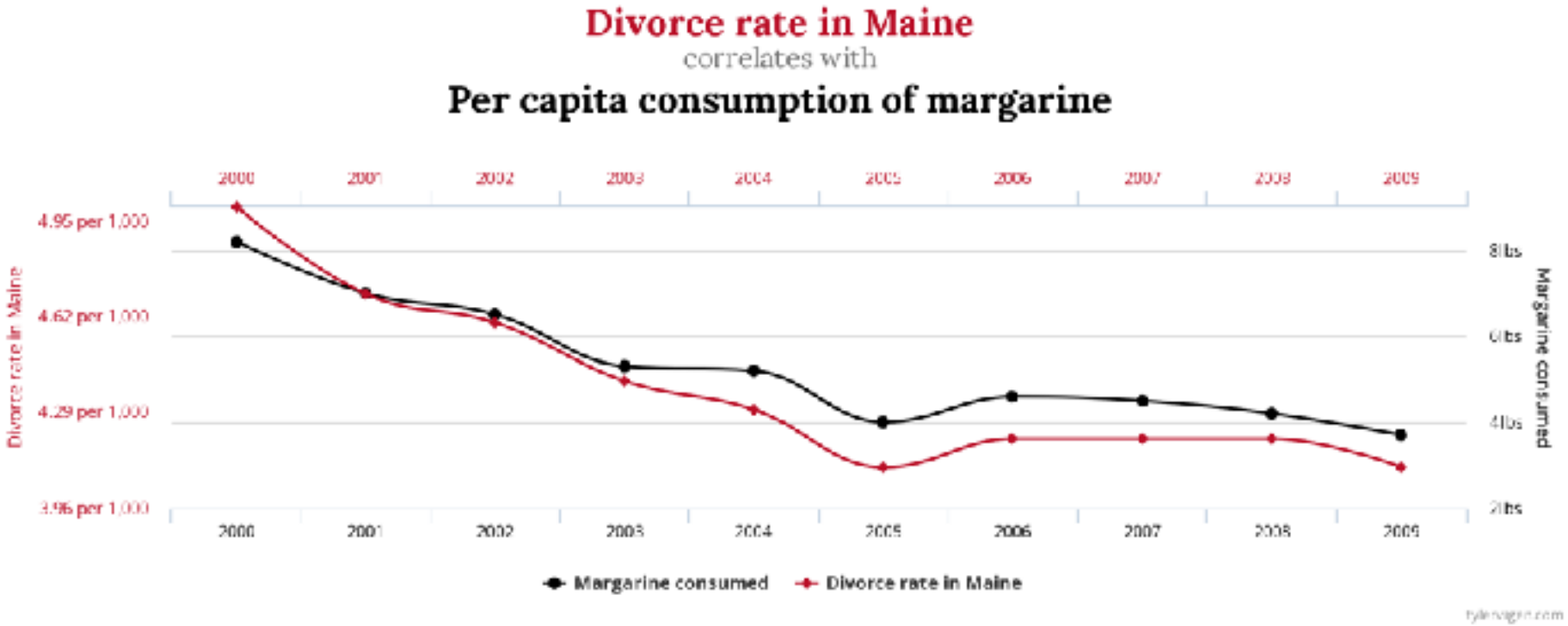
Correlation != Causation

Correlation establishes a relationship.

It does NOT establish causation.



http://www.tylervigen.com/spurious-correlations


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN CONDITIONS

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

t-test:

tests for difference in means between groups

William Sealy Gosset (13 June 1876 – 16 October 1937) was an English statistician, chemist and brewer who served as Head Brewer of Guinness and Head Experimental Brewer of Guinness and was a pioneer of modern statistics. He pioneered small sample experimental design and analysis with an economic approach to the logic of uncertainty. Gosset published under the pen name **Student** and developed most famously Student's t-distribution – originally called Student's "z" – and "Student's test of statistical significance".[1]

Contents [hide]

Life and career [edit]

Born in Canterbury, England the eldest son of Agnes Sealy Vidal and Colonel Frederic Gosset, R.E. Royal Engineers, Gosset attended Winchester College before matriculating as Winchester Scholar in natural sciences and mathematics at New College, Oxford. Upon graduating in 1899, he joined the brewery of Arthur Guinness & Son in Dublin, Ireland; he spent the rest of his 38-year career at Guinness.[1][2]

Gosset had three children with Marjory Gosset (née Philpotts). Harry Gosset (1907–1965) was a consultant paediatrician; Bertha Marian Gosset (1909–2004) was a geographer and nurse; the youngest, Ruth Gosset (1911–1953) married the Oxford mathematician Douglas Real and had five children.

In his job as Head Experimental Brewer at Guinness, the self-trained Gosset developed new statistical methods – both in the brewery and on the farm – now central to the design of experiments, to proper use of significance testing on repeated trials, and to analysis of economic significance (an early instance of decision theory interpretation of statistics) and more, such as his small-sample, stratified, and repeated balanced experiments on barley for proving the best yielding varieties.[3] Gosset acquired that knowledge by study, by trial and error, by cooperating with others, and by spending two terms in 1906–1907 in the Biometrics laboratory of Karl Pearson.[4] Gosset and Pearson had a good relationship.[4] Pearson helped Gosset with the mathematics of his papers, including the 1908 papers, but had little appreciation of their importance. The papers addressed the brewers concern with small samples; biometricians like Pearson, on the other hand, typically had hundreds of observations and saw no urgency in developing small-sample methods.[2]

Gosset's first publication came in 1907, "On the Error of Counting with a Haemacytometer," in which – unbeknownst to Gosset aka "Student" – he rediscovered the Poisson distribution.[3] Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness

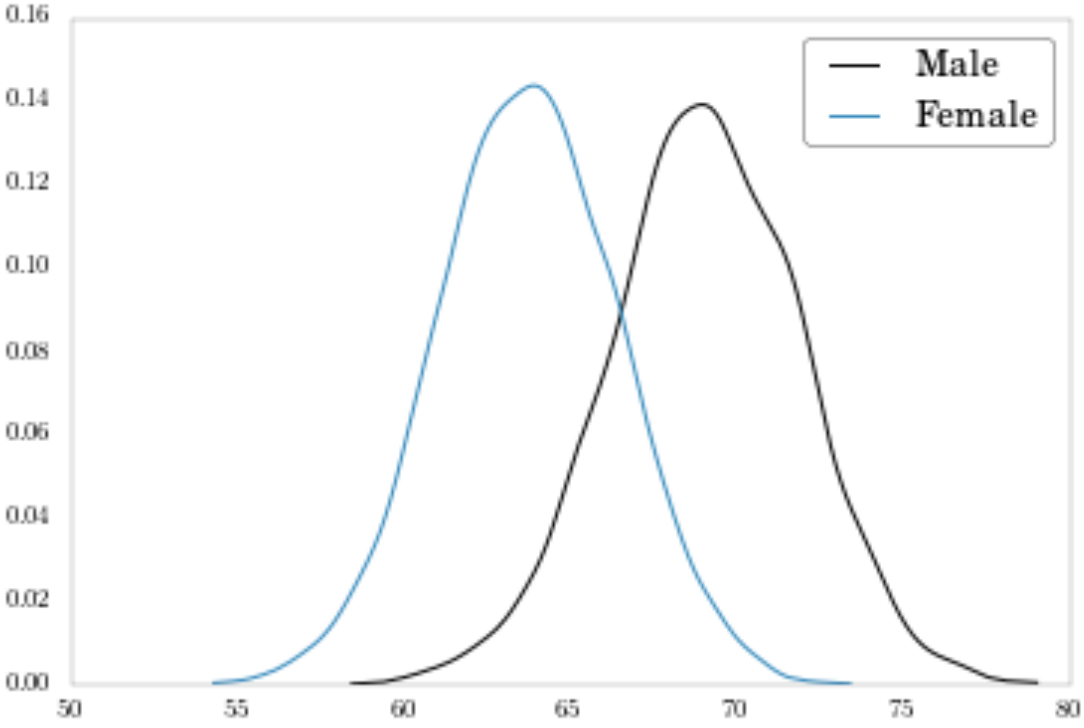
William Sealy Gosset

William Sealy Gosset (aka Student) in 1908 (age 32)

Born	13 June 1876 Canterbury, Kent, England
Died	16 October 1937 (aged 61) Beaconsfield, Buckinghamshire, England
Other names	Student
Alma mater	New College, Oxford, Winchester College
Known for	Student's t-distribution, statistical significance, design of experiments, Monte Carlo method, quality control, Modern synthesis, agricultural economics, econometrics
Children	5, including Isaac Henry Gosset

Scientific career

Do the heights between males and females differ?



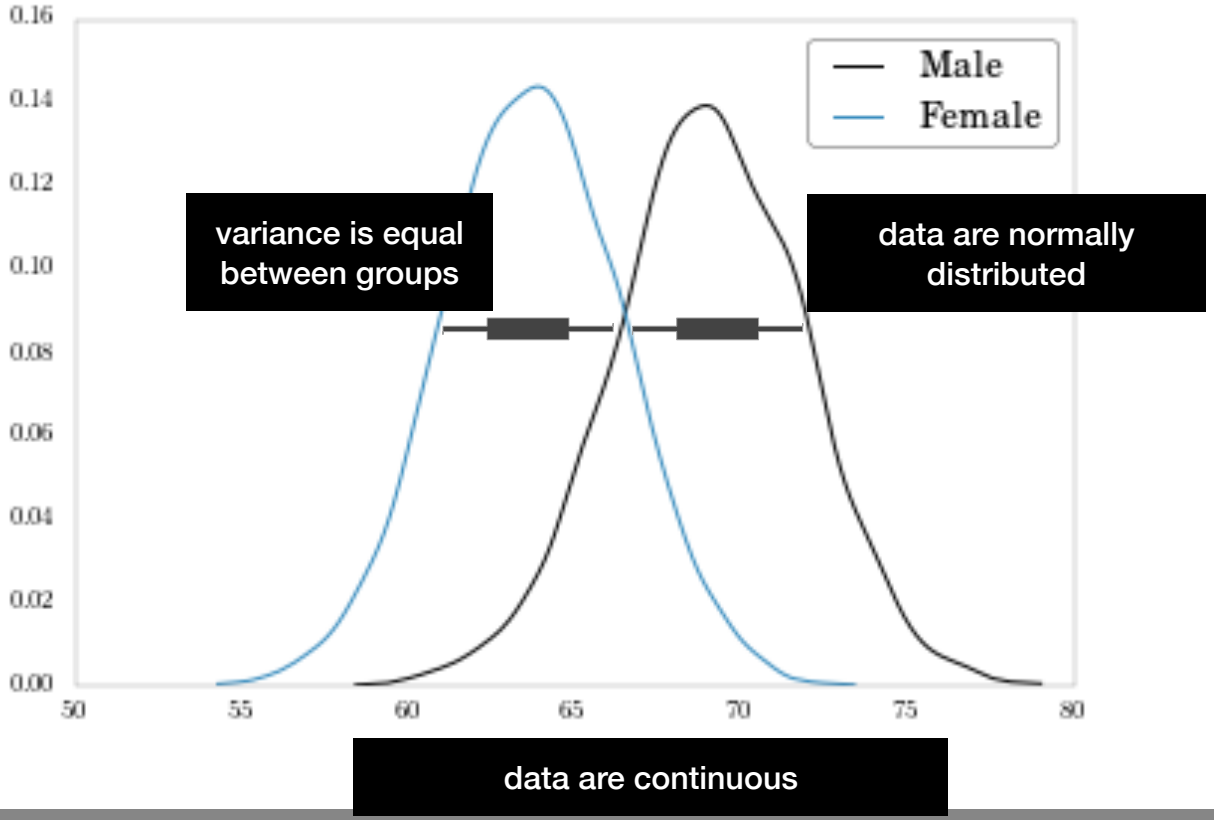
N=10,000

http://slendermeans.org/ml4h-ch2-p1.html

t-test Assumptions

1. Data are continuous
2. Normally distributed
3. Equal variance b/w groups (but can use Welch's test!)
4. Not paired (will talk more about this later)

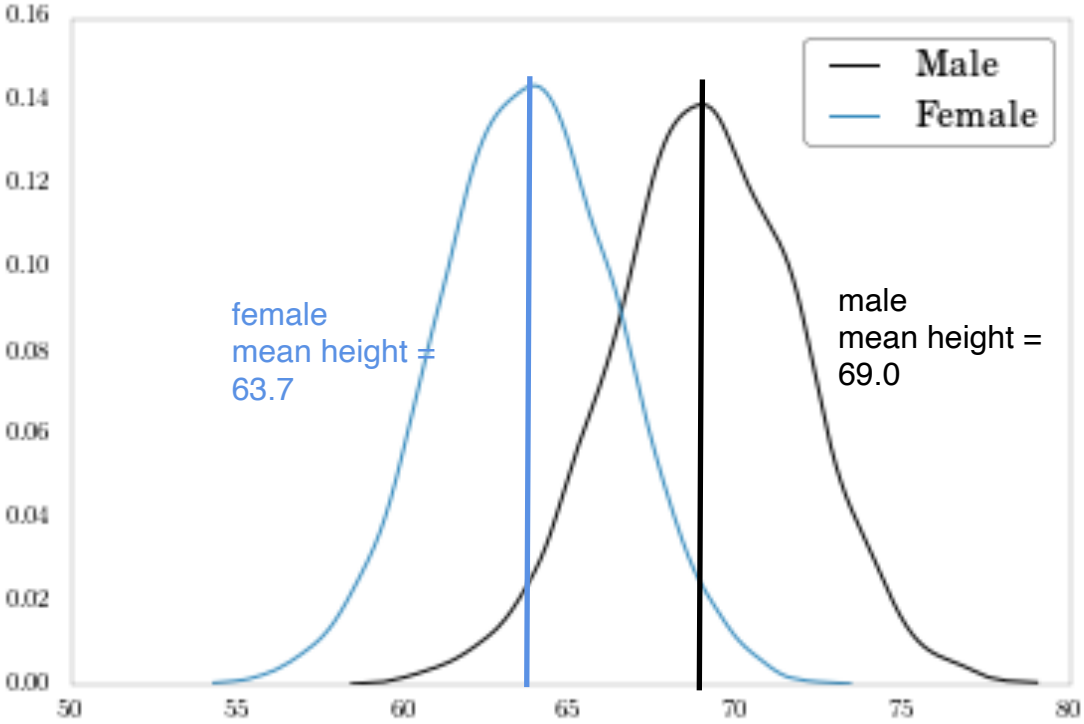
Do the heights between males and females differ?



sample size affects statistic

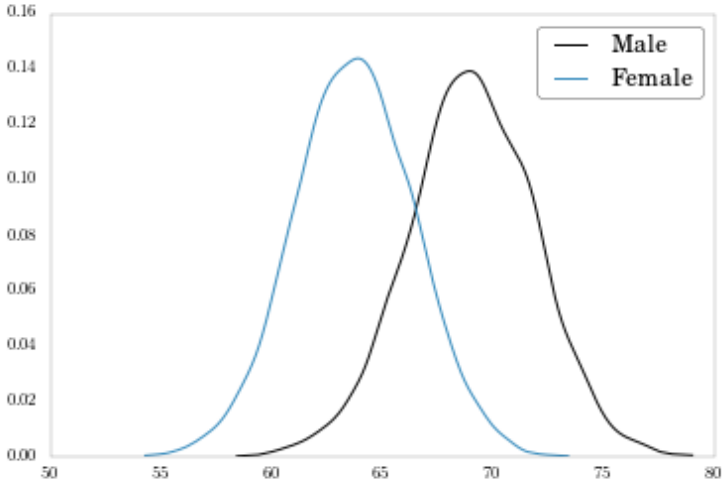
N=10,000

Do the heights between males and females differ?



N=10,000

Do the heights between males and females differ?



t-statistic: -95.6

p-value << 0.001

95% CI for true difference in means [-5.43, -5.21]

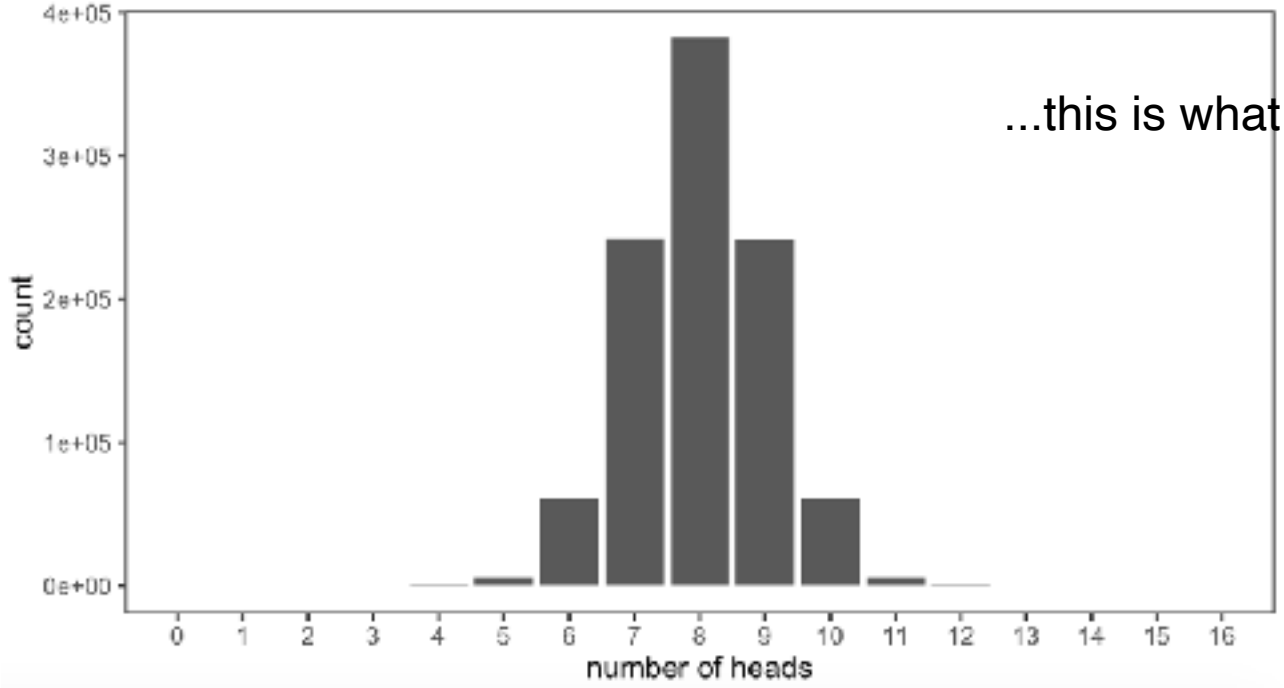
Yes.

p-value : the probability of getting the observed results (or results more extreme) by chance alone

