

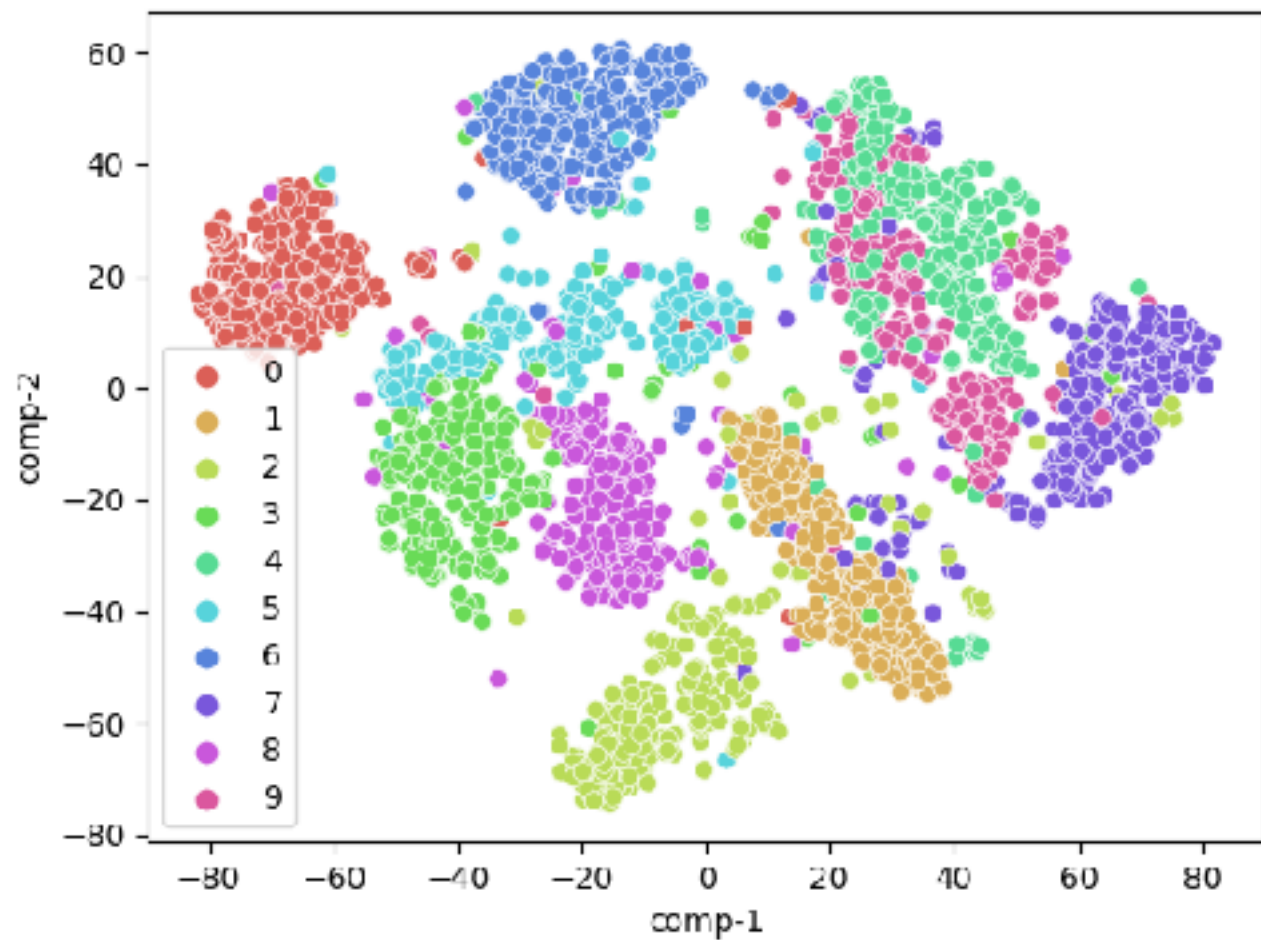
Course Announcements

- Due Friday (11:59 PM)
 - D4
 - A2

Grading underway: Project Proposals



MNIST data T-SNE projection



1. Hyperparameters really matter
2. Cluster sizes in a UMAP plot mean nothing
3. Distances between clusters might not mean anything
4. Random noise doesn't always look random.
5. You may need more than one plot

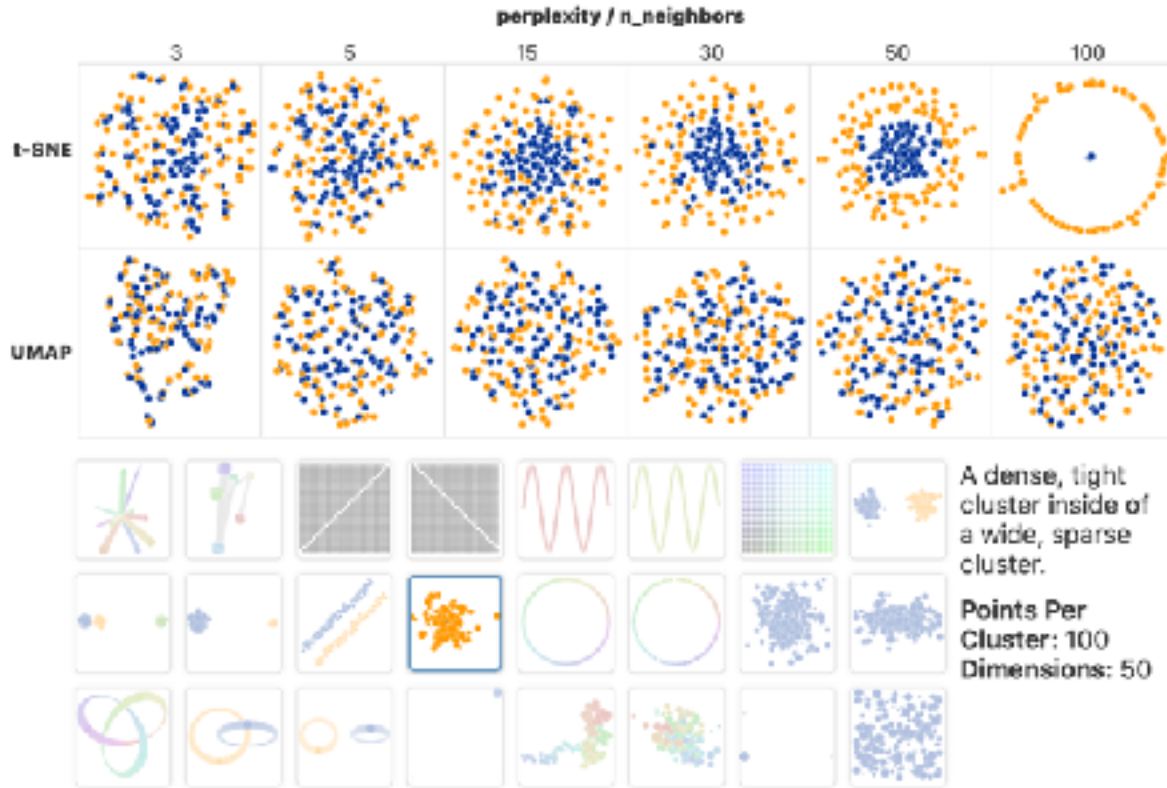


Figure 7: Comparison between UMAP and t-SNE projecting various toy datasets.

Inferential analysis

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

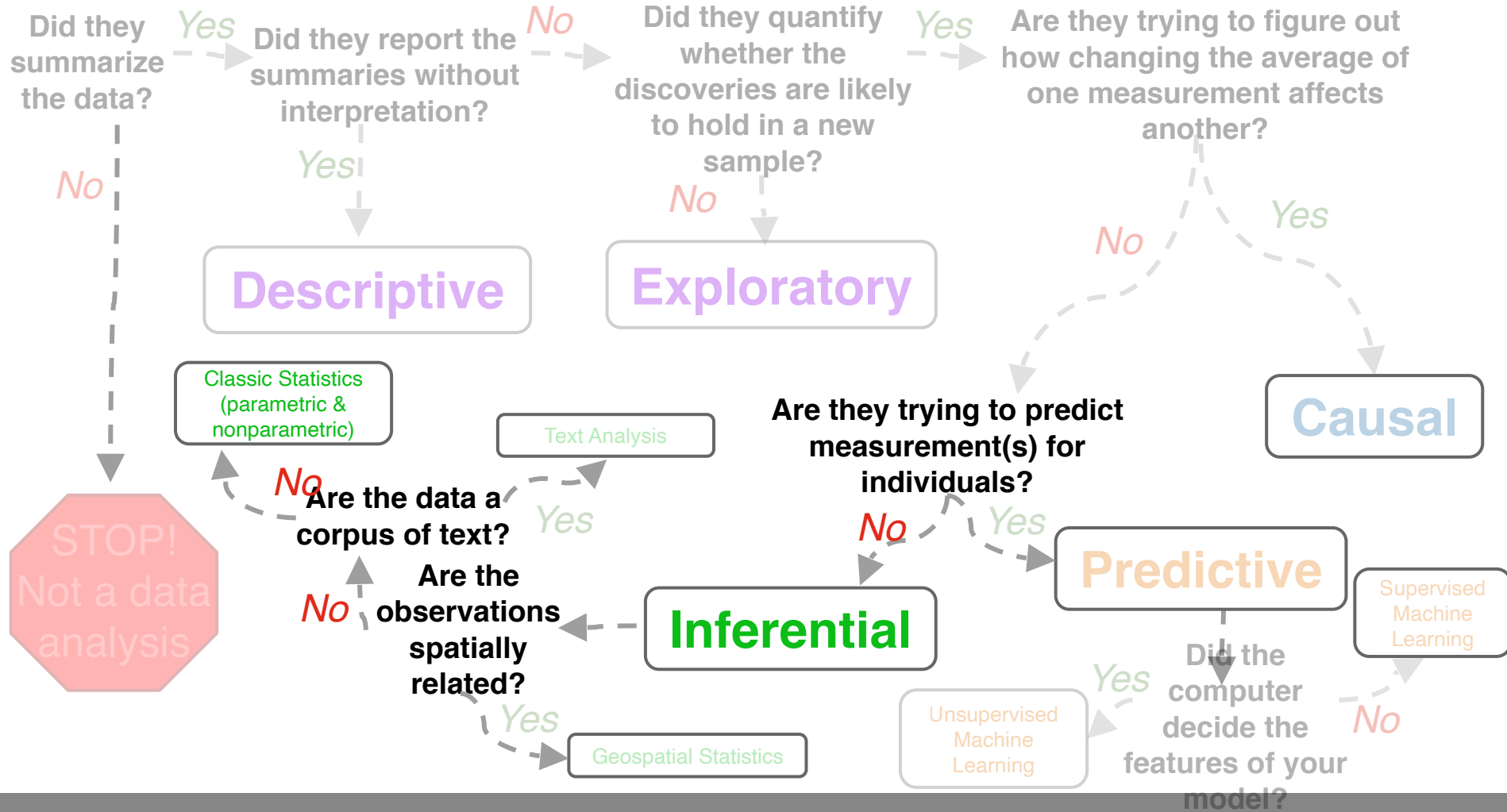
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>