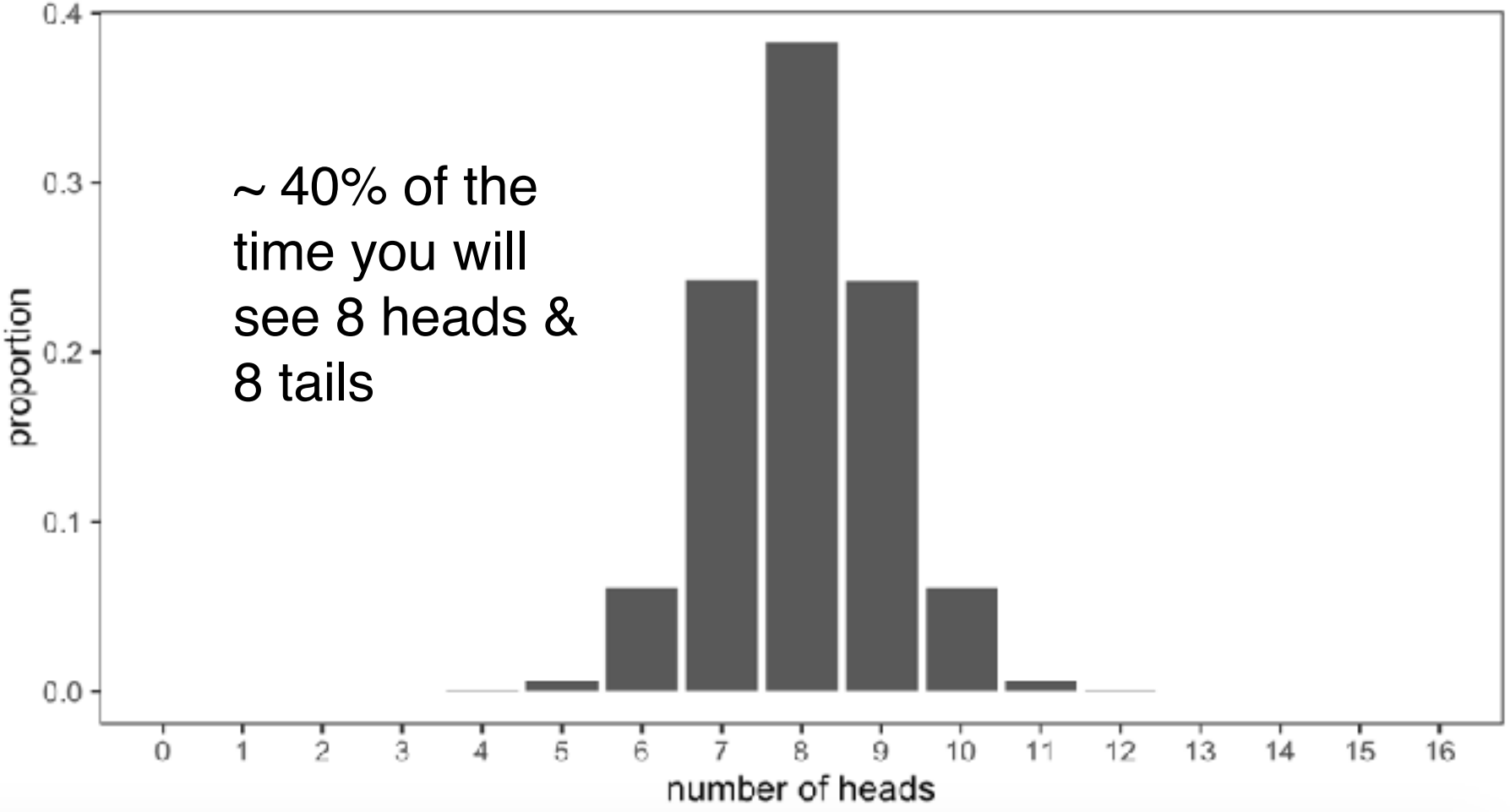
https://forms.gle/6MCyp7qFsaHgGKi5A

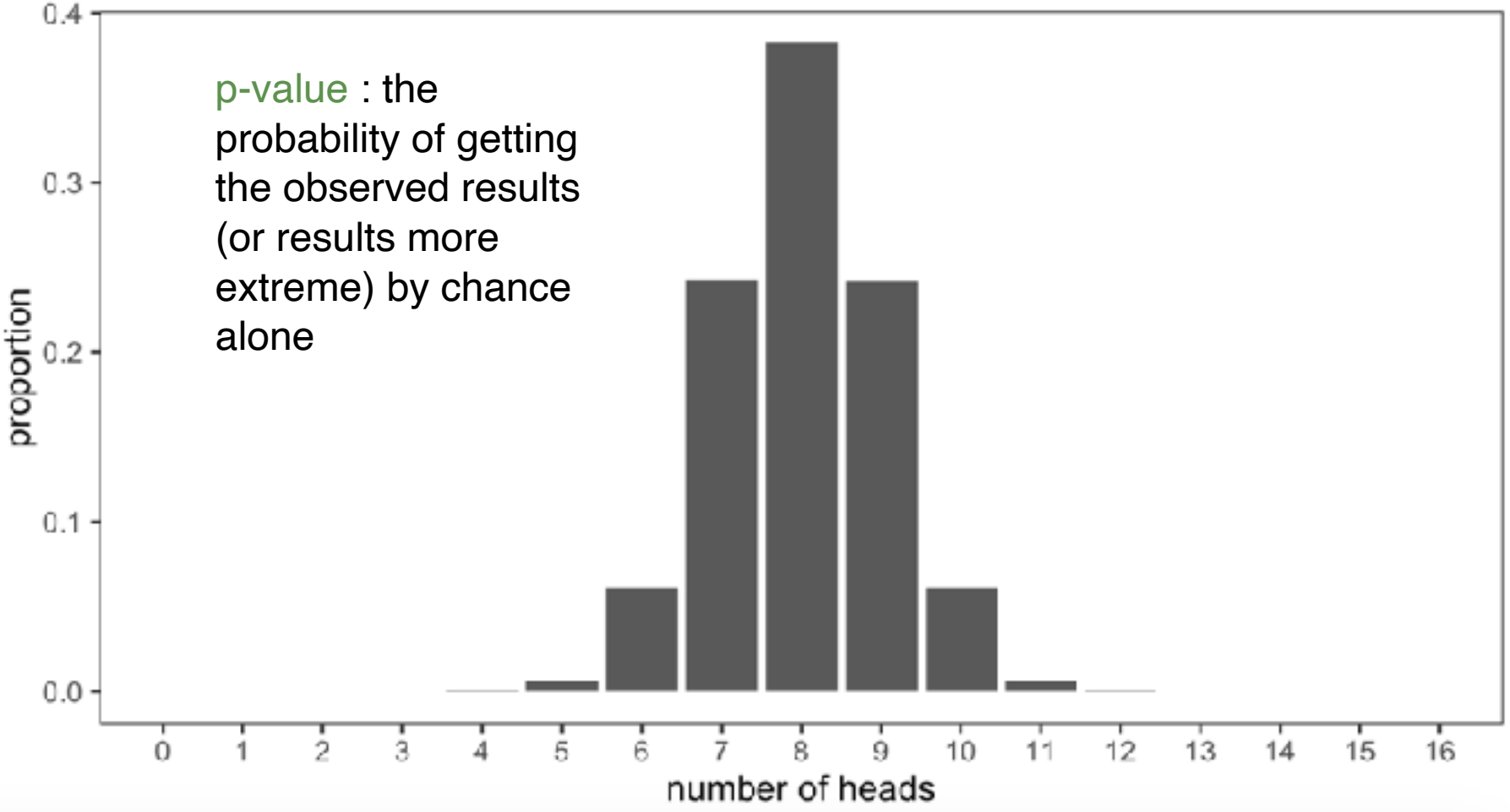
If we flip a coin 16 times and record the number of heads....

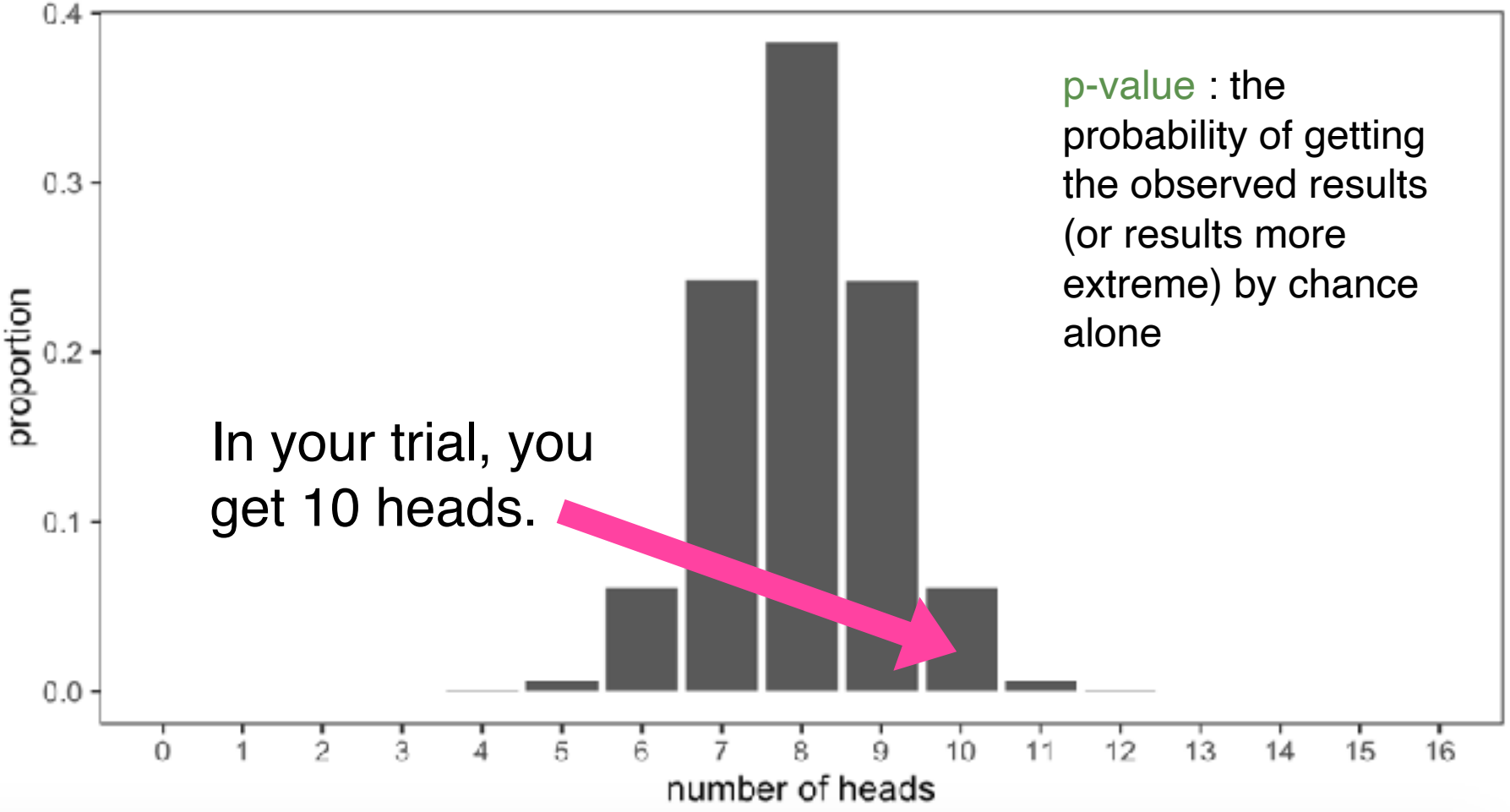
….and then do that 1M times

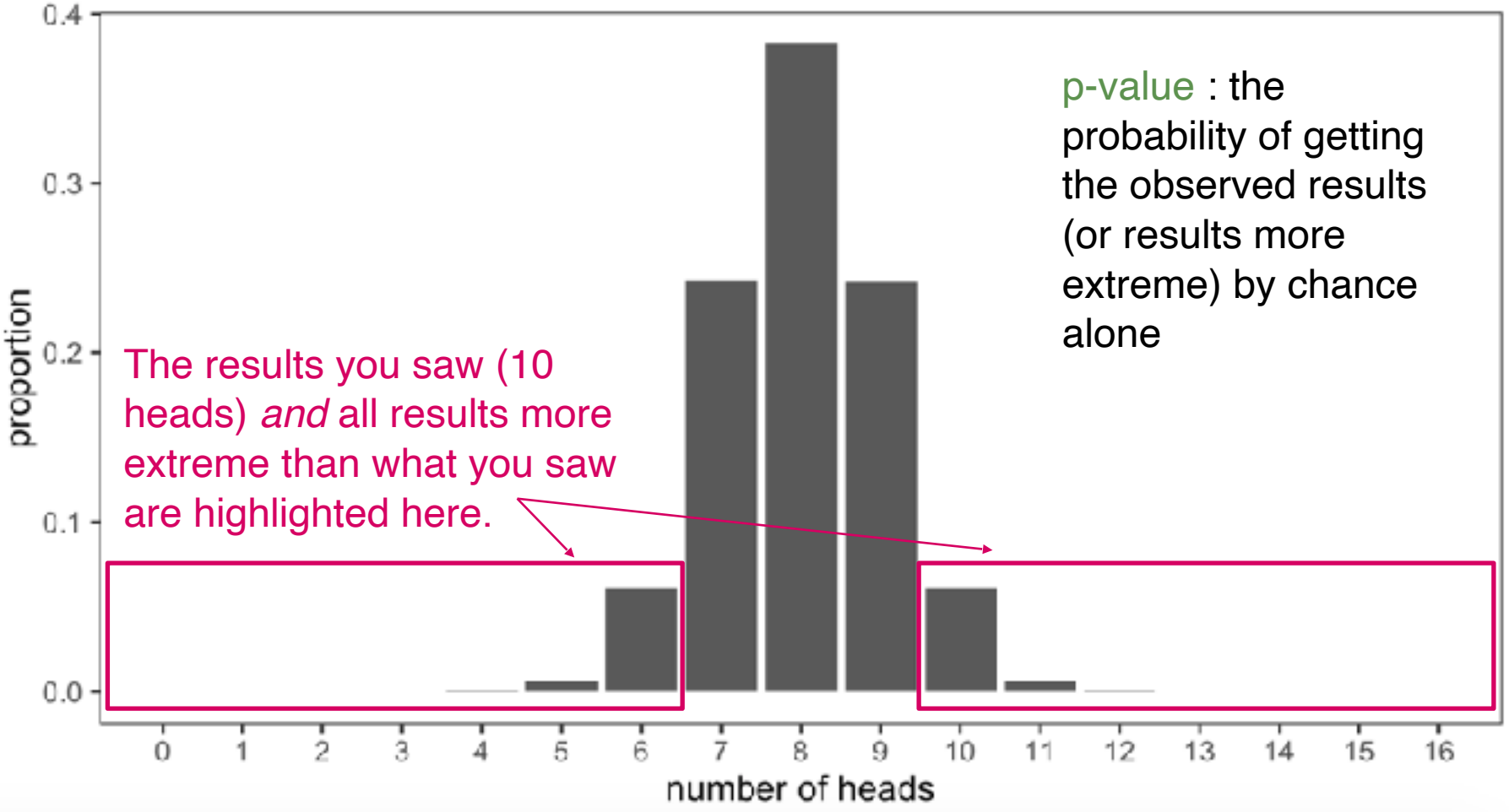


...this is what happens by chance alone.


~ 40% of the time you will see 8 heads & 8 tails
proportion
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

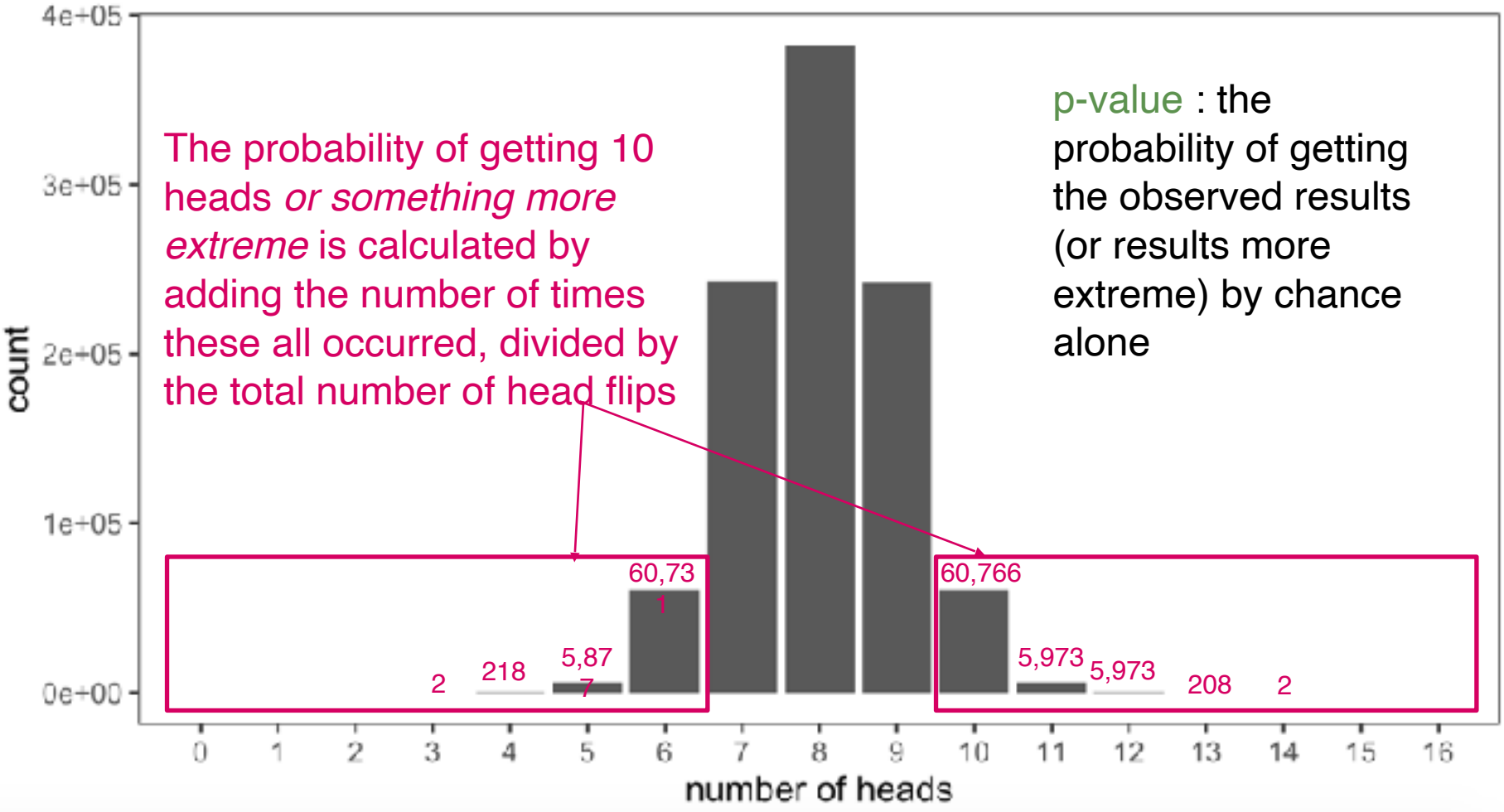

p-value : the probability of getting the observed results (or results more extreme) by chance alone
proportion
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads


0.4
0.3
0.2
0.1
0.0
proportion
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads
In your trial, you get 10 heads.
p-value : the probability of getting the observed results (or results more extreme) by chance alone



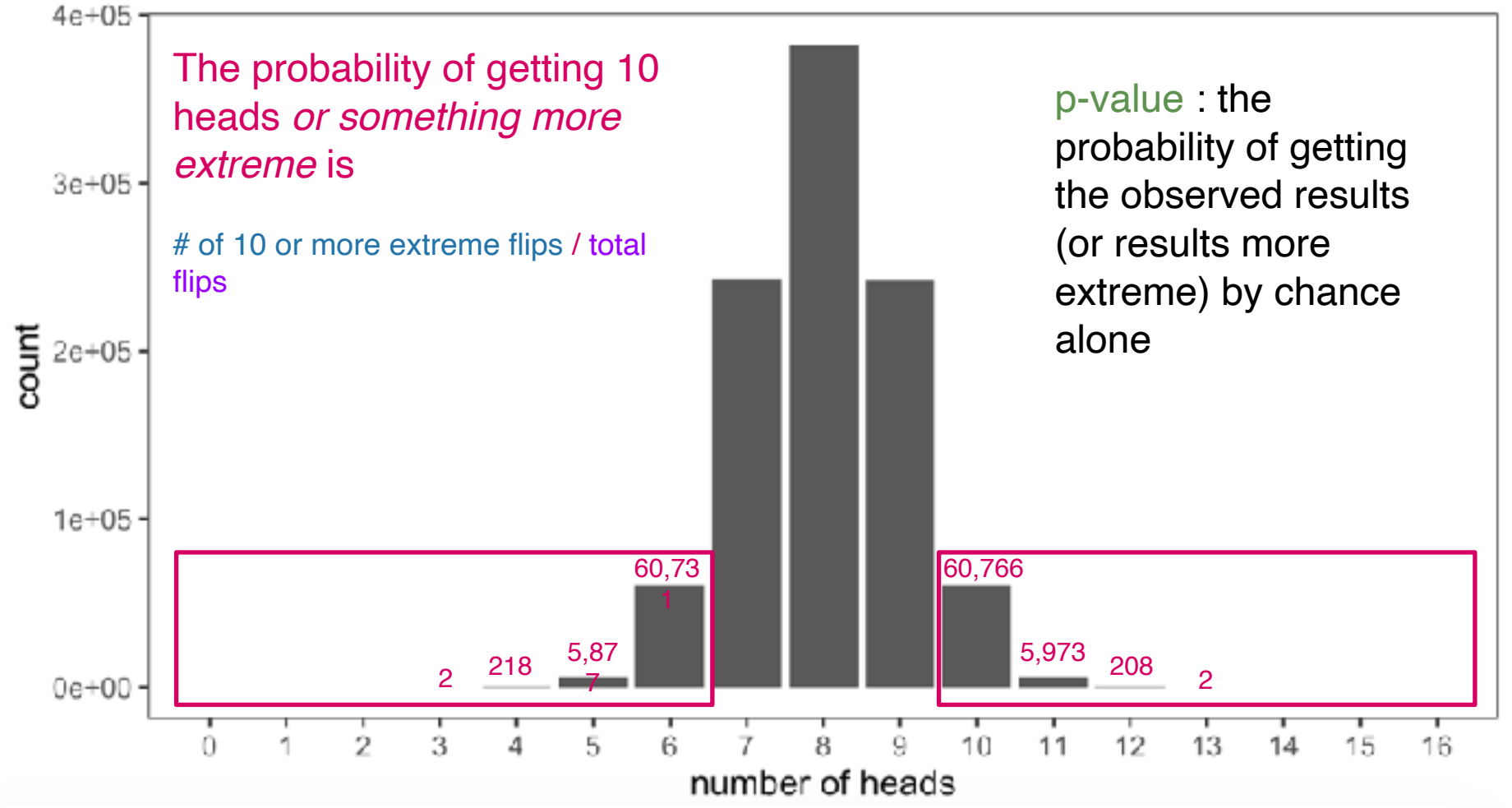
p-value : the probability of getting the observed results (or results more extreme) by chance alone

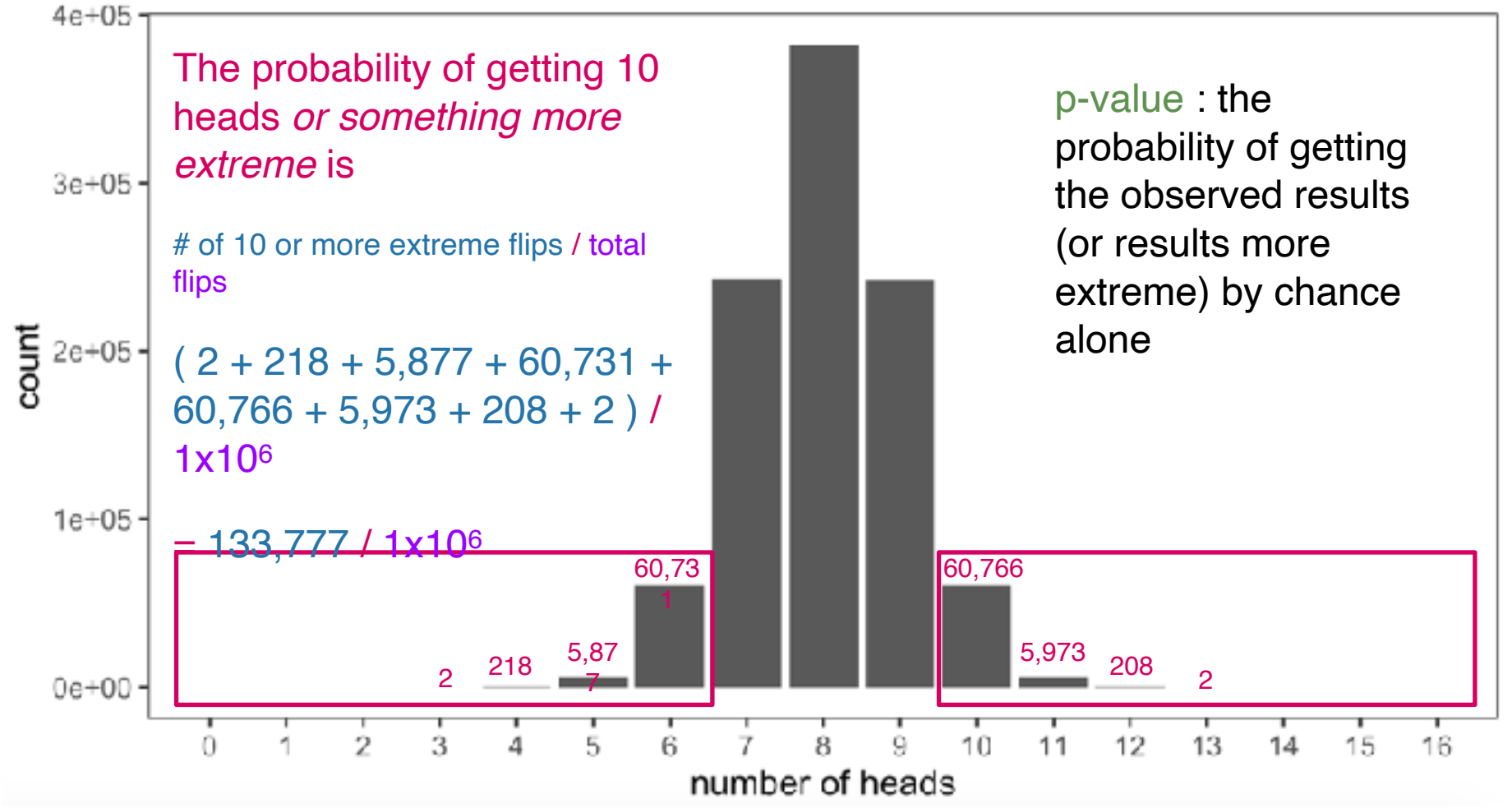
The results you saw (10 heads) *and* all results more extreme than what you saw are highlighted here.

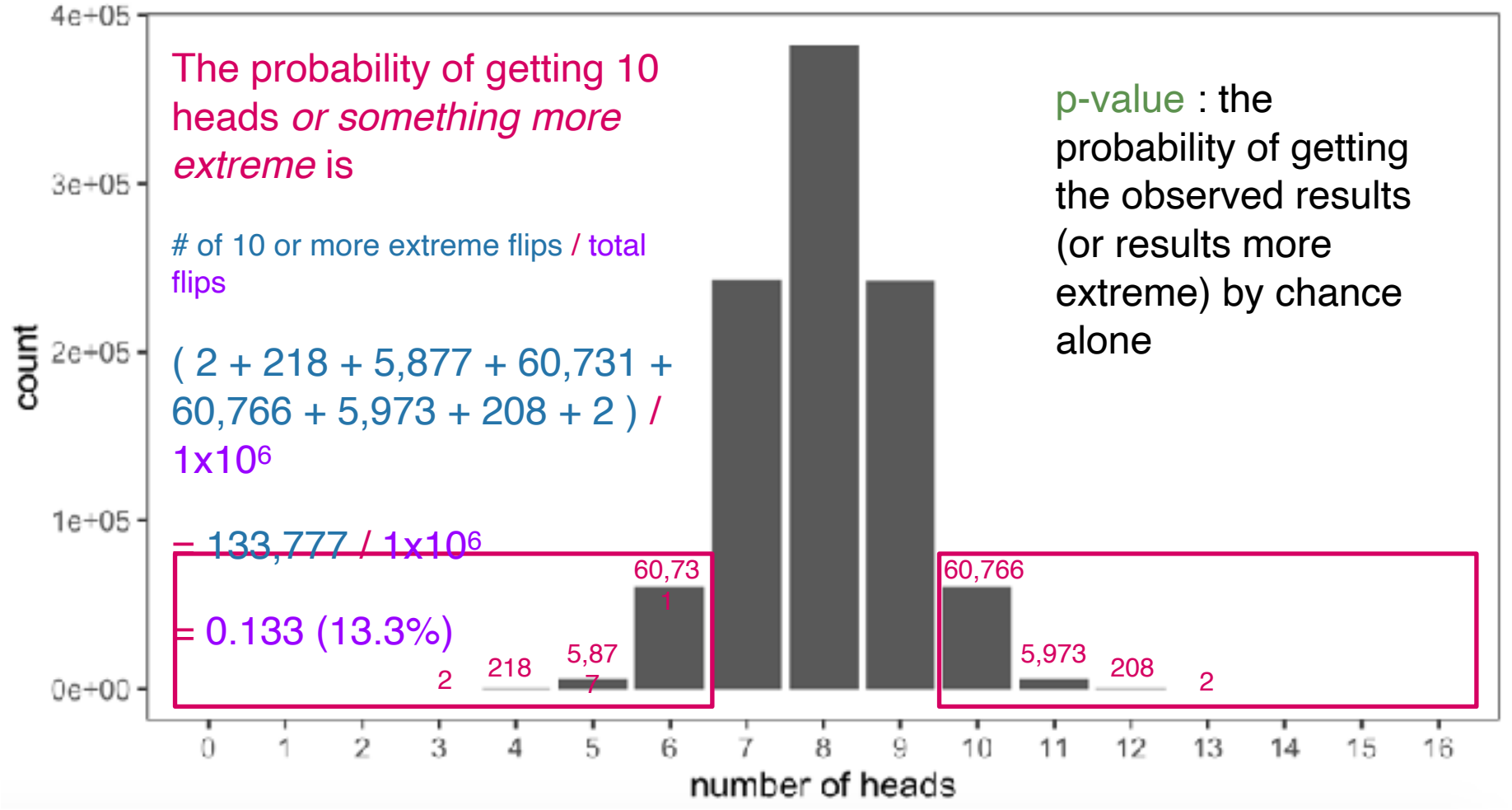


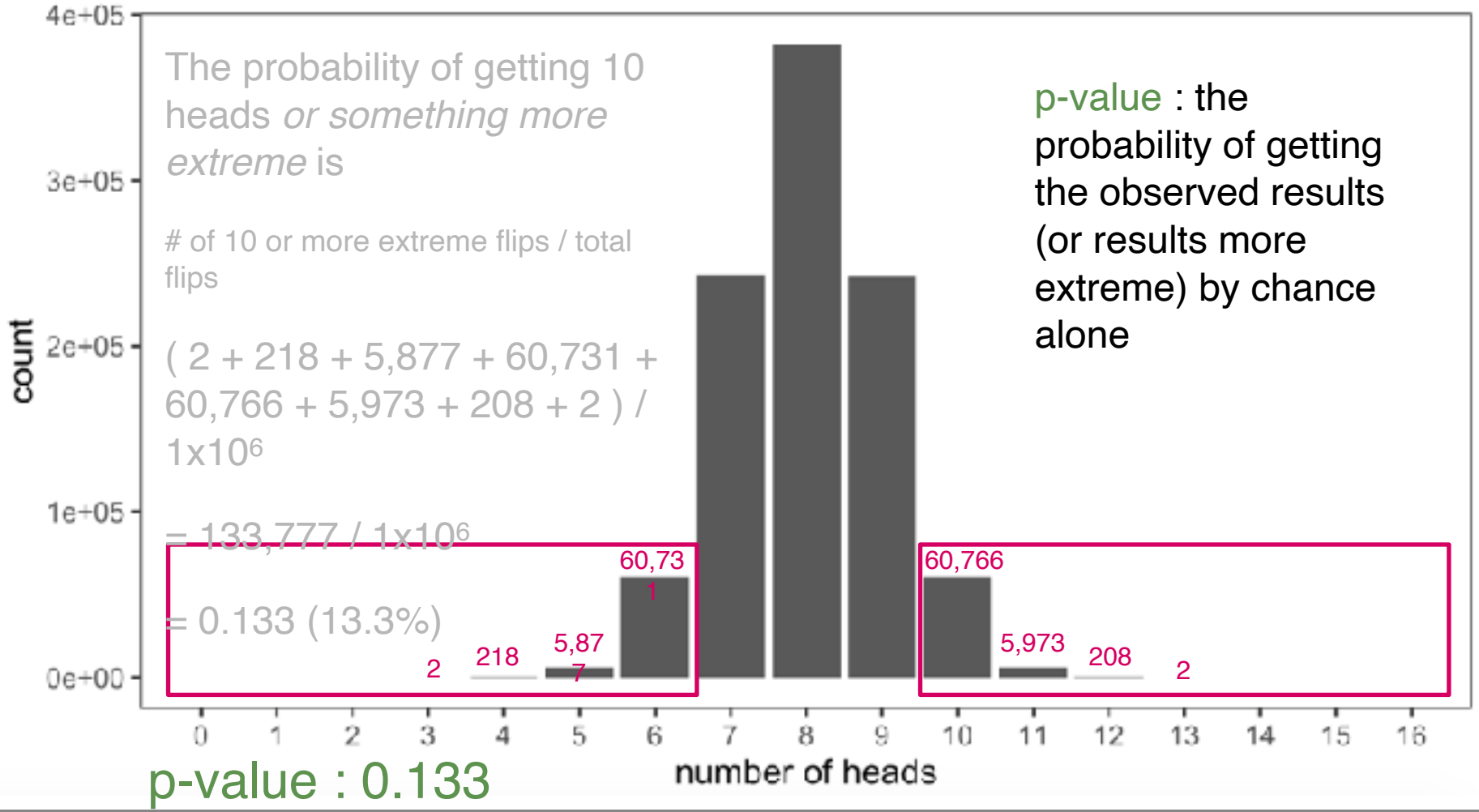
The probability of getting 10 heads *or something more extreme* is calculated by adding the number of times these all occurred, divided by the total number of head flips

p-value : the probability of getting the observed results (or results more extreme) by chance alone


The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,87
7
60,73
1
60,766
5,973
208
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads


The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
p-value : the probability of getting the observed results (or results more extreme) by chance alone
4e+05
3e+05
2e+05
1e+05
0e+00
count
2
218
5,87
7
60,73
1
60,766
5,973
208
2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads


The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
= 0.133 (13.3%)
p-value : the probability of getting the observed results (or results more extreme) by chance alone
count
4e+05
3e+05
2e+05
1e+05
0e+00
2
218
5,87
7
60,73
1
60,766
5,973
208
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads



The probability of getting 10 heads *or something more extreme* is

# of 10 or more extreme flips / total flips

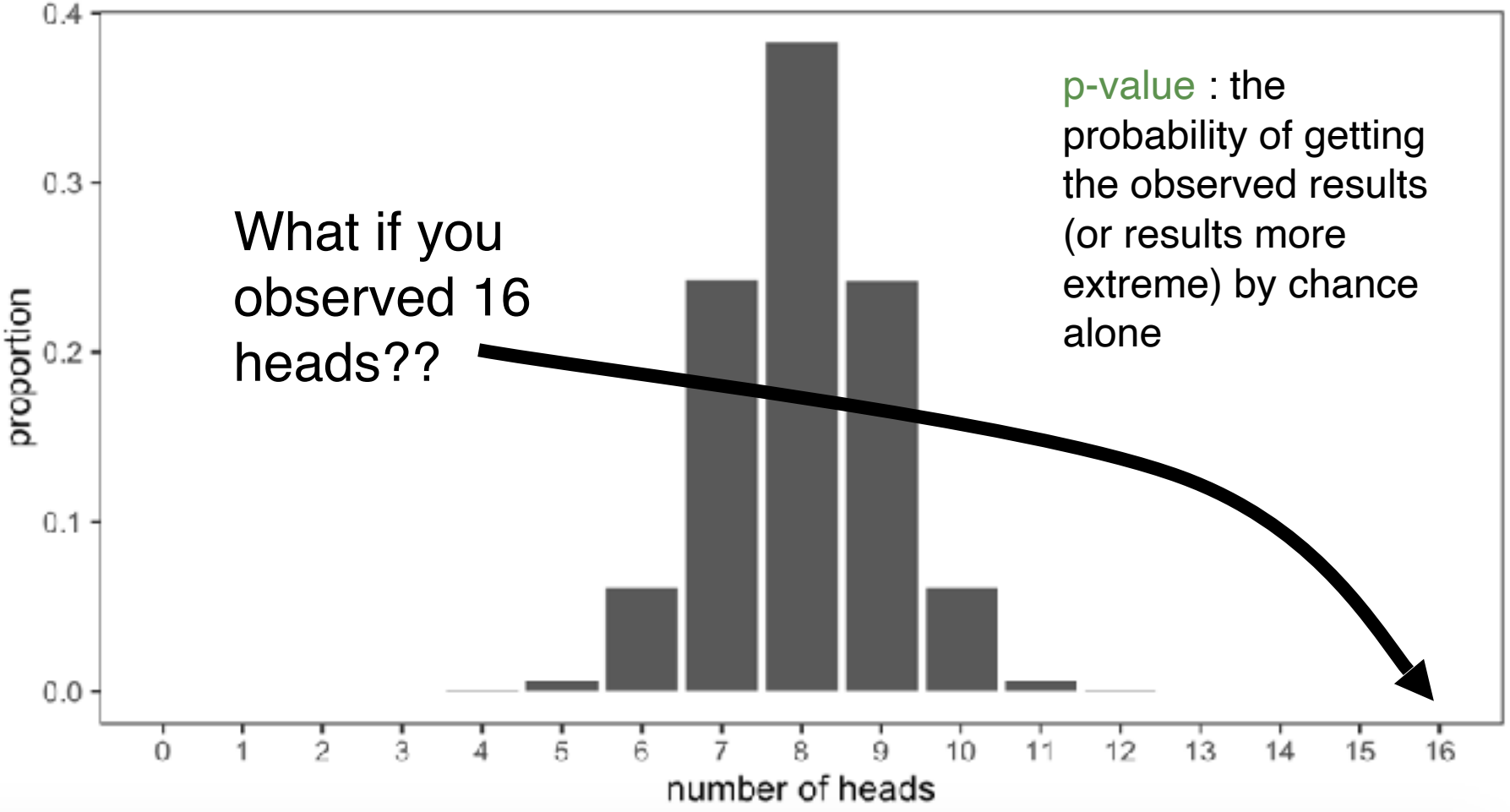
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / $1x10^6$

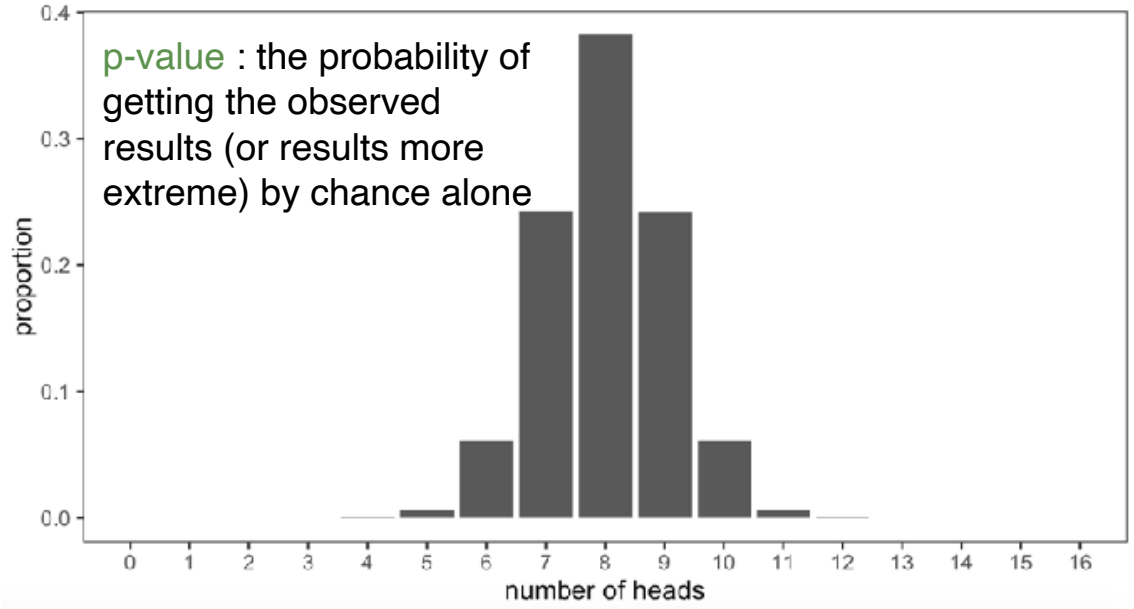
= 133,777 / $1x10^6$

= 0.133 (13.3%)

p-value : the probability of getting the observed results (or results more extreme) by chance alone

p-value : 0.133


What if you observed 16 heads??
p-value : the probability of getting the observed results (or results more extreme) by chance alone
proportion
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads



What would be the p-value of you flipping 16 heads?