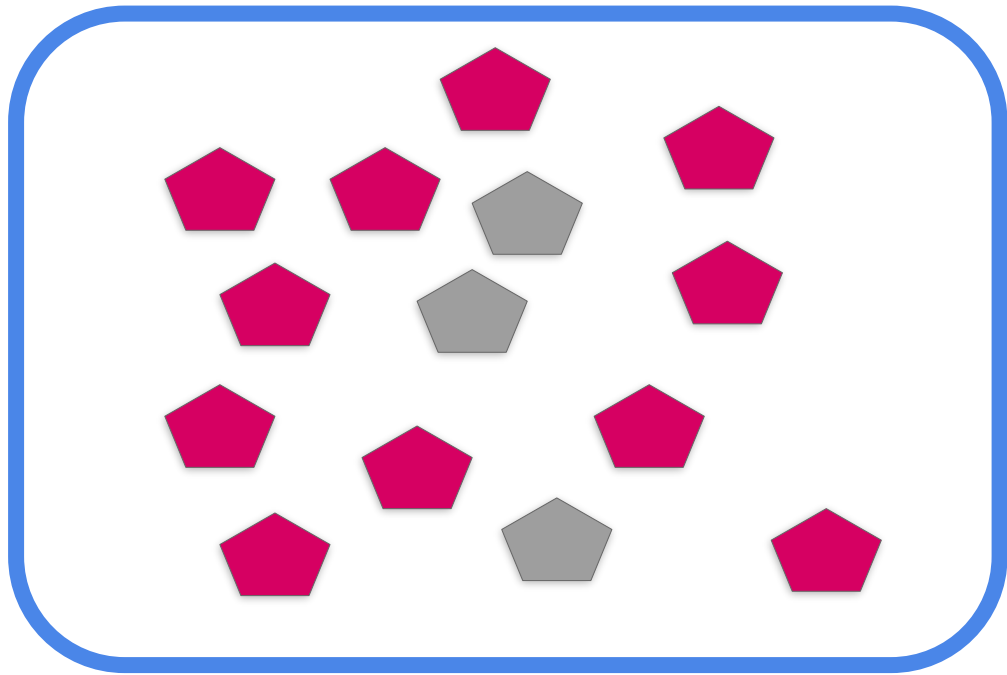
- **Problem:** Does Sesame Street affect kids brain development?
- **Data science question:** Is there a relationship between watching Sesame Street and test scores among children?
- **Type of analysis:** Inferential analysis



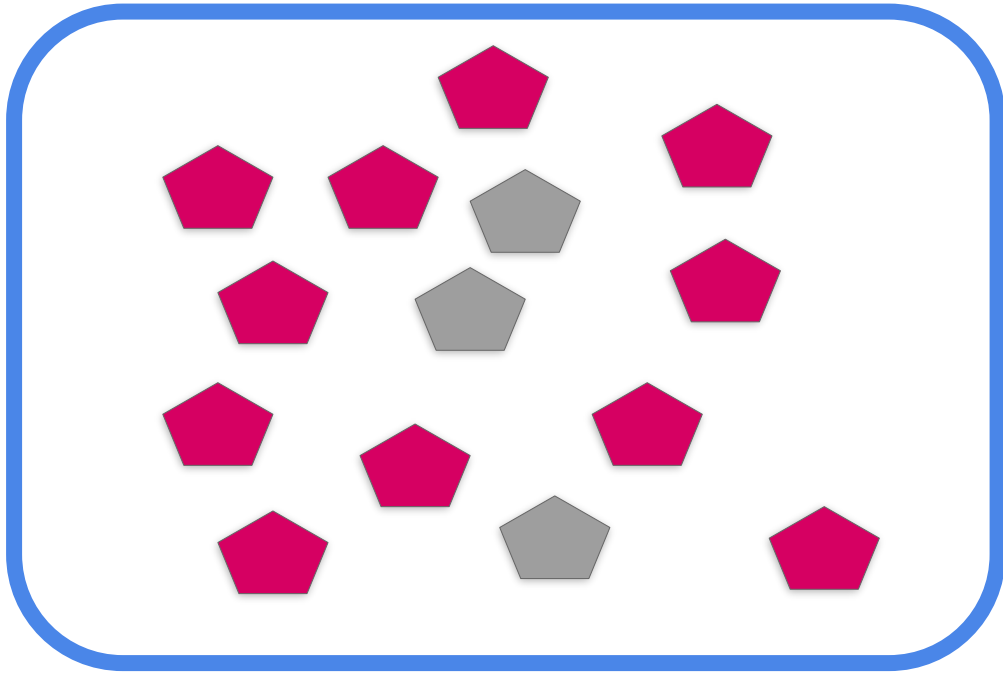
Sesame Street
viewership

??

Test scores



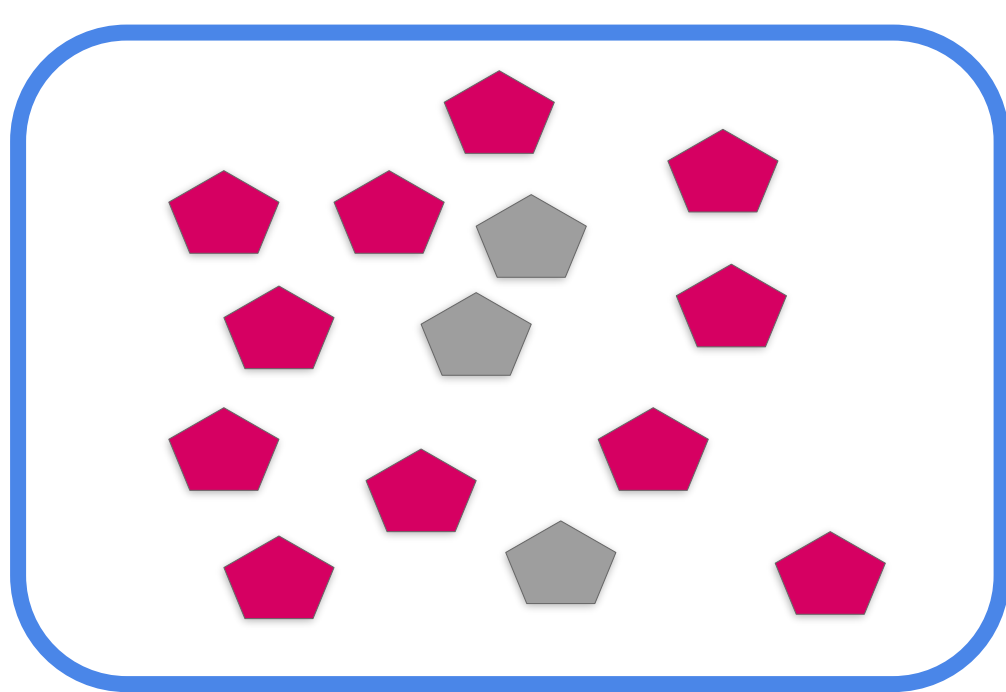
Population



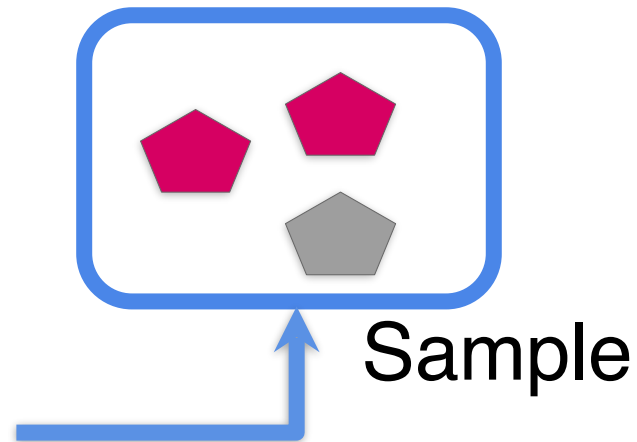
Population



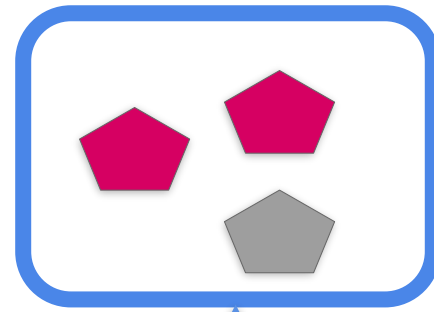
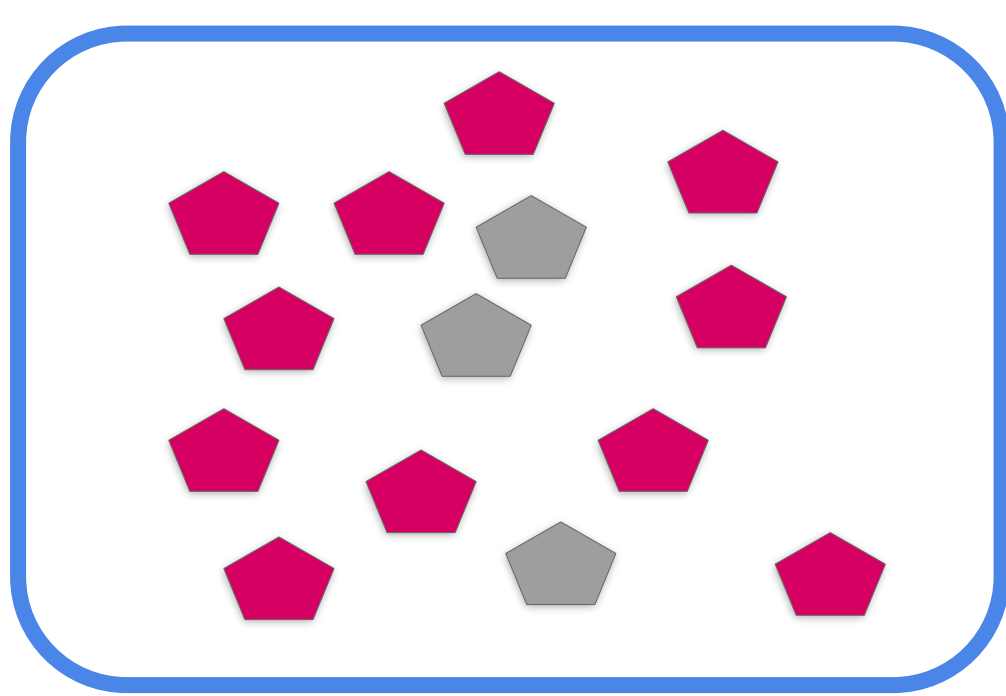
In our Sesame street example, the population would be all children



Population



Sample

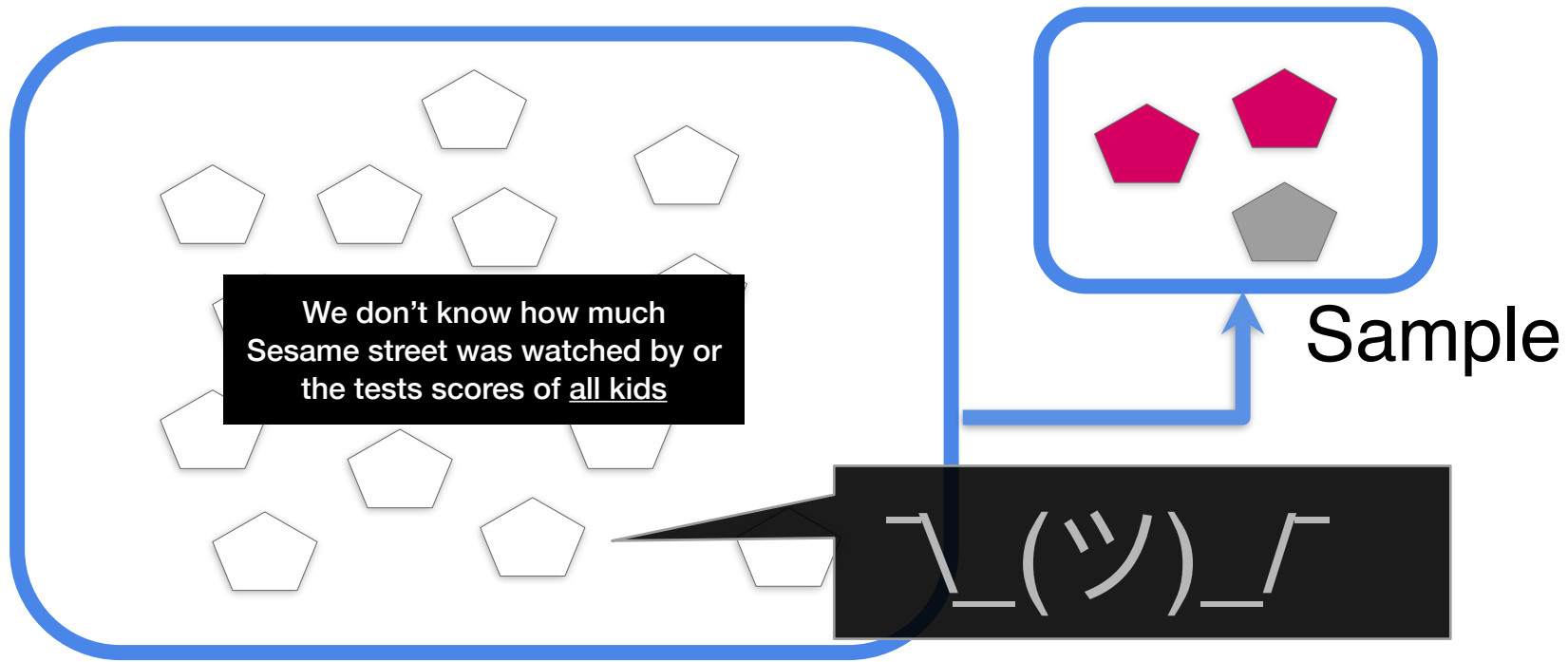


Sample

Population



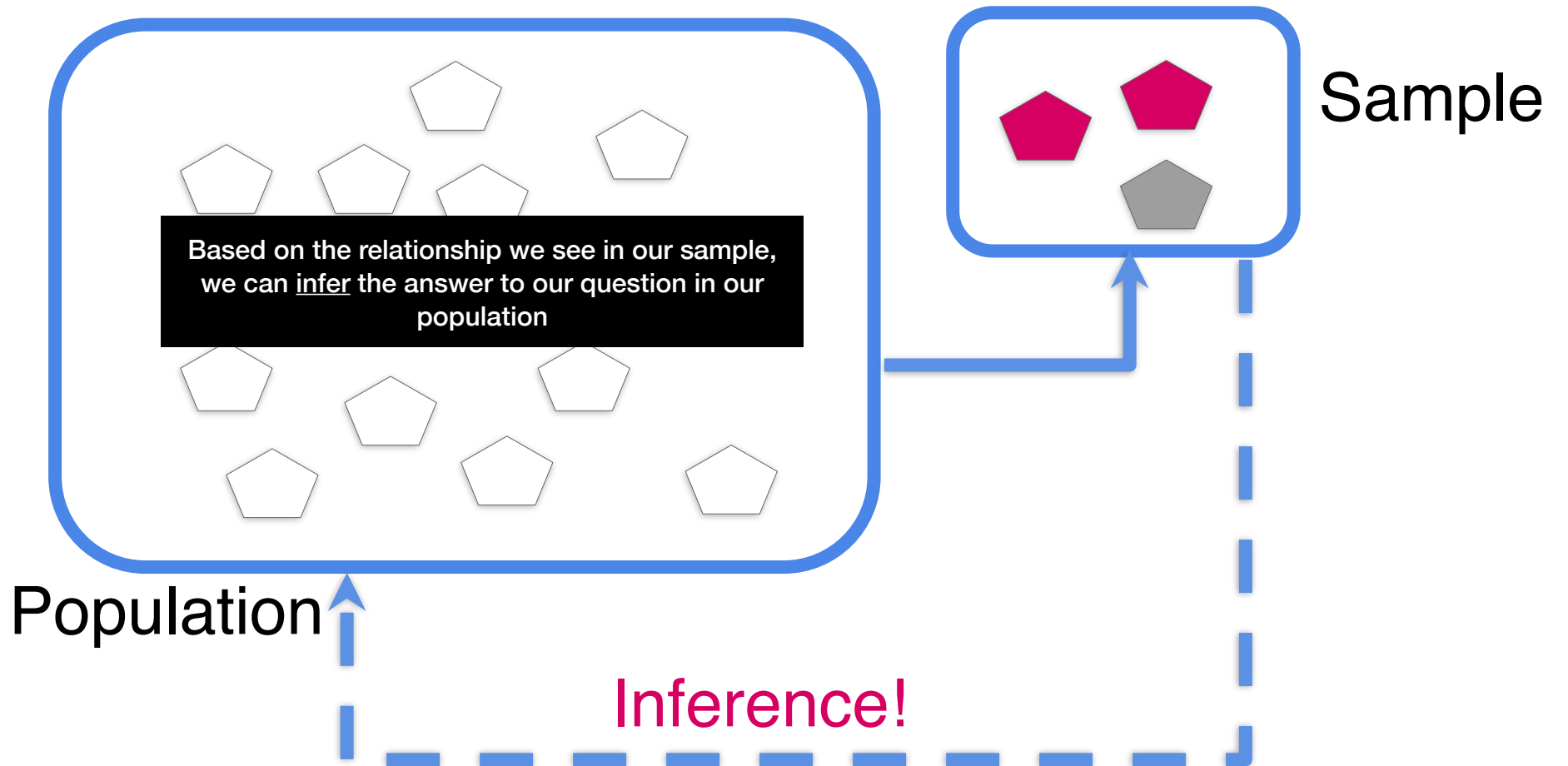
In our Sesame street example, the population would be all children