A
< 0.13

B
> 0.13

Do the heights between males and females differ?

t-statistic: -95.6

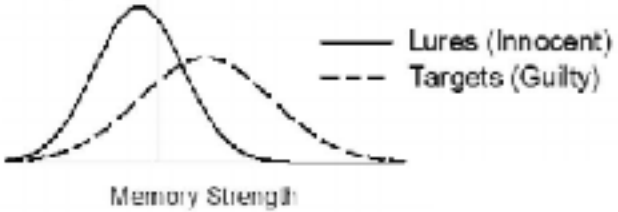
p-value << 0.001

The probability of seeing this difference in means by random chance alone is much less than 1 in 1000

95% CI for true difference in means [-5.43, -5.21]

Yes.

Difference in Means



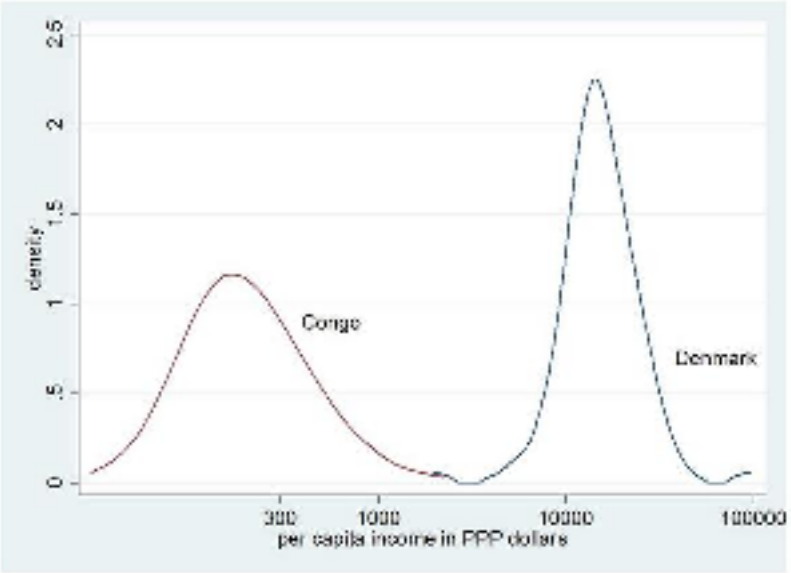
Why would a t-test *not* be appropriate for these data?

A
Not normally distributed

B
Unequal variances

C
Small sample size

D
Data are not continuous



Would a t-test find a significant difference in means?

A
t-test not appropriate

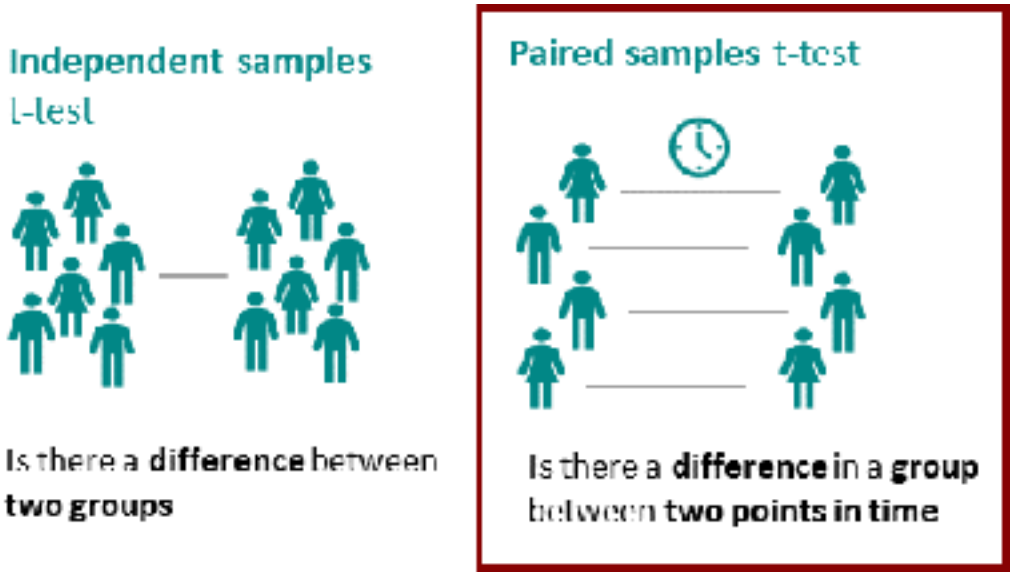
B
Yes

C
No

D
Need more information

https://twitter.com/brankomilan/status/406105955080757249

Paired data



https://datatab.net/tutorial/paired-t-test

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

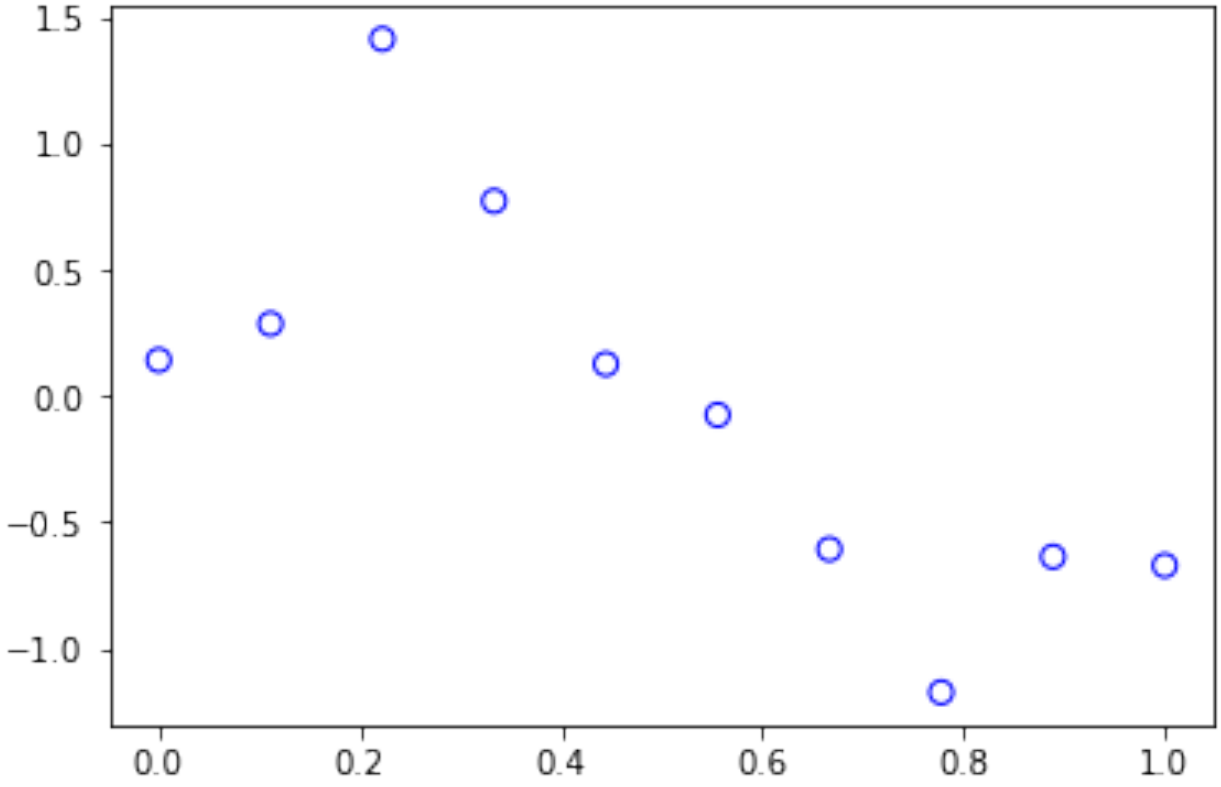
DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test


1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

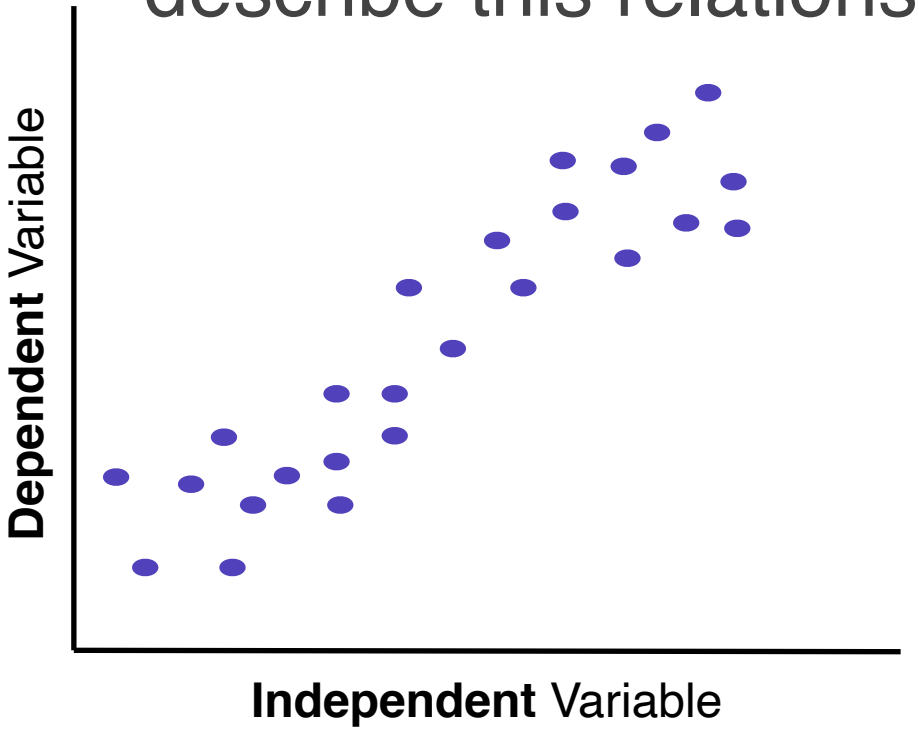
I.e. simple regression, multiple regression

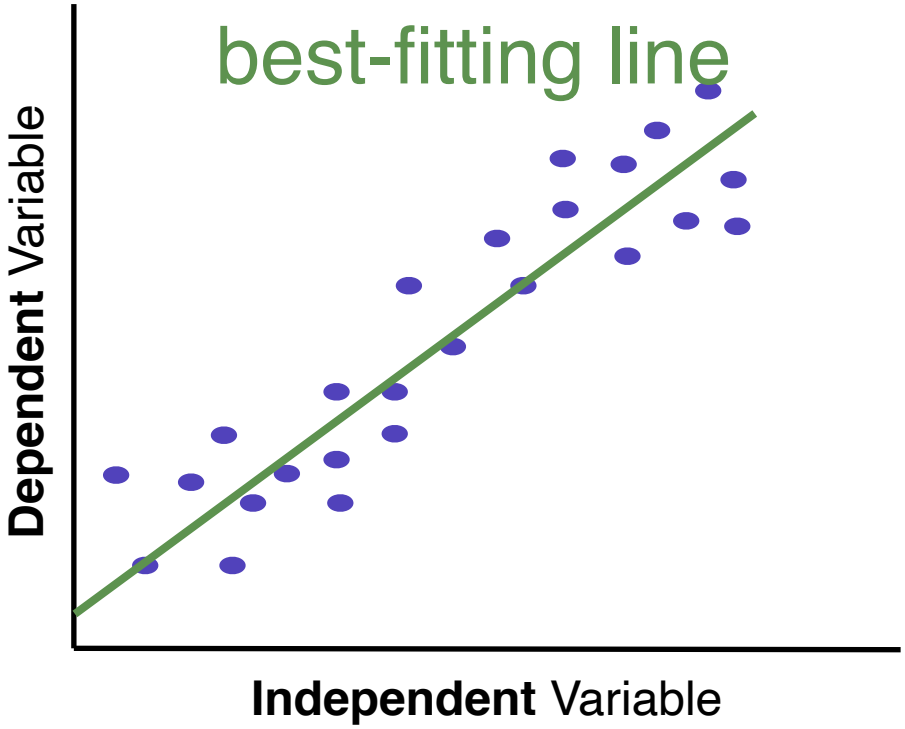
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

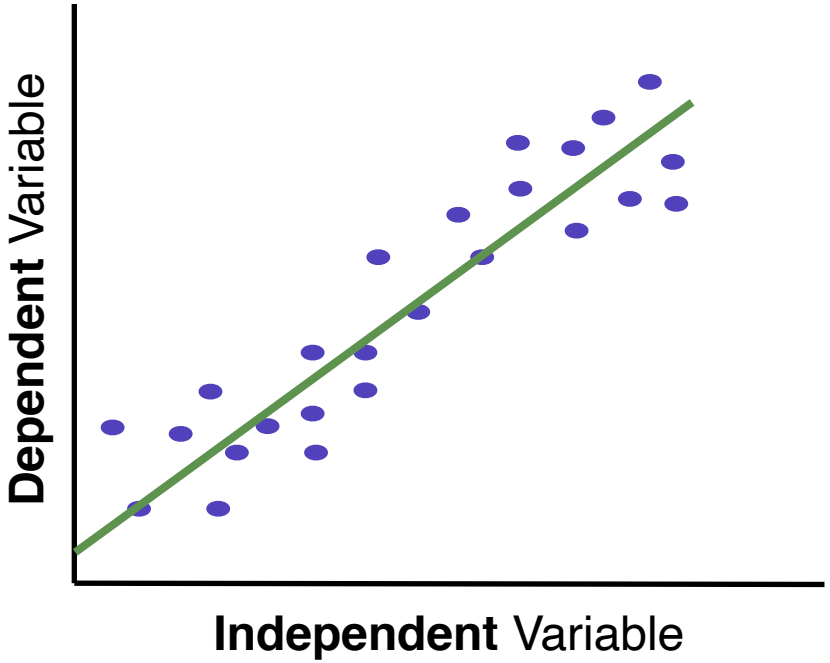
i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

Linear regression can be used to describe this relationship

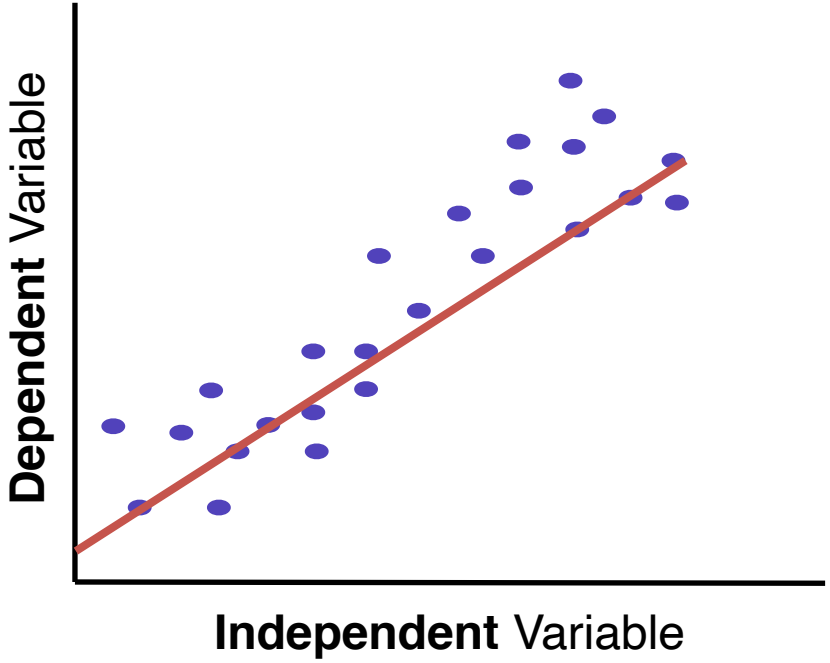



best-fitting line
Dependent Variable
Independent Variable

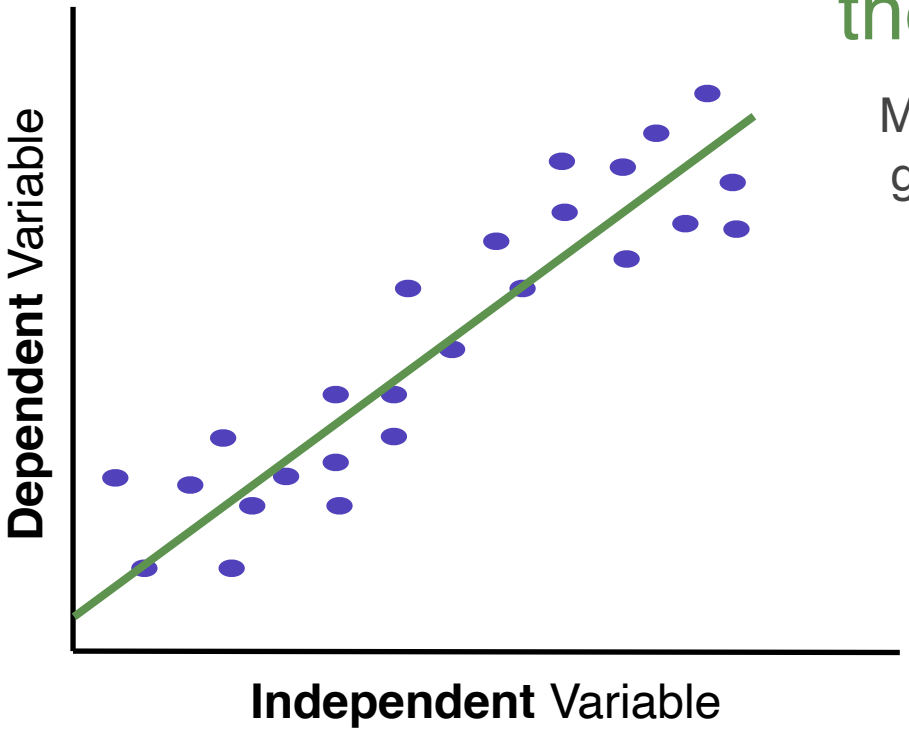
Best-fitting line



NOT a best-fitting line

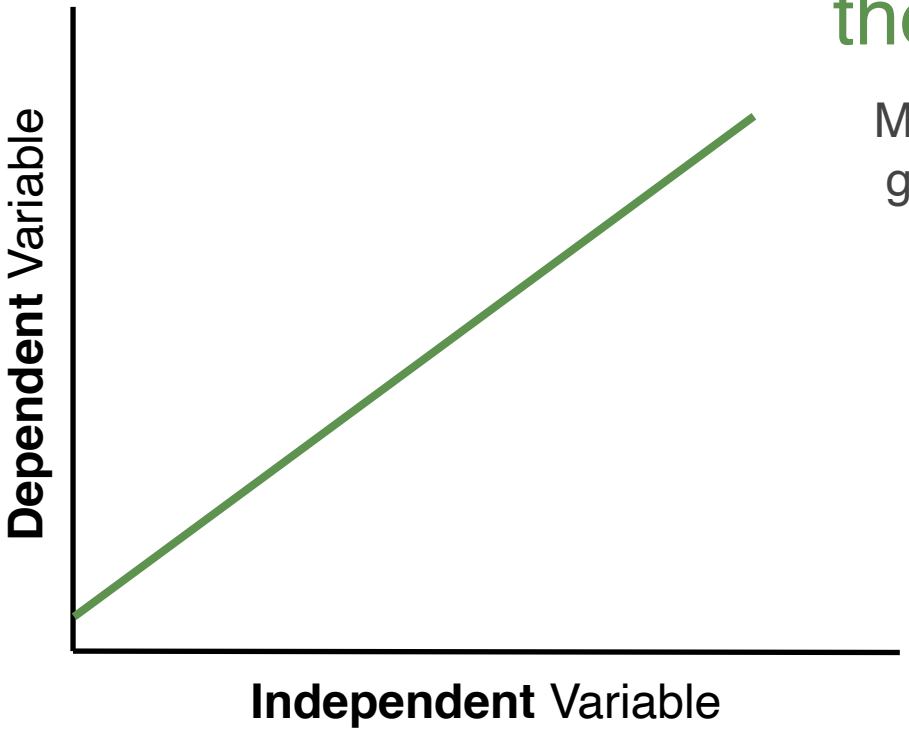


This line is a model of the data



Models are mathematical equations generated to *represent* the real life situation

This line is a model of the data



Models are mathematical equations generated to *represent* the real life situation