Population



In our Sesame street example, the population would be all children

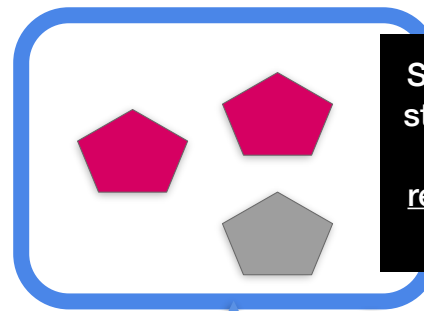
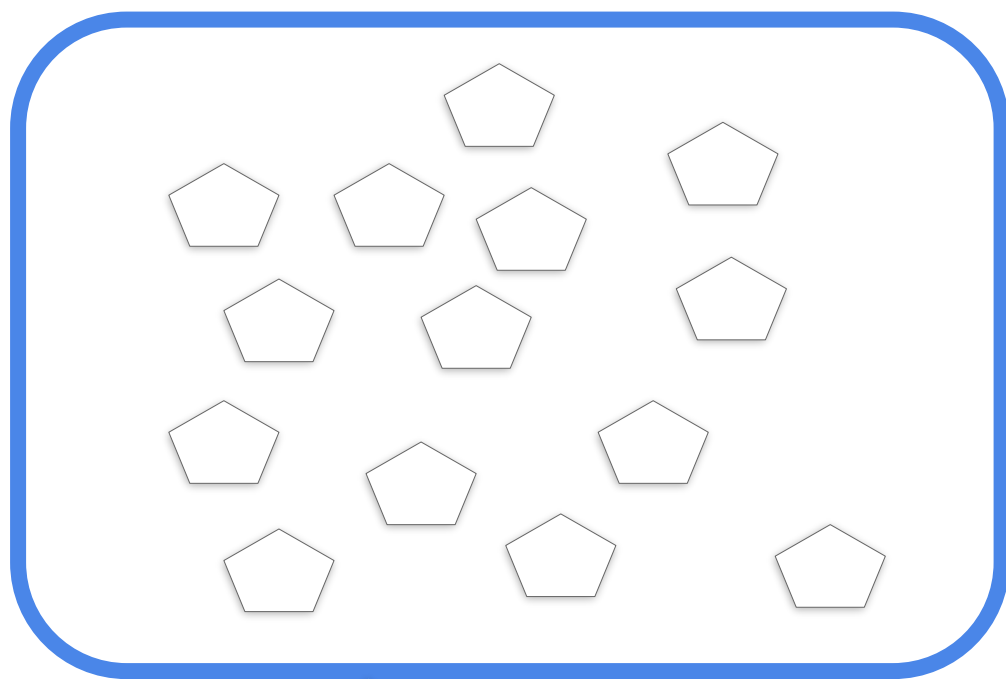


Based on the relationship we see in our sample,
we can infer the answer to our question in our
population

Sample

Population

Inference!



So we look at Sesame
street viewing and test
scores in a
representative sample
of kids



Could be this

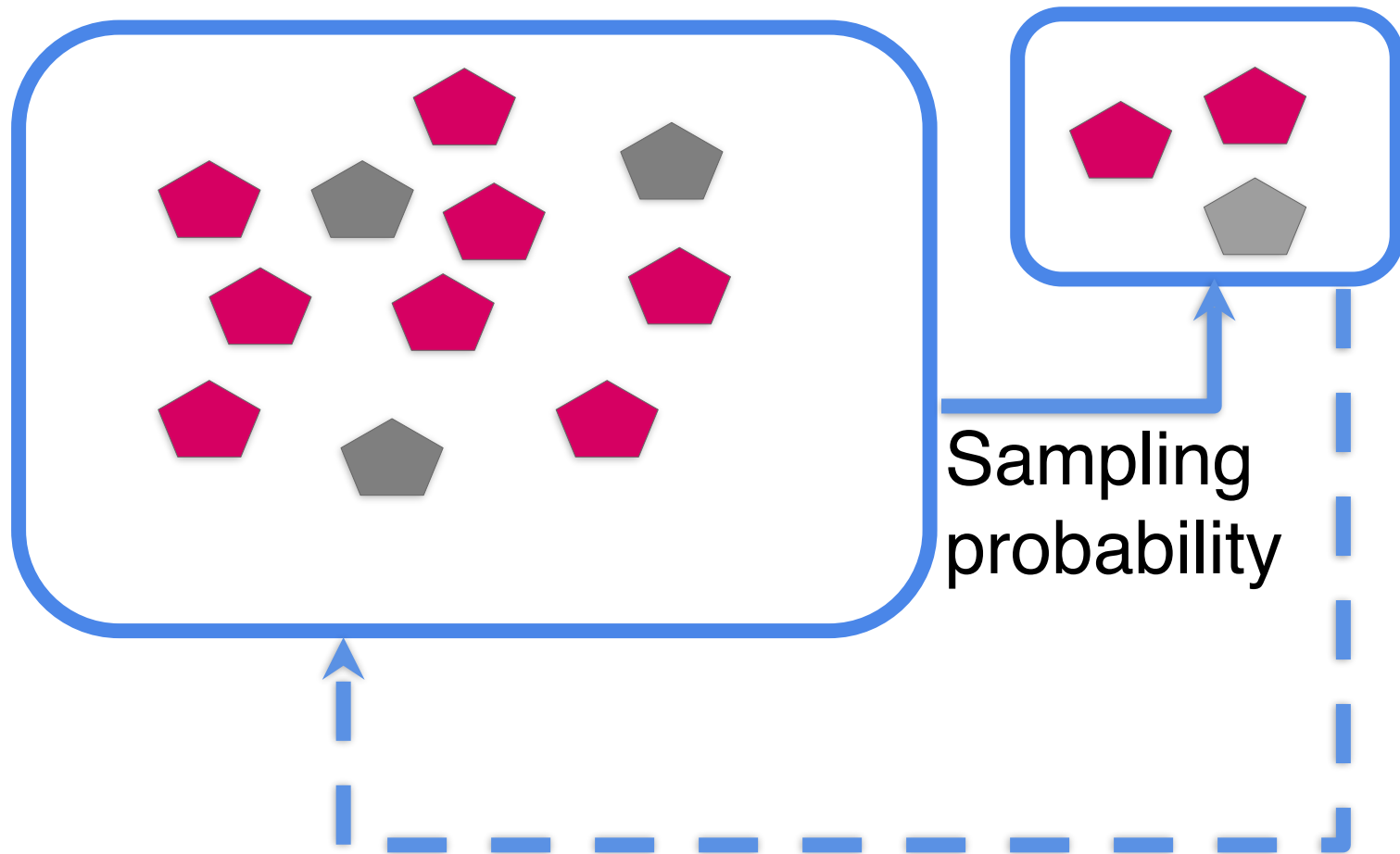
So we look at Sesame
street viewing and test
scores in a
representative sample
of kids

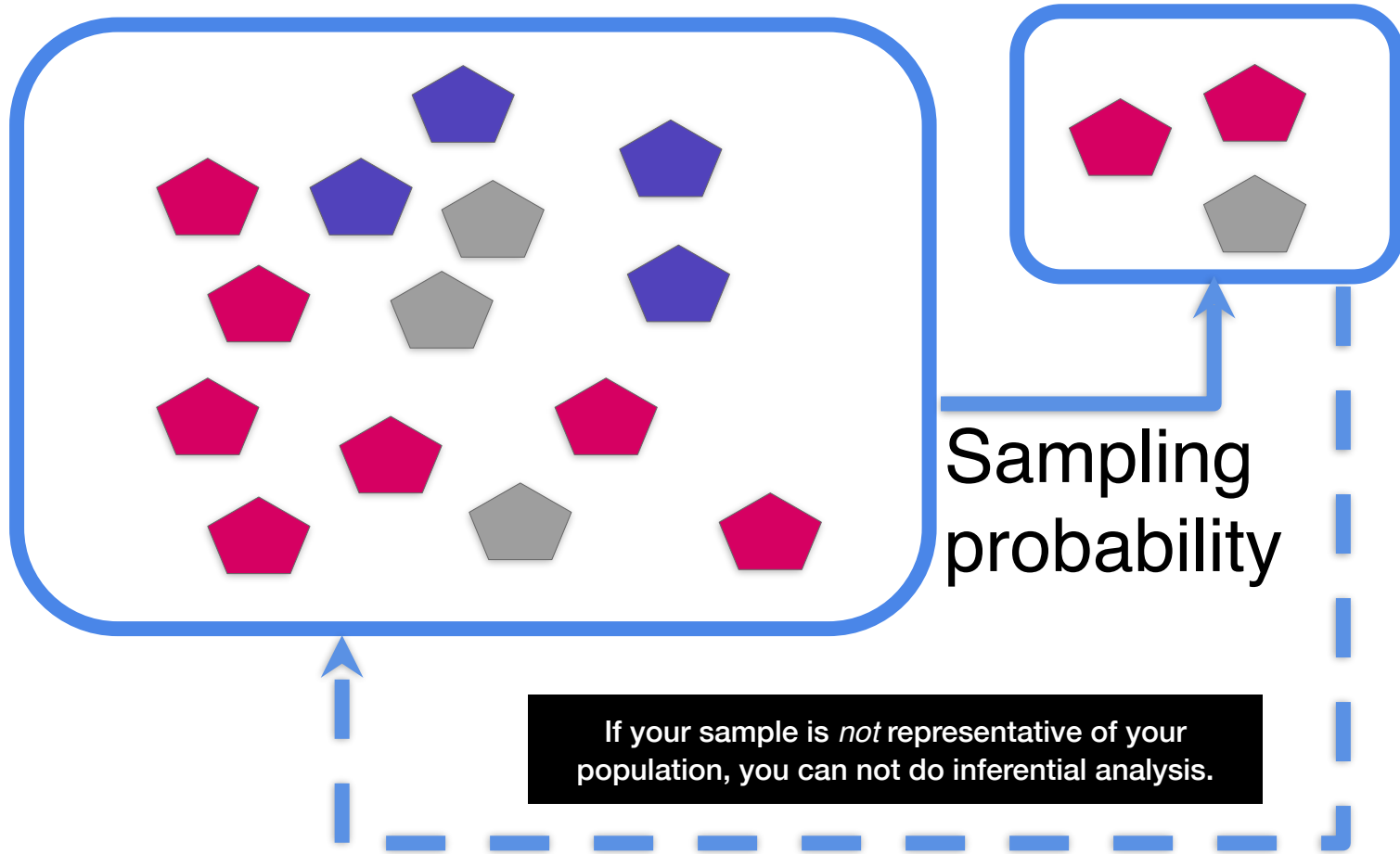


... or this

So we look at Sesame
street viewing and test
scores in a
representative sample
of kids









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. Cornea,

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

G. 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

Air pollution
control

??

Lifespan

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?

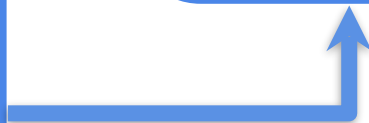


The population:
every individual in the USA

The population:
every individual in the USA



The sample:
545 US counties



Random Sampler

“Girl Scouts were randomly sampled from across the United States”

What would be the likely population given this sample?

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

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

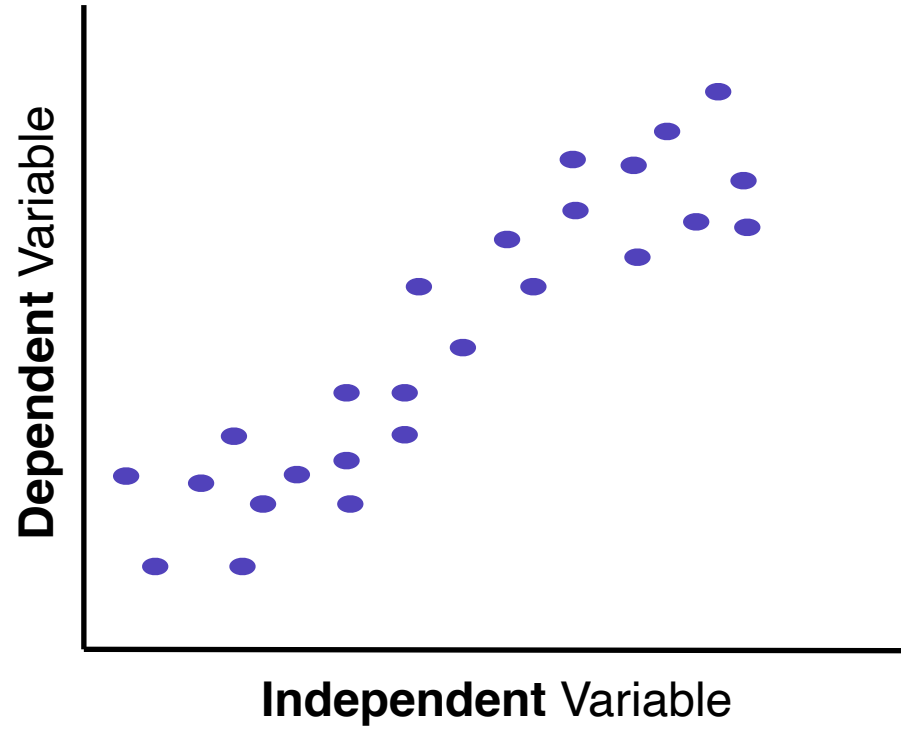
DOES CHANGE IN ONE
VARIABLE MEAN CHANGE
IN ANOTHER?

I.e. simple
regression, multiple
regression

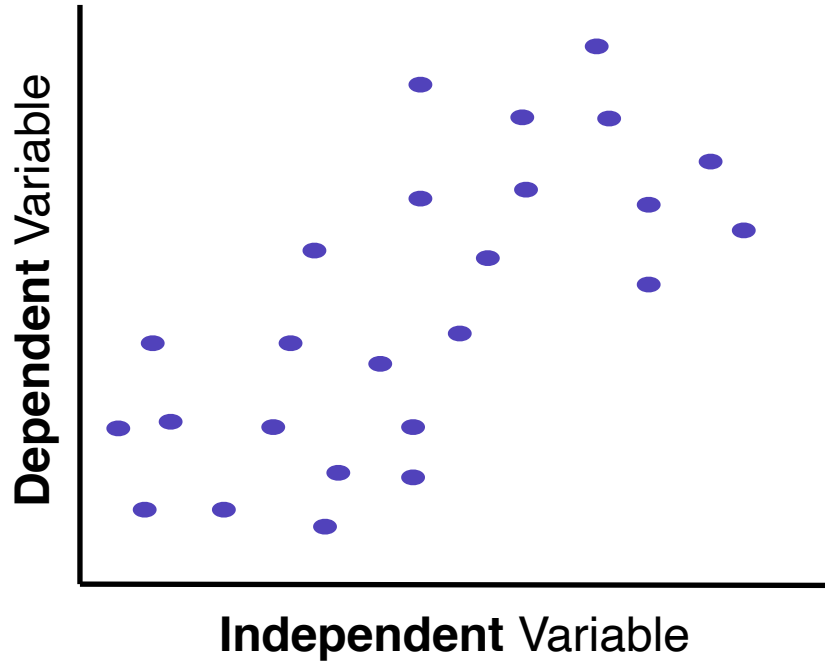
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS
IN THESE OTHER 3
CATEGORIES ARE NOT
MET

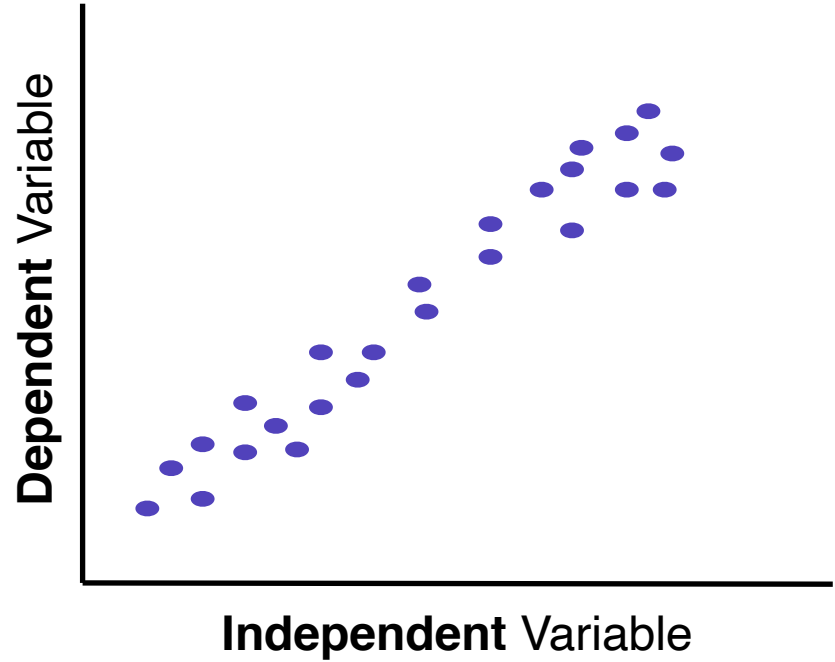
i.e. Wilcoxon rank-
sum test, Wilcoxon
sign-rank test, sign
test