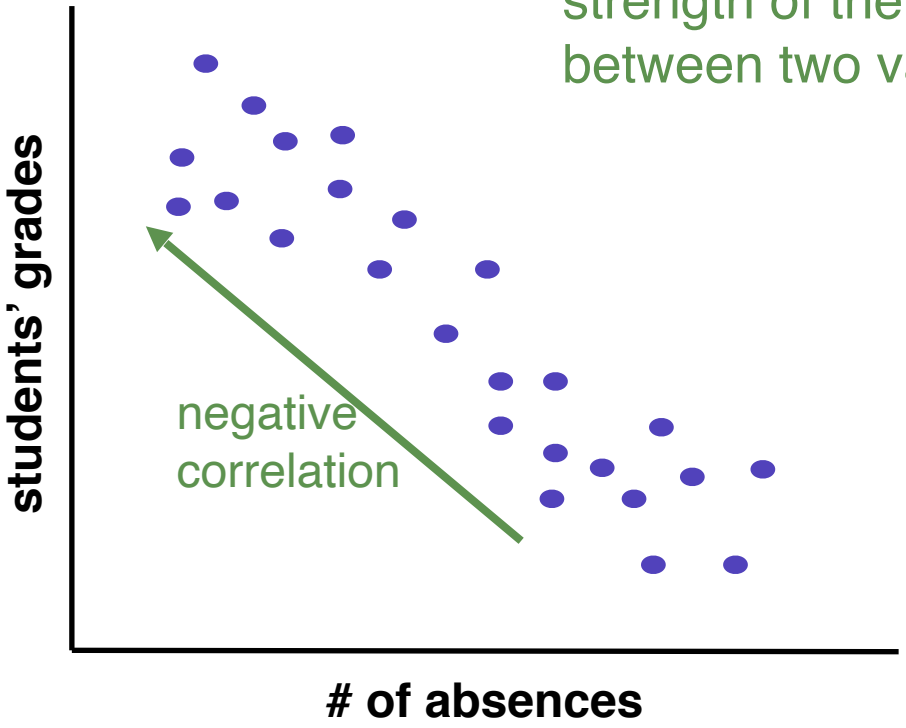
2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

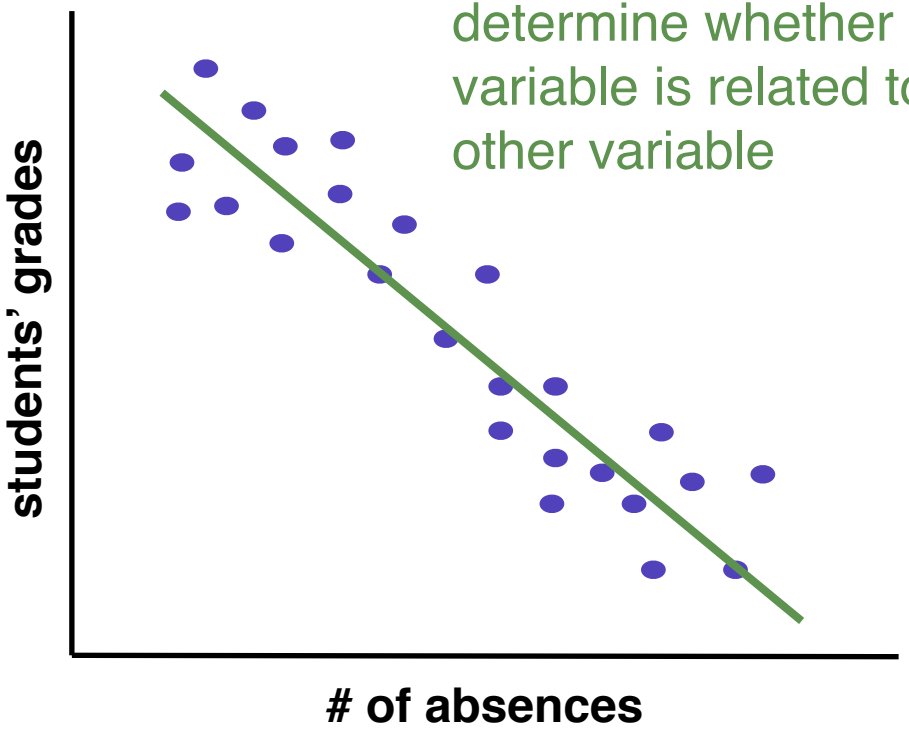
2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

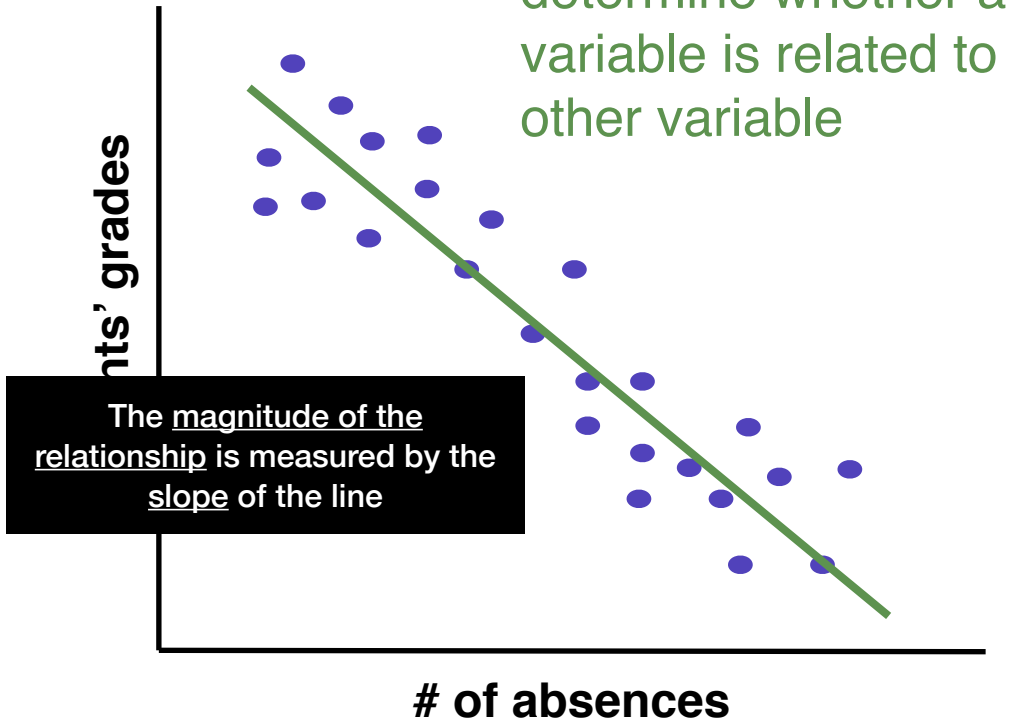
Correlation measures the strength of the linear relationship between two variables



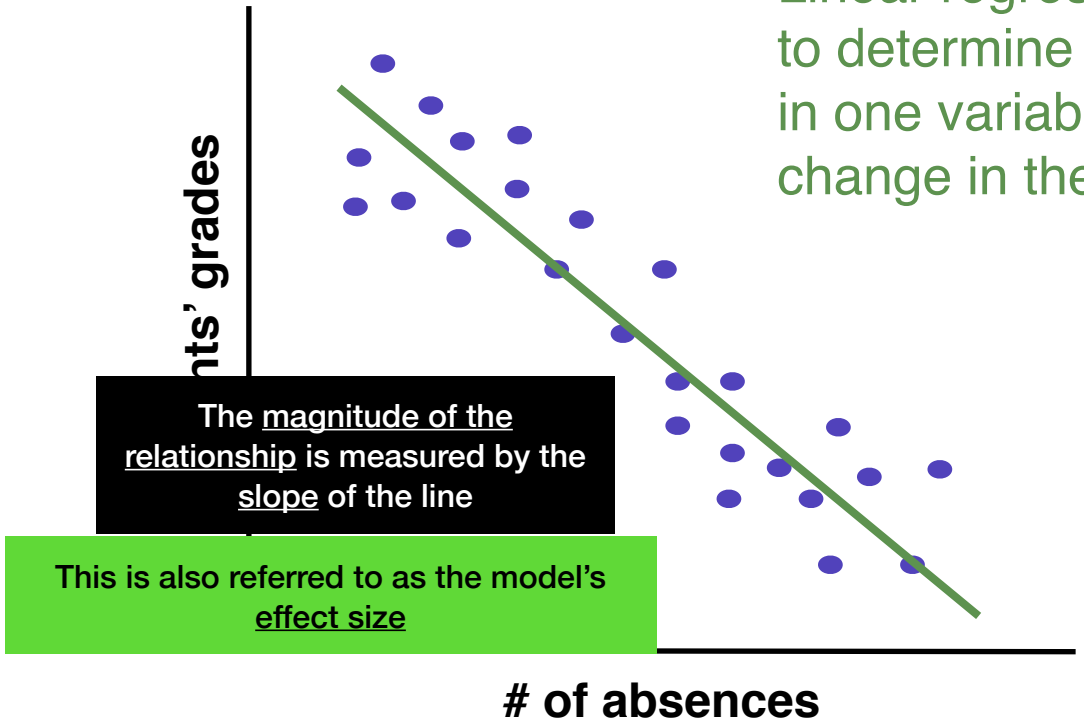
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



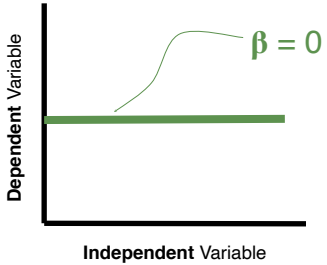
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



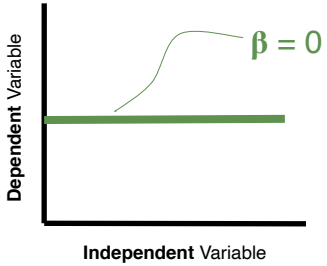
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable

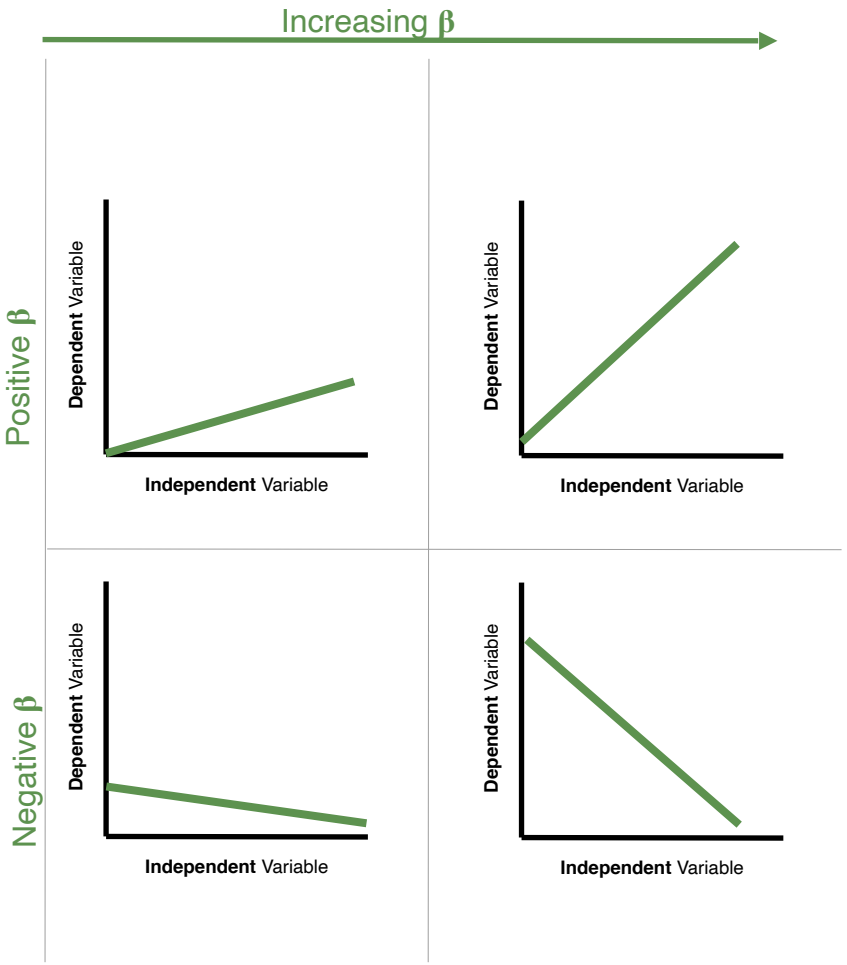


Effect size (β) can be estimated using the slope of the line



Effect size (β) can be estimated using the slope of the line

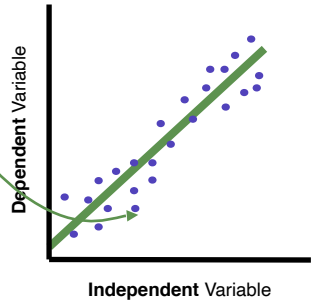


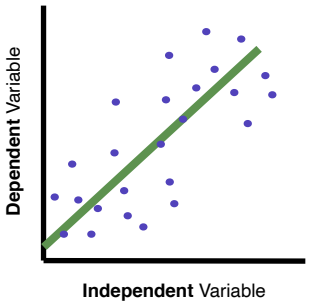


The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

increasing standard error (SE)

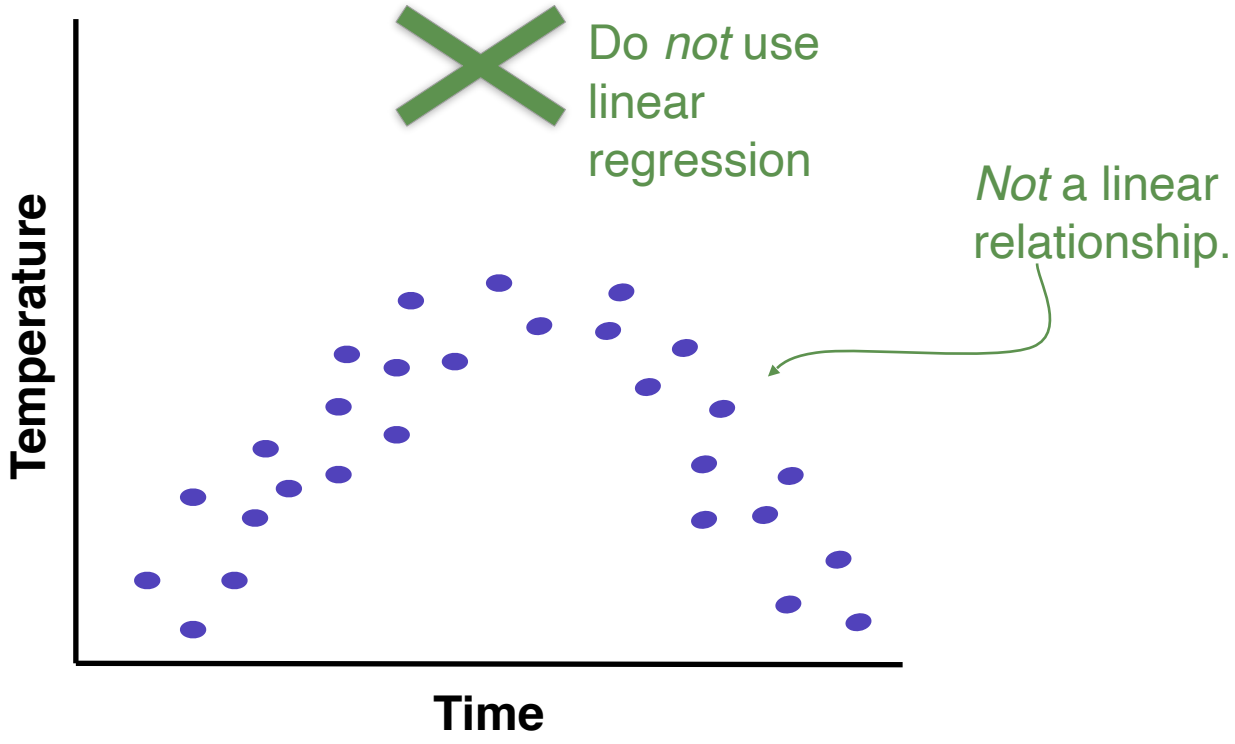
decreasing correlation coefficient



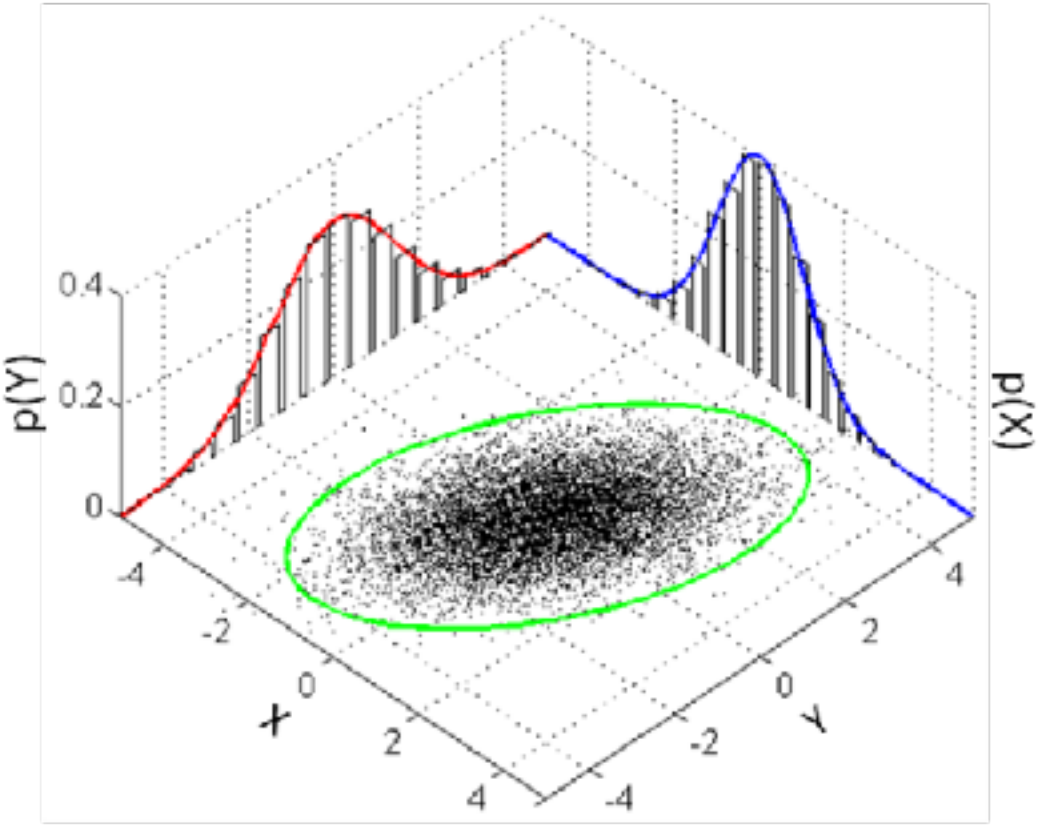


Assumptions of linear regression

1. Linear relationship
2. Multivariate normality
3. No multicollinearity
4. No autocorrelation
5. Homoscedasticity

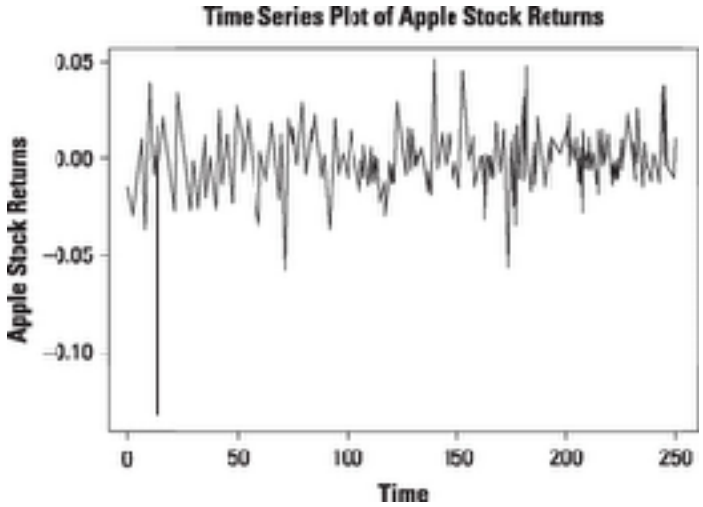

Do *not* use linear regression
Not a linear relationship.
Temperature
Time

A multivariate normal probability distribution (joint normal)