weaker relationship

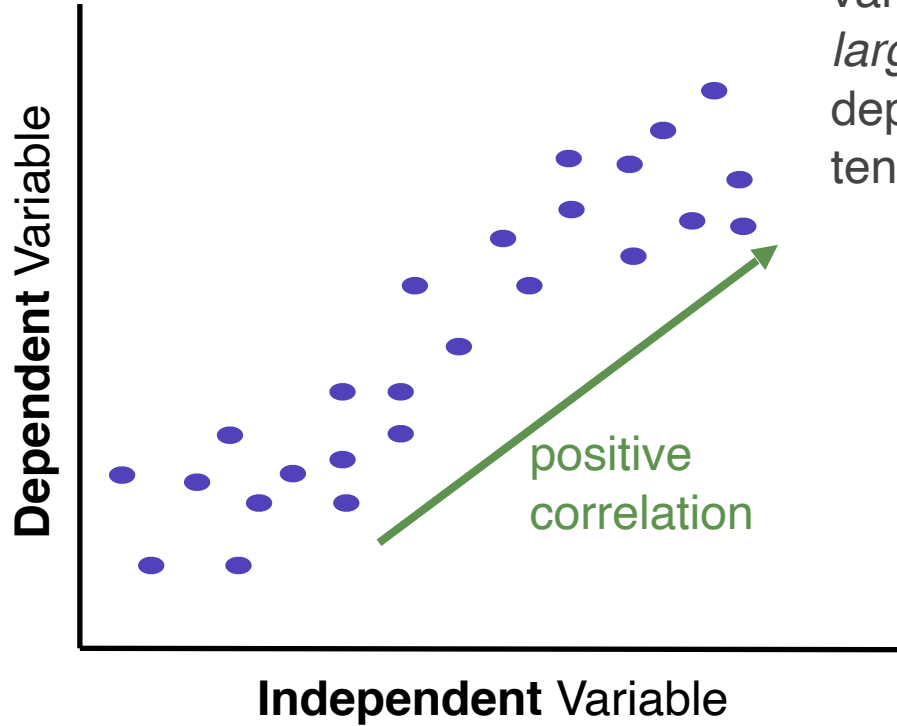


stronger relationship



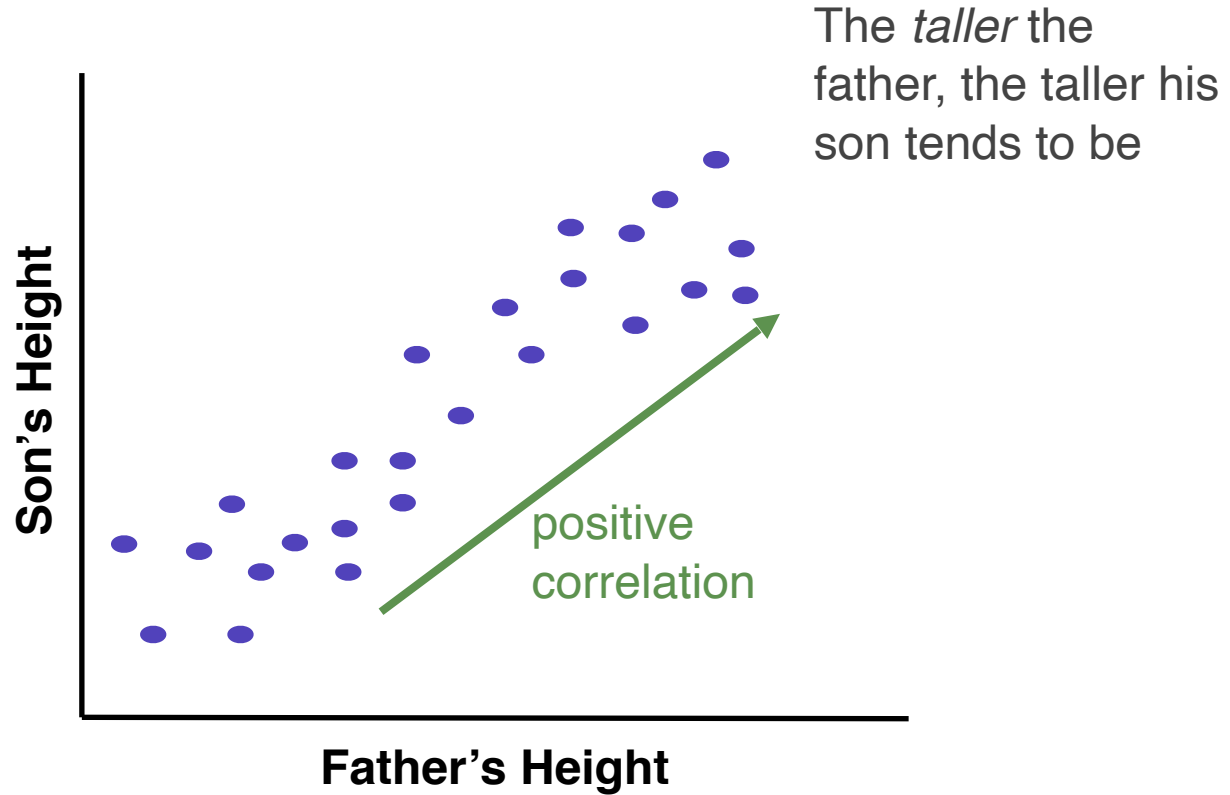
stronger relationship = higher correlation

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

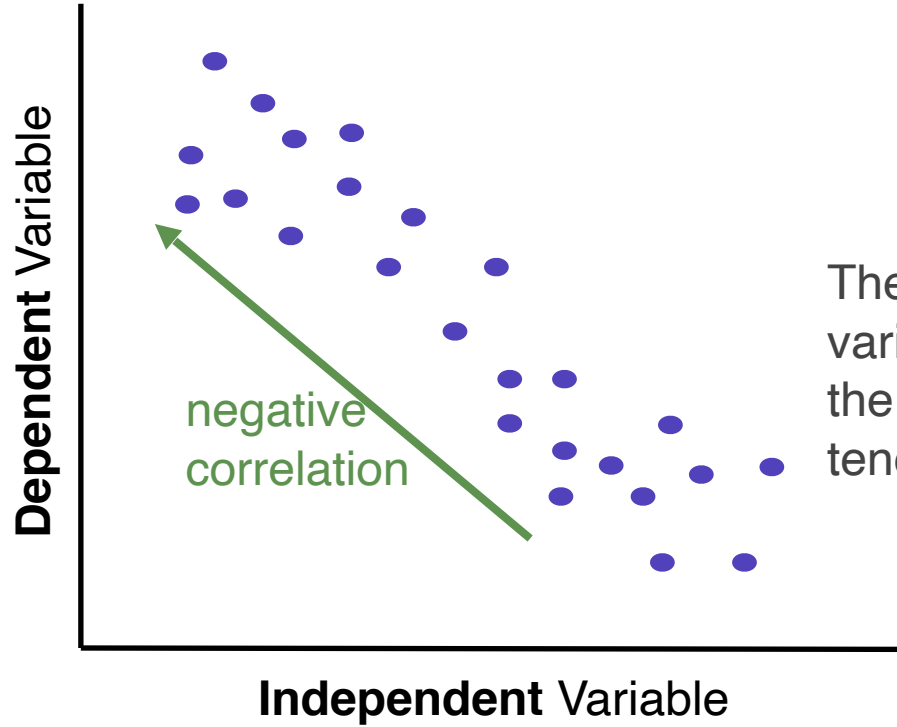


The *larger* the independent variable value, the *larger* the dependent variable tends to be

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

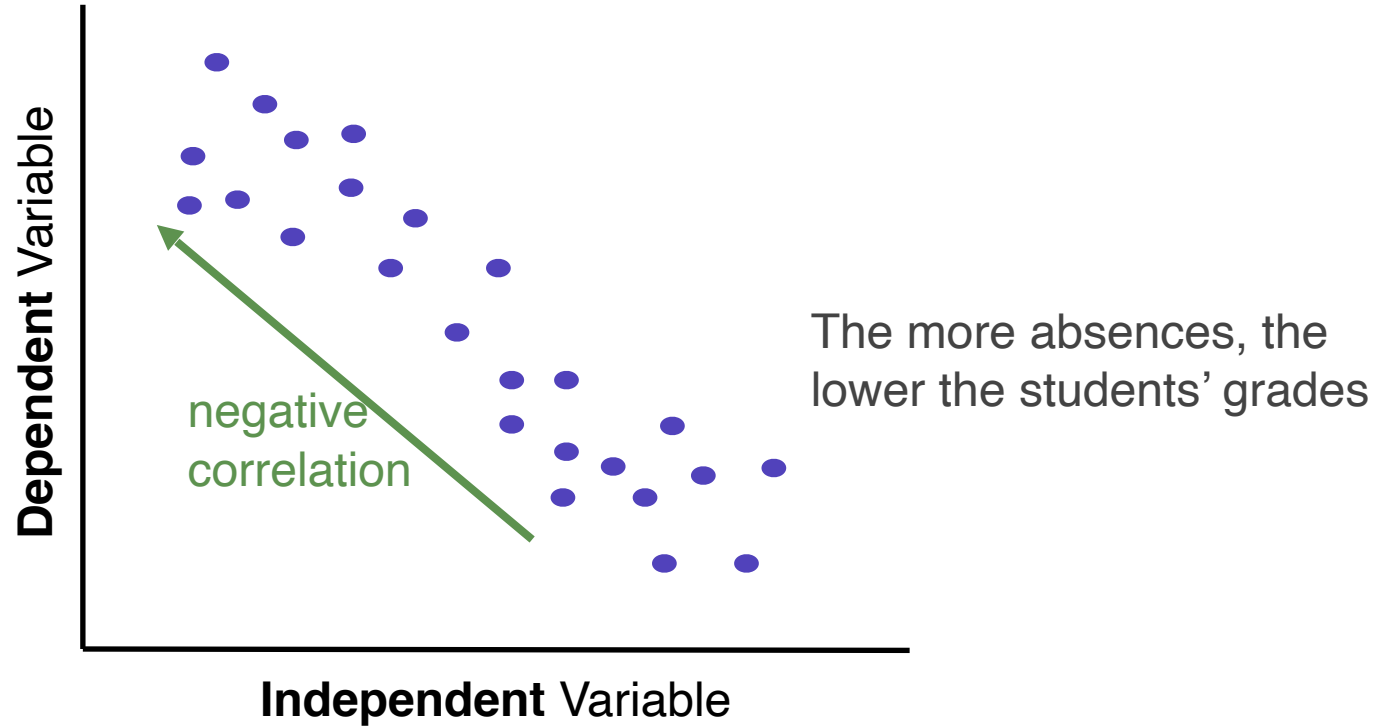


The *smaller* the independent variable value, the *larger* the dependent variable tends to be



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



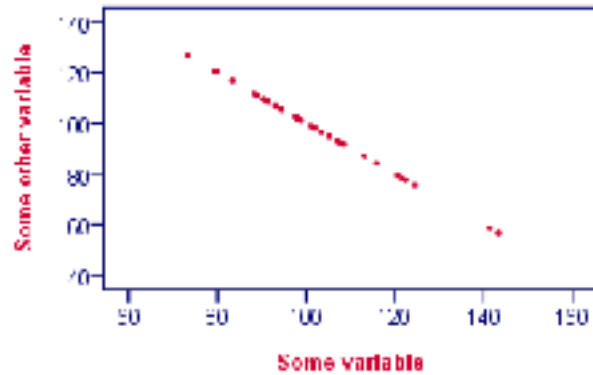
Pearson's r :

linear correlation between two variables

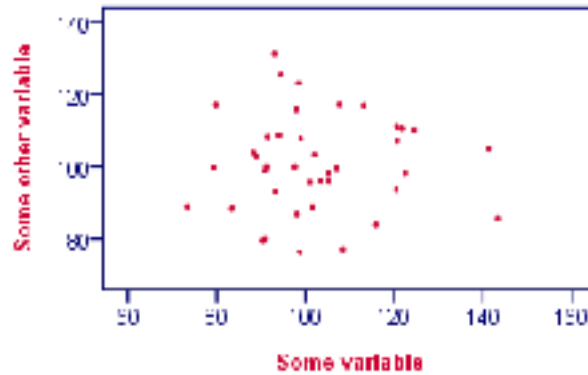
takes values $[-1, 1]$

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

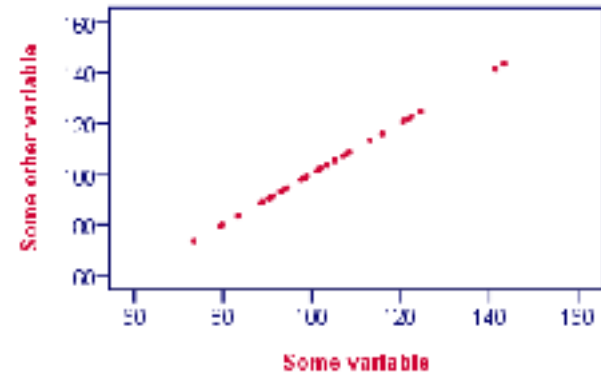
Correlation Coefficient = -1

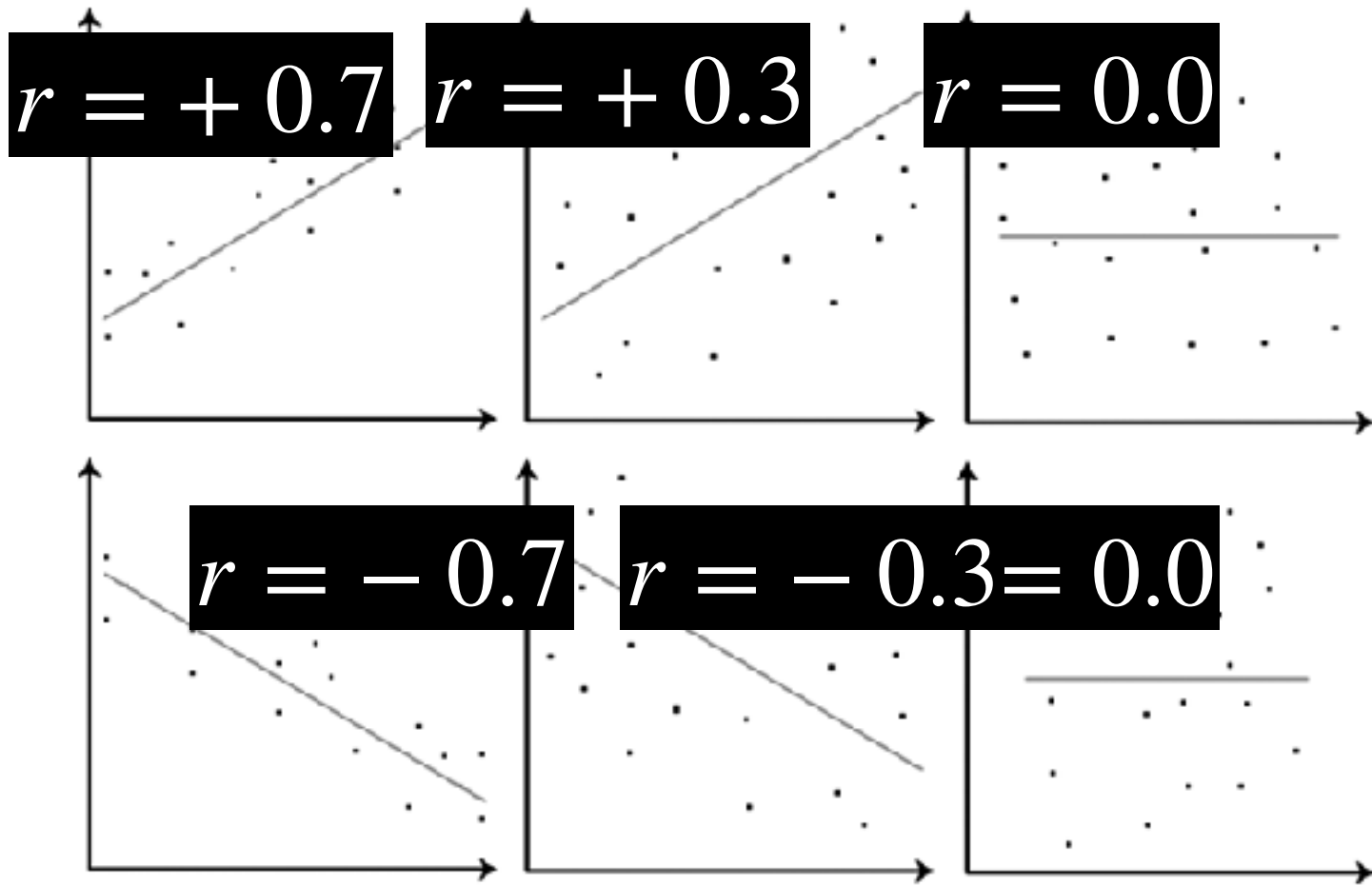


Correlation Coefficient = 0



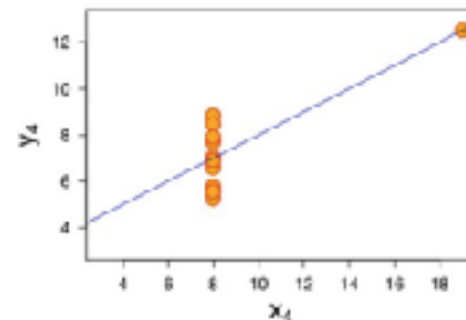
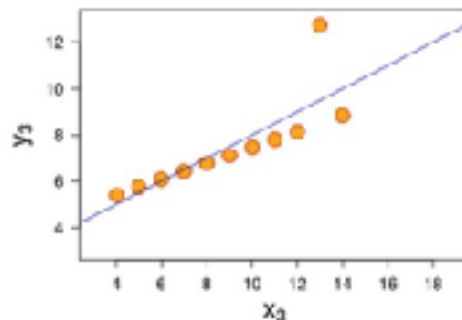
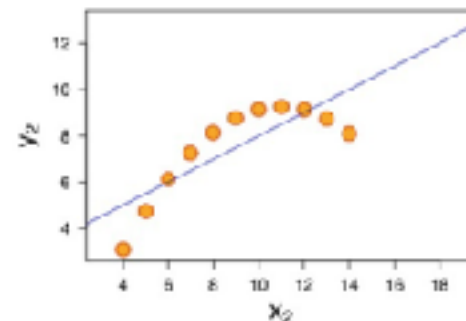
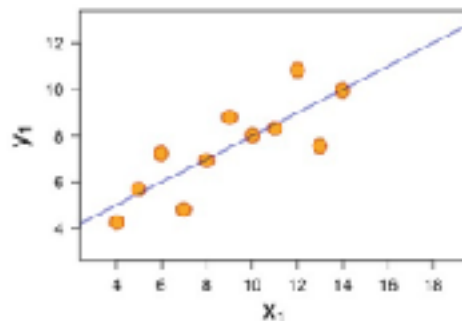
Correlation Coefficient = 1



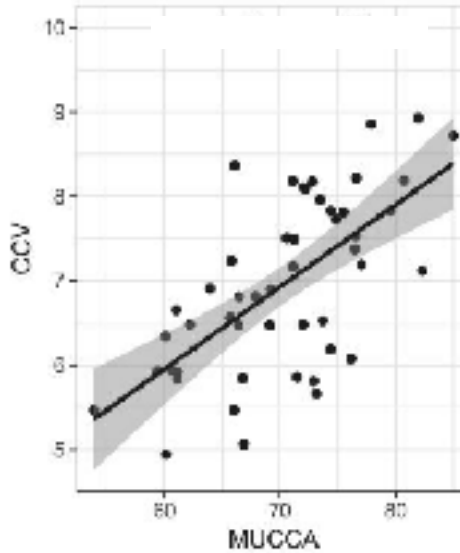


Anscombe's Quartet

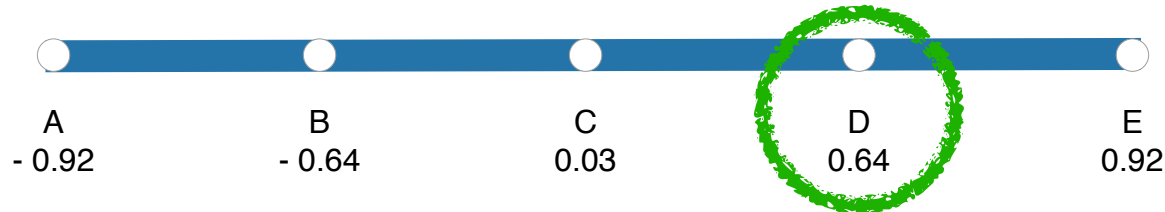
Property	Value
Mean of x in each case	9 (score)
Variance of x in each case	11 (score)
Mean of y in each case	7.50 (to 3 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case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