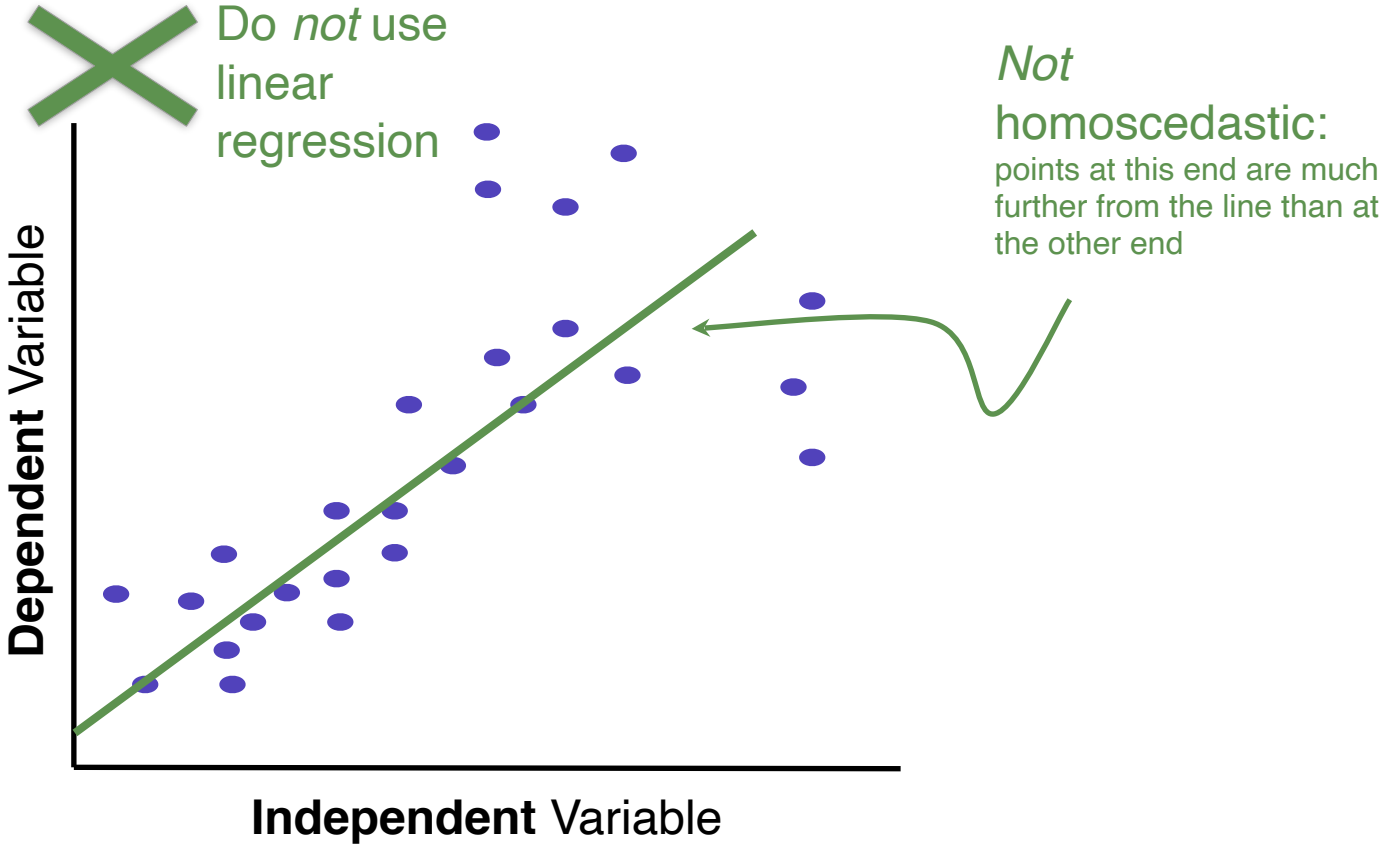
Linear regression assumes no multicollinearity. Multicollinearity occurs when the independent variables (in multiple linear regression) are too highly correlated with each other.

Daily returns are
$\ln(\text{price}_t / \text{price}_{t-1})$

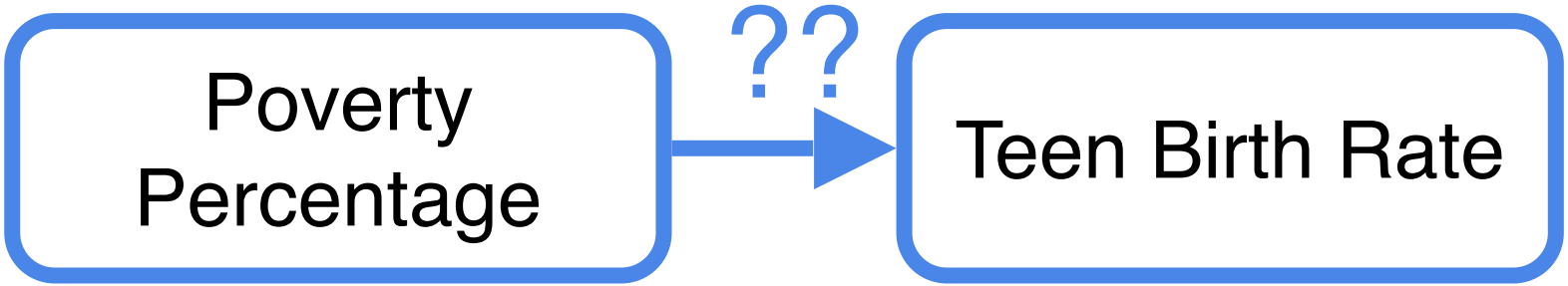


Autocorrelation occurs when the observations are *not* independent of one another (i.e. stock prices)

https://www.dummies.com/programming/big-data/data-science/autocorrelation-plots-graphical-technique-for-statistical-data//


Do *not* use linear regression
Not homoscedastic:
points at this end are much further from the line than at the other end
Dependent Variable
Independent Variable

Does Poverty Percentage affect Teen Birth Rate?


Poverty Percentage
??
Teen Birth Rate

What is the relationship between Poverty Percentage & Teen Birth Rate?

What's your hypothesis?

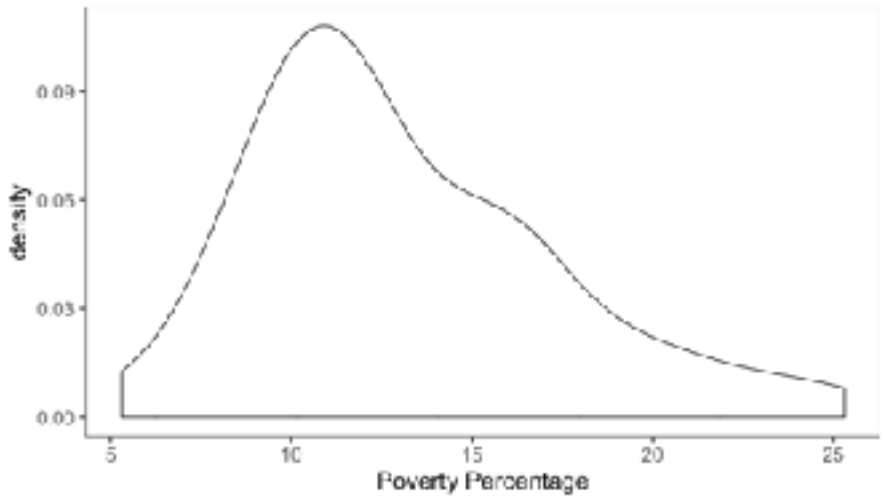
A
No relationship

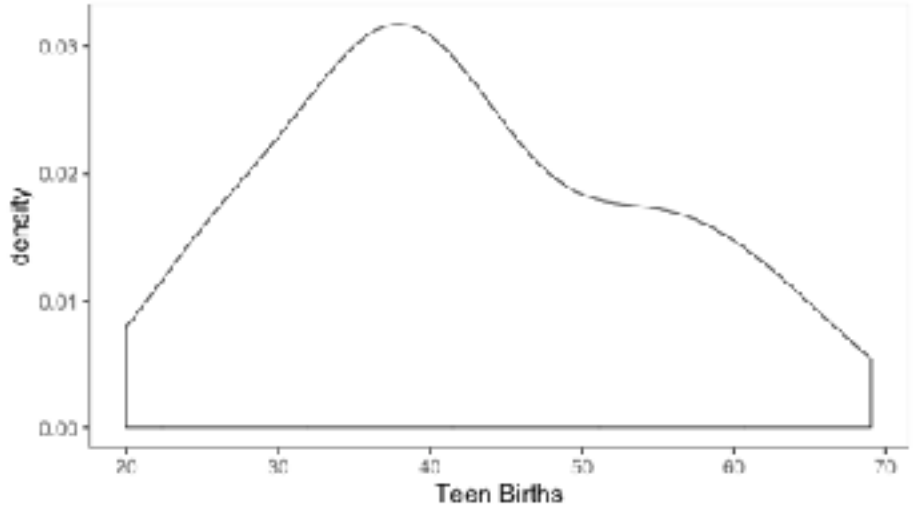
B
Positive relationship

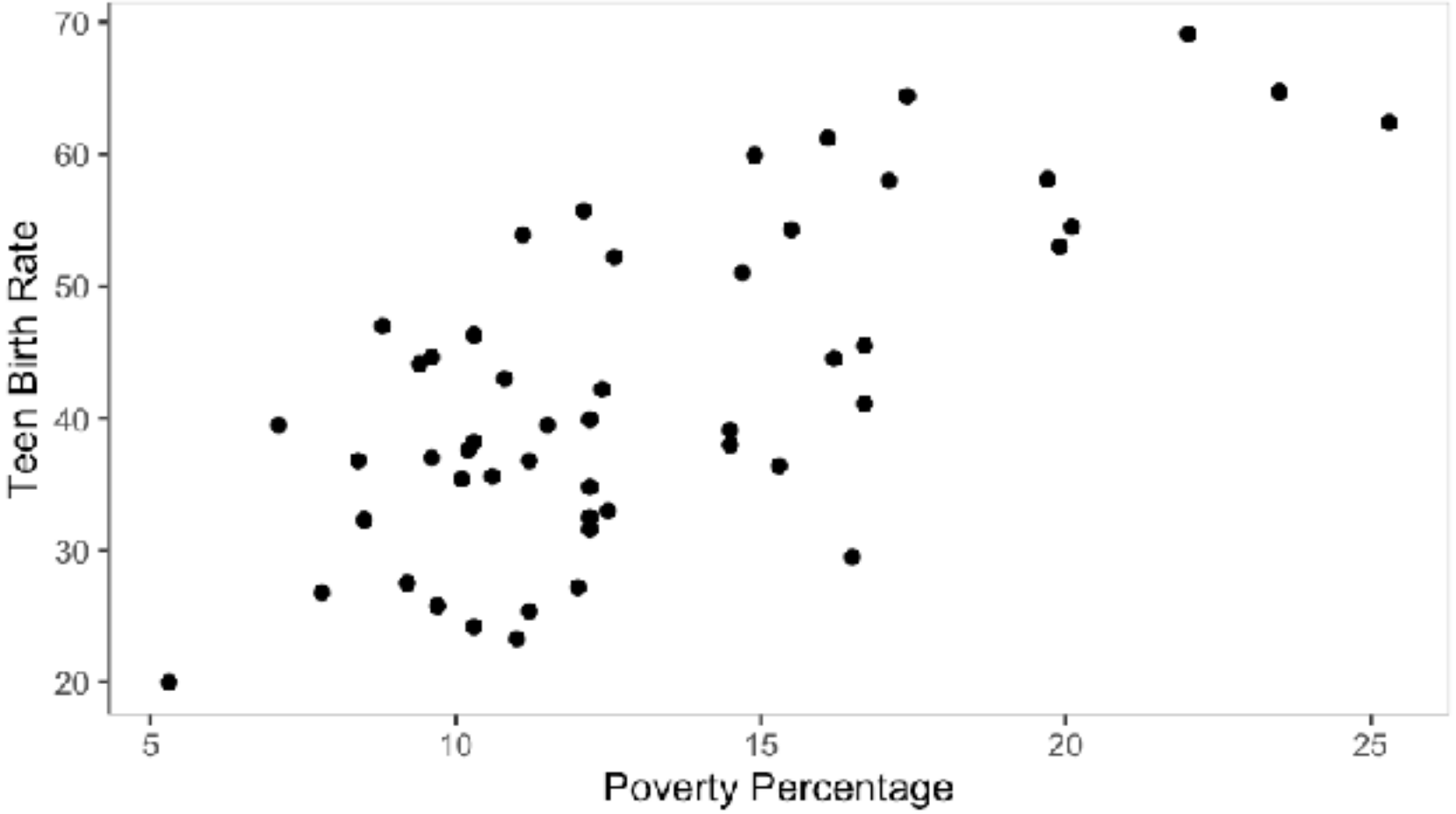
C
Negative relationship

	Location	PovPct	Brth15to17	Brth18to19	ViolCrime	TeenBrth
1	Alabama	20.1	31.5	88.7	11.2	54.5
2	Alaska	7.1	18.9	73.7	9.1	39.5
3	Arizona	16.1	35.0	102.5	10.4	61.2
4	Arkansas	14.9	31.6	101.7	10.4	59.9
5	California	16.7	22.6	69.1	11.2	41.1
6	Colorado	8.8	26.2	79.1	5.8	47.0
7	Connecticut	9.7	14.1	45.1	4.6	25.8
8	Delaware	10.3	24.7	77.8	3.5	46.3
9	District_of_Columbia	22.0	44.8	101.5	65.0	69.1
10	Florida	16.2	23.2	78.4	7.3	44.5
11	Georgia	12.1	31.4	92.8	9.5	55.7
12	Hawaii	10.3	17.7	66.4	4.7	38.2
13	Idaho	14.5	18.4	69.1	4.1	39.1
14	Illinois	12.4	23.4	70.5	10.3	42.2
15	Indiana	9.6	22.6	78.5	8.0	44.6
16	Iowa	12.2	16.4	55.4	1.8	32.5
17	Kansas	10.8	21.4	74.2	6.2	43.0

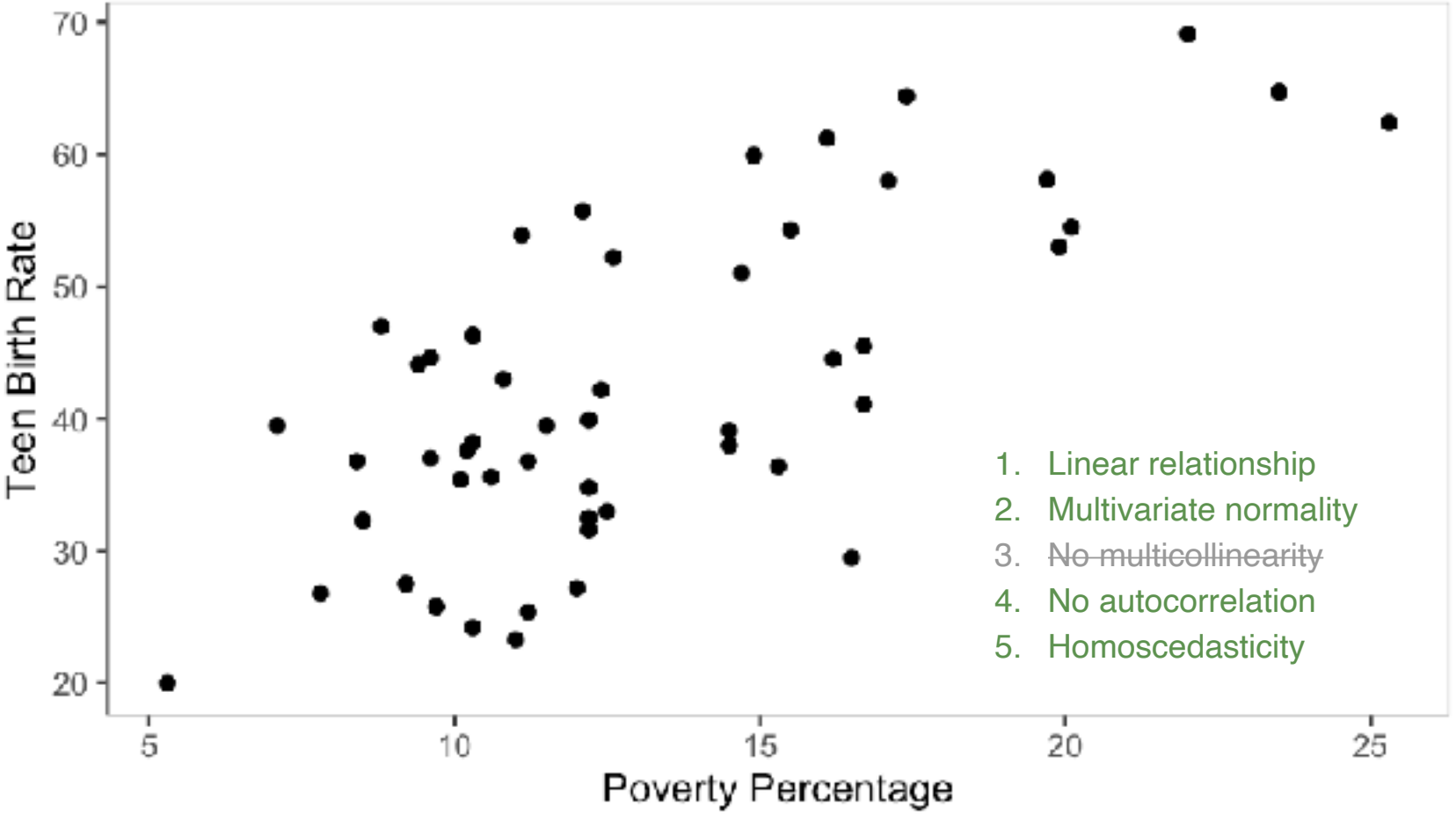
Normal(ish) distributions



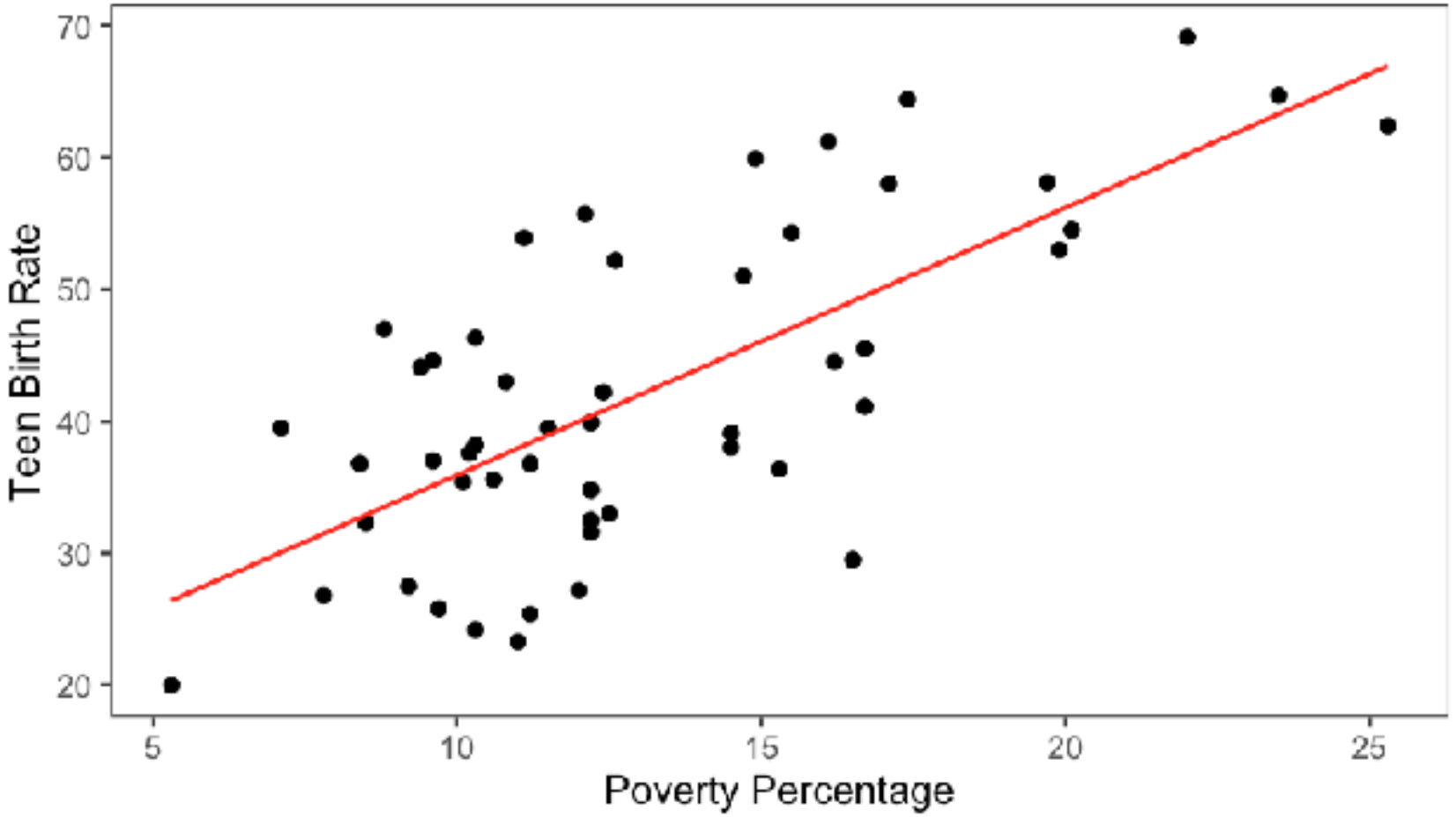


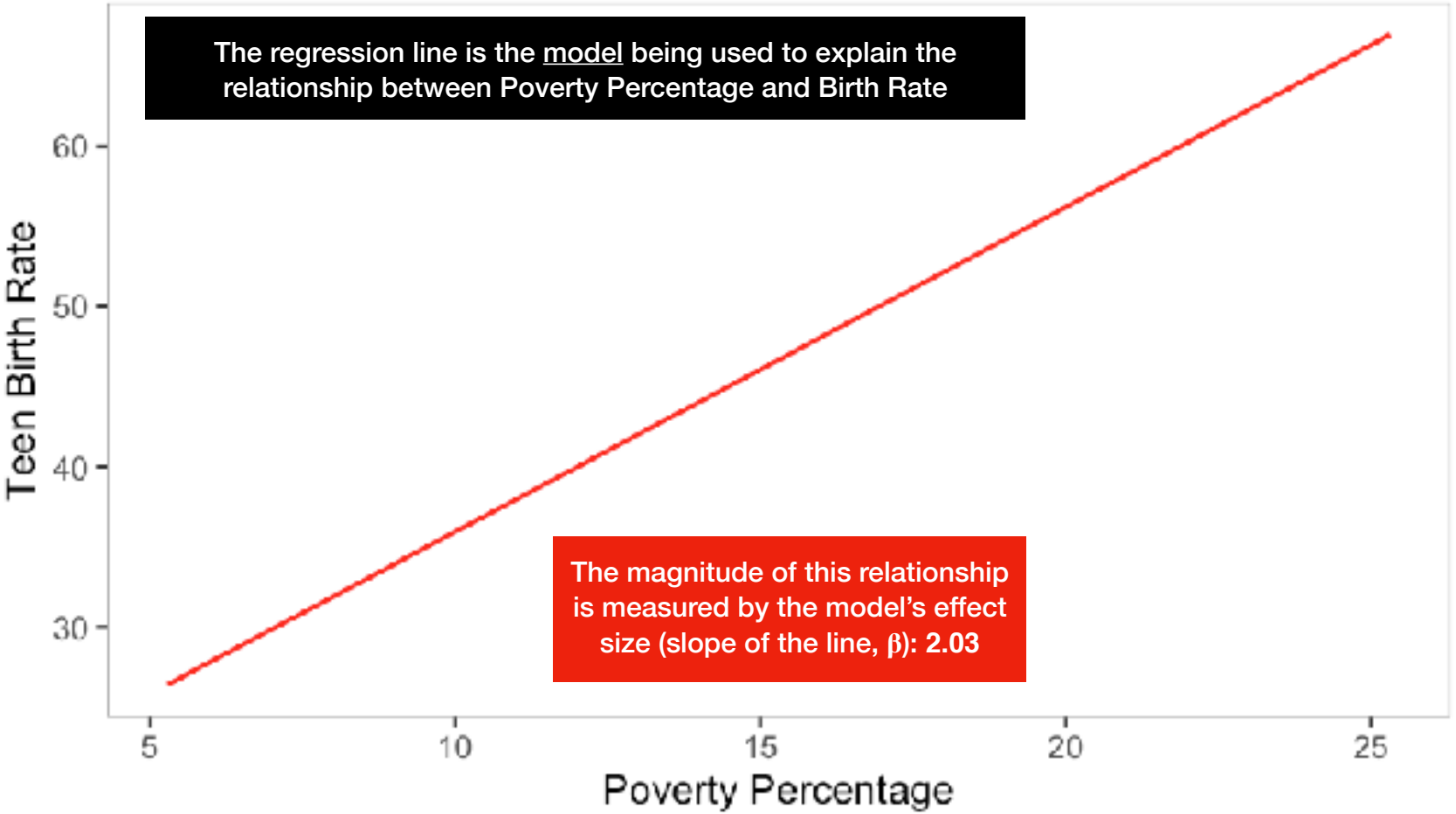


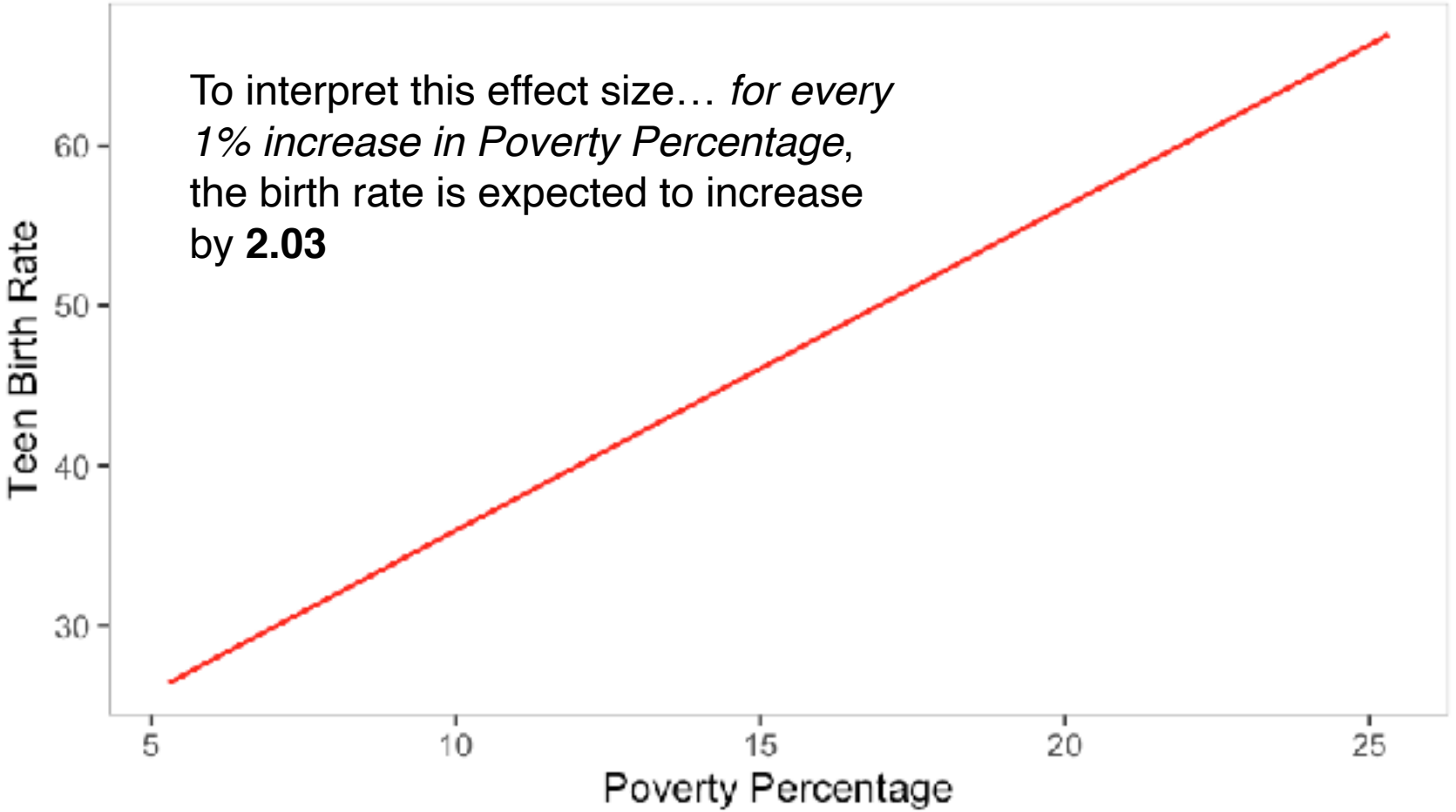
Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.



Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.

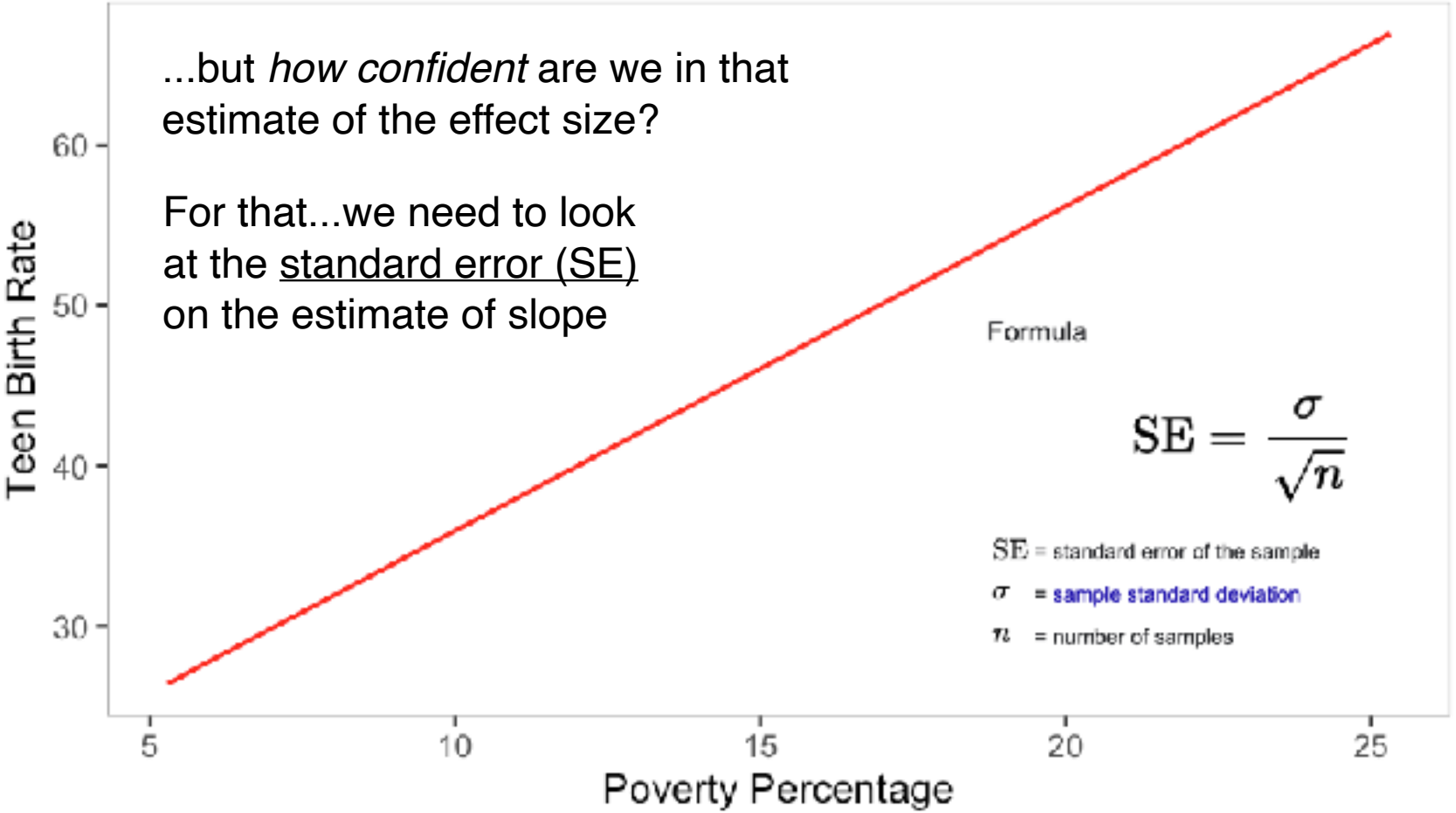

Teen Birth Rate
70
60
50
40
30
20
5
10
15
20
25
Poverty Percentage

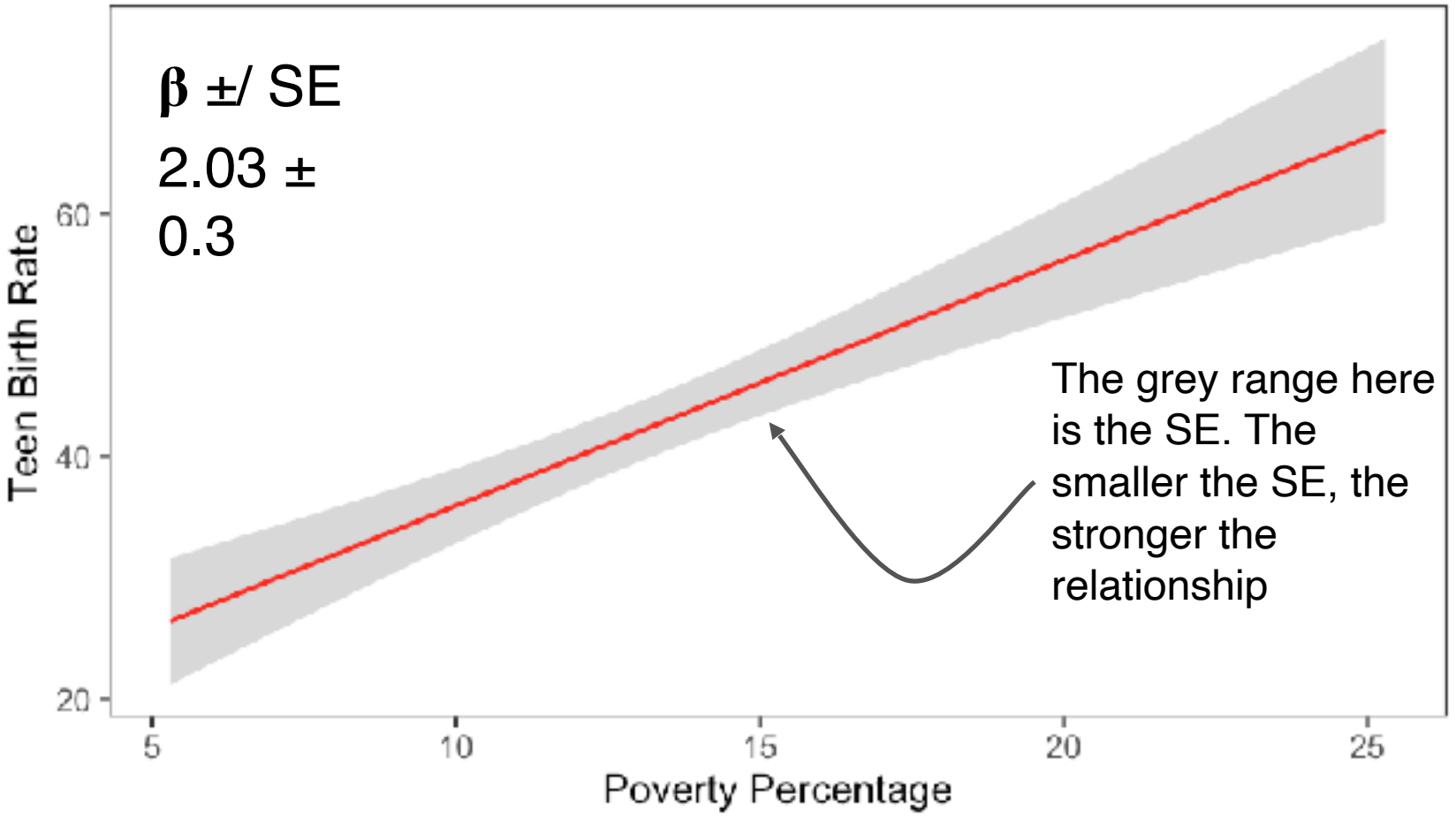

The regression line is the model being used to explain the relationship between Poverty Percentage and Birth Rate
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage
The magnitude of this relationship is measured by the model's effect size (slope of the line, β): 2.03

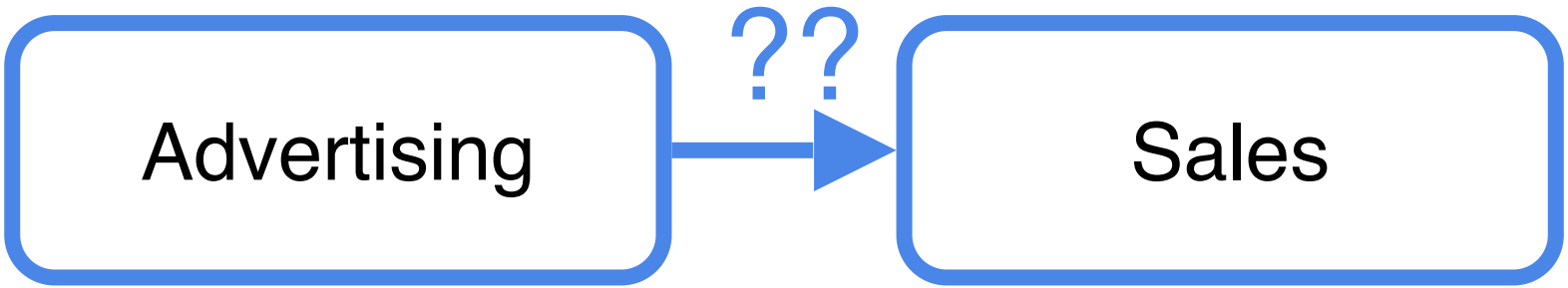

To interpret this effect size… *for every 1% increase in Poverty Percentage*, the birth rate is expected to increase by **2.03**
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage

...but *how confident* are we in that estimate of the effect size?

For that...we need to look at the standard error (SE) on the estimate of slope

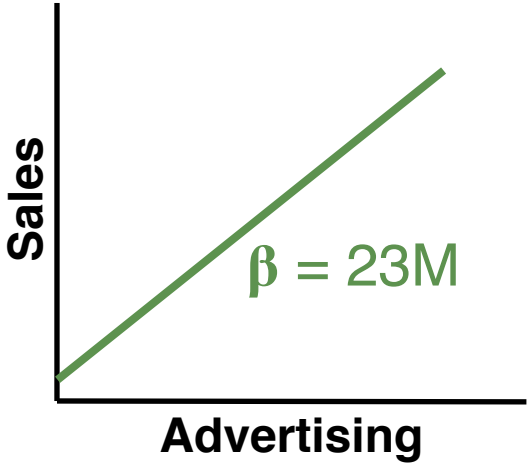



β ±/ SE
2.03 ±
0.3
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage
The grey range here is the SE. The smaller the SE, the stronger the relationship


Advertising
??
Sales

Effect size interpretation

Sales (Million Euro)	Advertising (Million Euro)
651	23
762	26
856	30
1,063	34
1,190	43
1,298	48
1,421	52
1,440	57
1,518	58



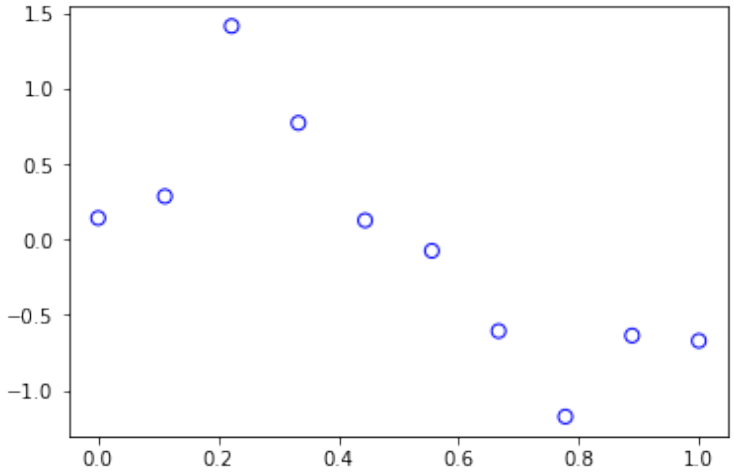
The effect size (β) between the advertising and sales is 23M. What does this mean?

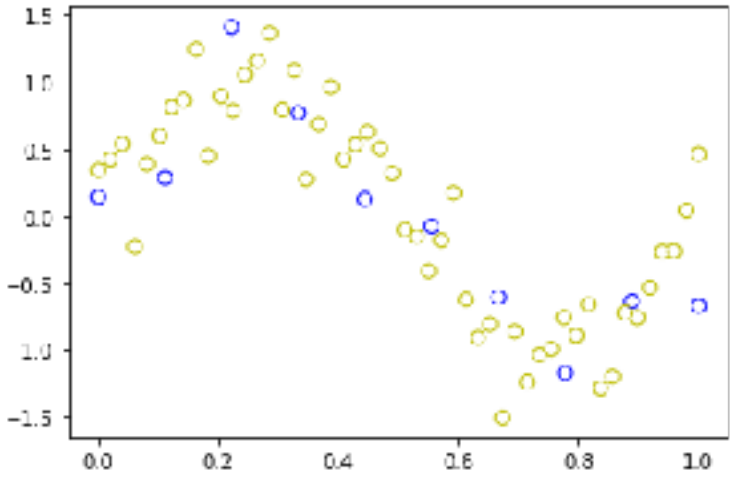
A
For every 1M Euro spent on advertising, the company sees 23M more in sales

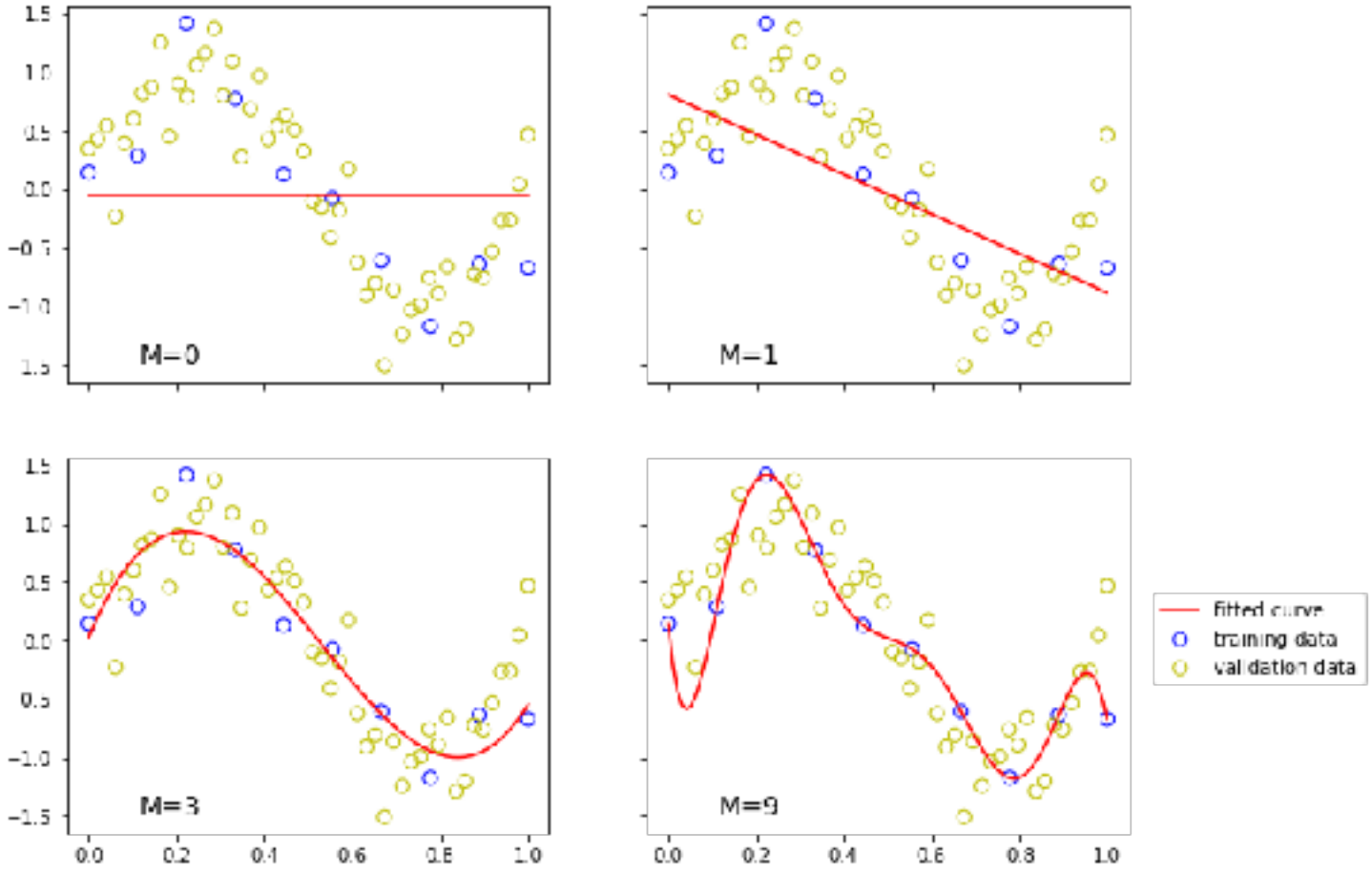
B
For every 1M Euro spent in sales, the company spends 23M more in advertising

C
For every 1M Euro spent on advertising, the company sees 24M less in sales

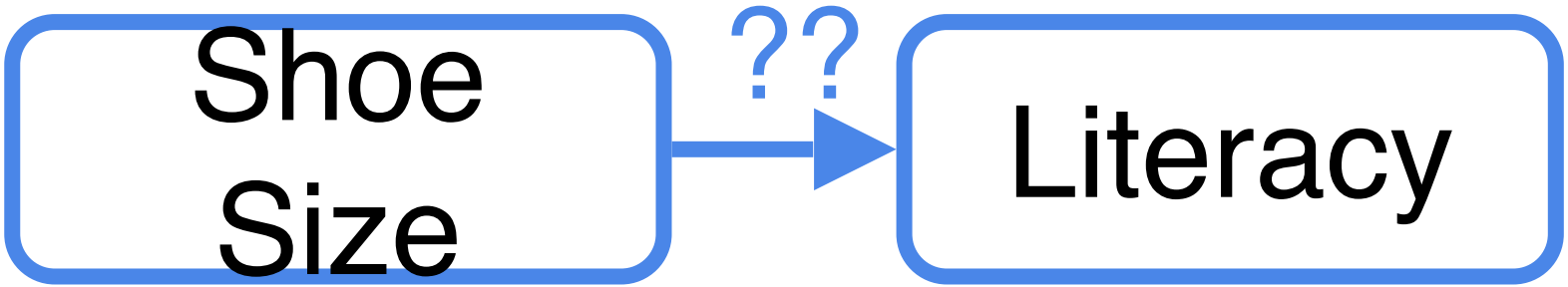
D
For every 1M Euro spent in sales, the company spends 23M less in advertising

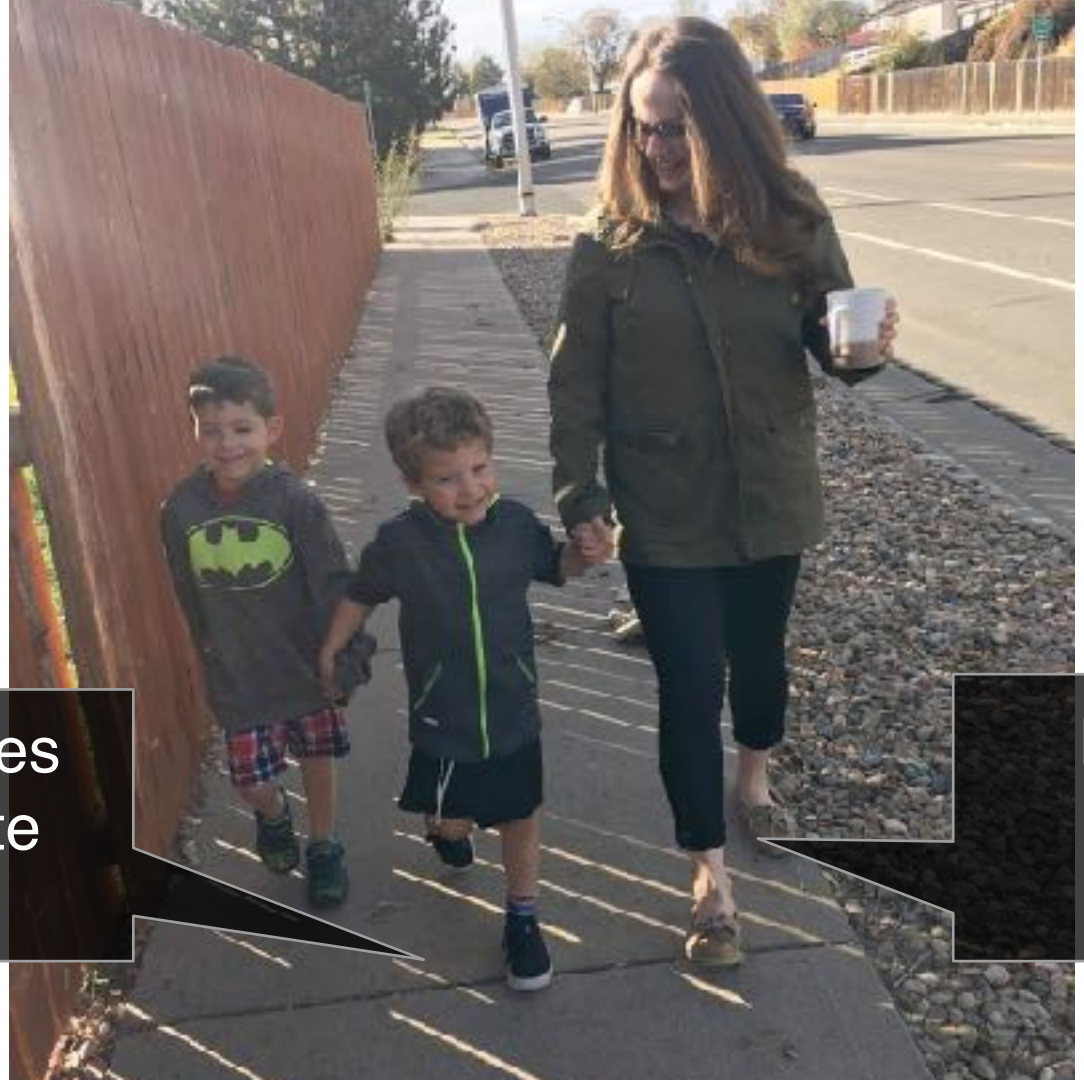

1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

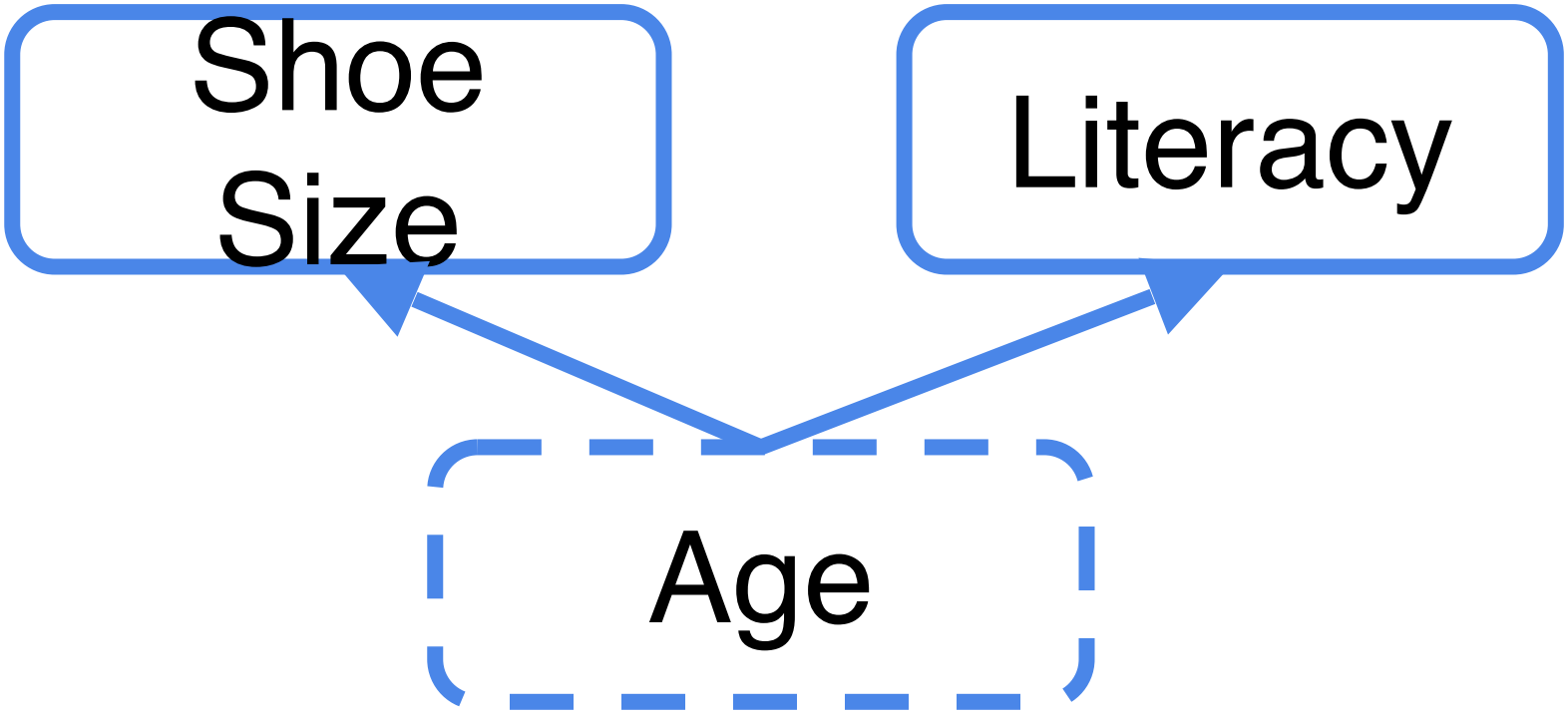

1.5
1.0
0.5
0.0
−0.5
1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

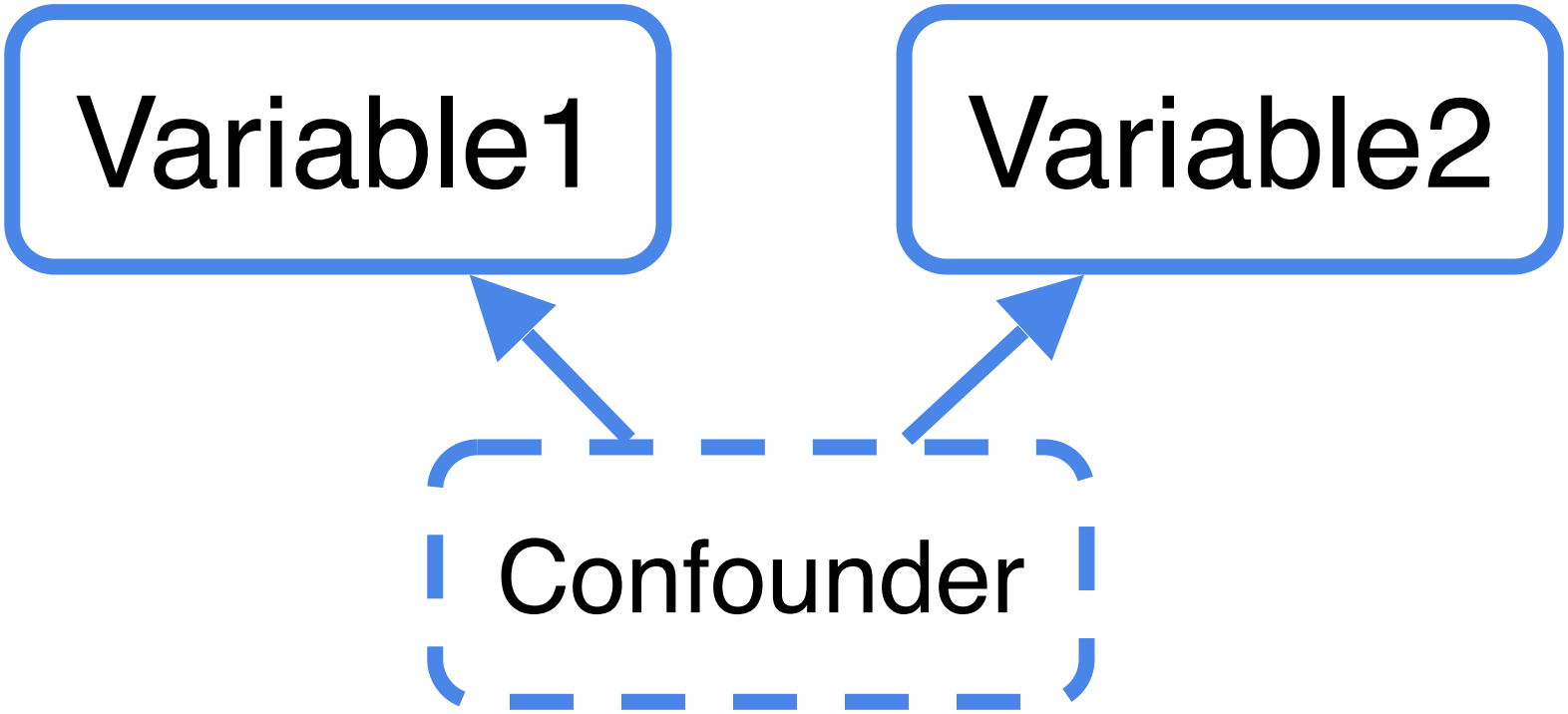

M=0
M=1
M=3
M=9
fitted curve
training data
validation data

Confounding


Shoe Size
??
Literacy


Small shoes
Not literate
Child
Big shoes
Literate
Adult

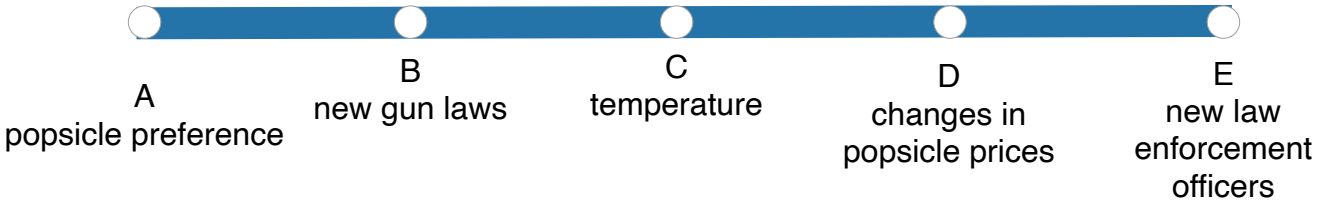

Shoe Size
Literacy
Age


Variable1
Variable2
Confounder

Confounding

popsicles → crime rate

Your analysis sees an increase in crime rate whenever popsicle sales increase. What could confound this analysis?



You can plan ahead to avoid confounding and/or include confounders in your models to account for their role on the outcome variable.

Ignoring confounders will lead you to draw incorrect conclusions

Spine Surgery Results

Sample: 400 patients with index vertebral fractures

…looks like vertebroplasty was *way* worse for patients!

Vertebroplasty	Conservative care	Relative risk (95% confidence interval)
30/200 (15%)	15/200 (7.5%)	2.0 (1.1–3.6)

subsequent fractures

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

But wait...at time of initial fracture...

	Vertebroplasty N = 200	Conservative care N = 200
Age, y, mean ± SD	78.2 ± 4.1	79.0 ± 5.2
Weight, kg, mean ± SD	54.4 ± 2.3	53.9 ± 2.1
Smoking status, No. (%)	110 (55)	16 (8)

Age and weight are similar between groups. **Smoking Status** differs vastly.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

So...let's stratify those results real quick

Smoke			No smoke		
Vertebroplasty	Conservative	RR (95% confidence interval)	Vertebroplasty	Conservative	RR (95% confidence interval)
23/110 (21%)	3/16 (19%)	1.1 (0.4, 3.3)	7/90 (8%)	12/184(7%)	1.2 (0.5, 2.9)

Risk of re-fracture is now similar within group

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/