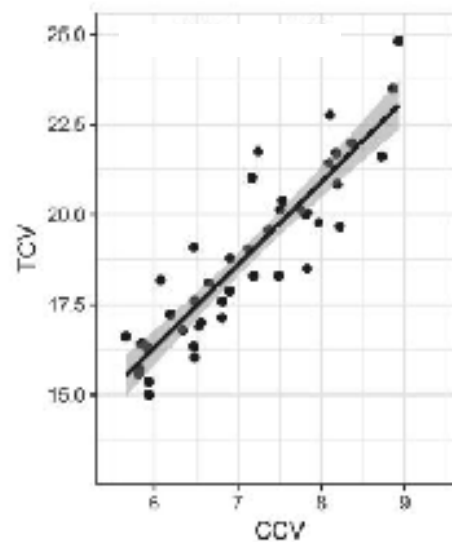
Correlation Champ



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



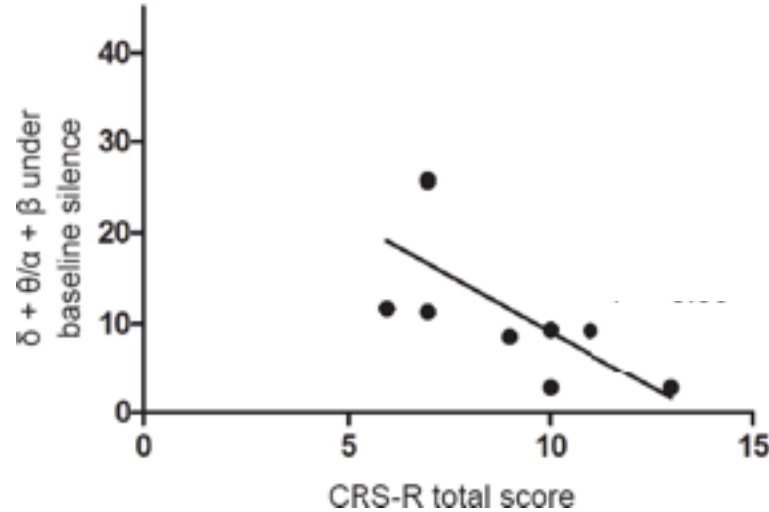
Correlation Champ



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

- A horizontal blue bar with five white circles representing radio buttons. The options are labeled below each circle:
- A -0.92
 - B -0.64
 - C 0.03
 - D 0.64
 - E 0.92
- The option E (0.92) is circled in green, indicating it is the correct answer.

Correlation Champ



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

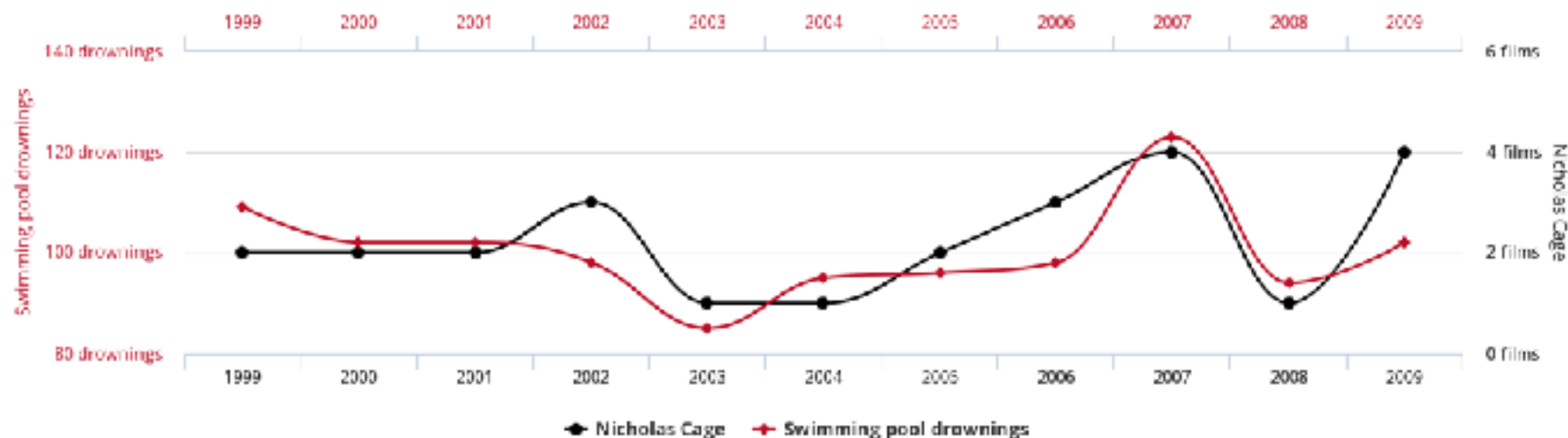
- A horizontal blue bar with five white circles representing options A, B, C, D, and E. Option B is circled in green.
- A -0.91
 - B -0.68**
 - C 0.03
 - D 0.68
 - E 0.90

Correlation \neq Causation

Correlation establishes a relationship.

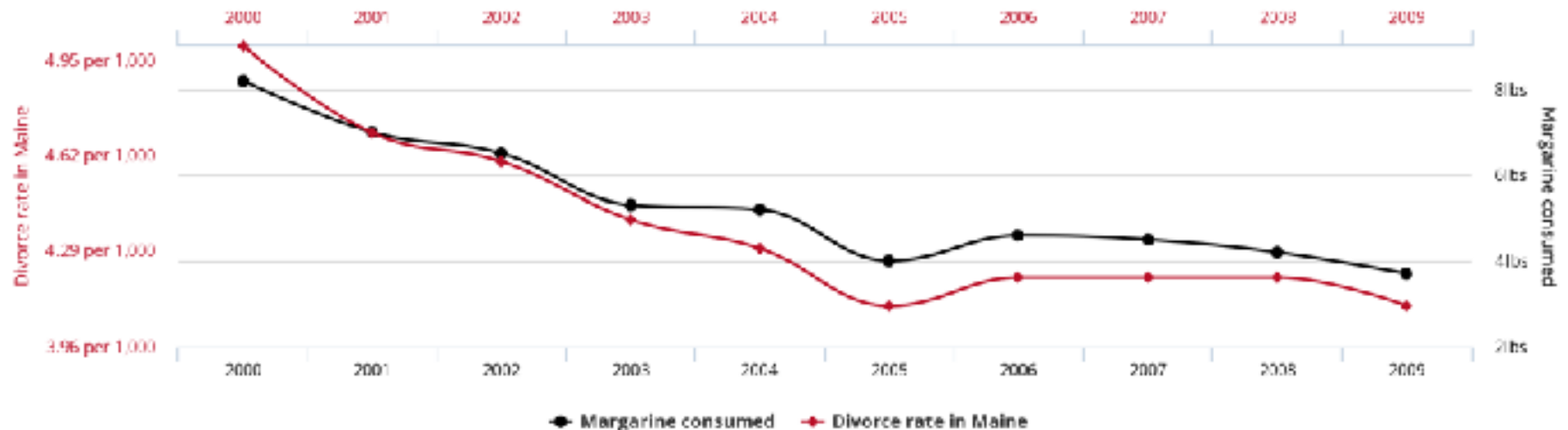
It does NOT establish causation.

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



tylervigen.com

Divorce rate in Maine
correlates with
Per capita consumption of margarine



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

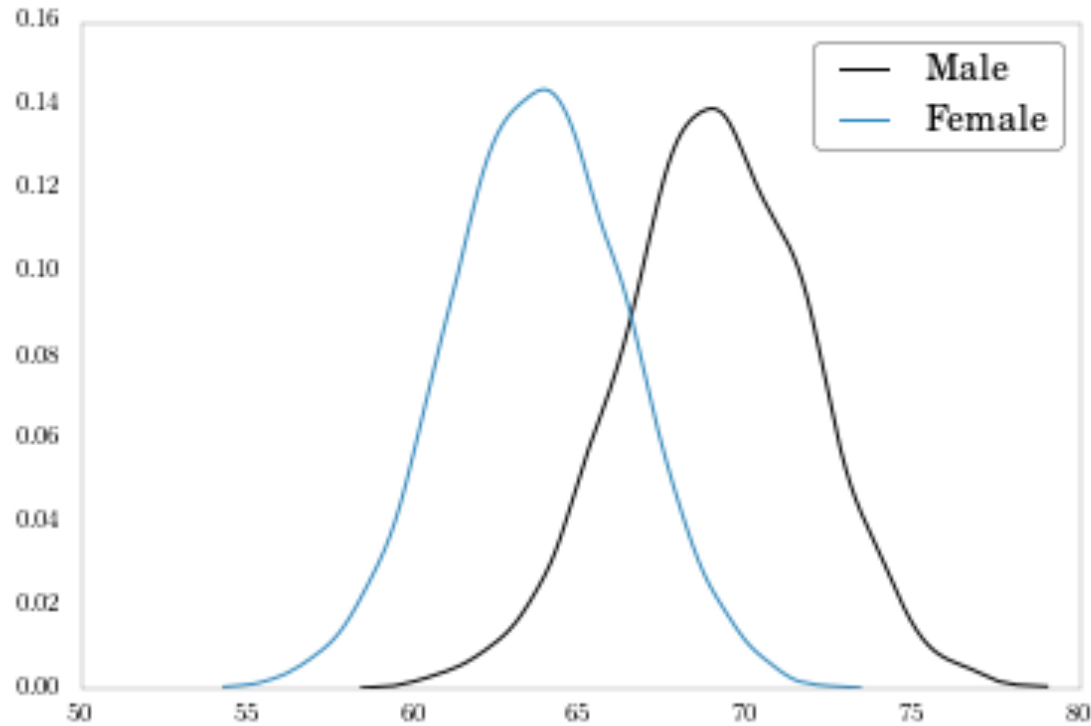
FOR WHEN ASSUMPTIONS
IN THESE OTHER 3
CATEGORIES ARE NOT
MET

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

t-test:

tests for difference in means between groups

Do the heights between males and females differ?

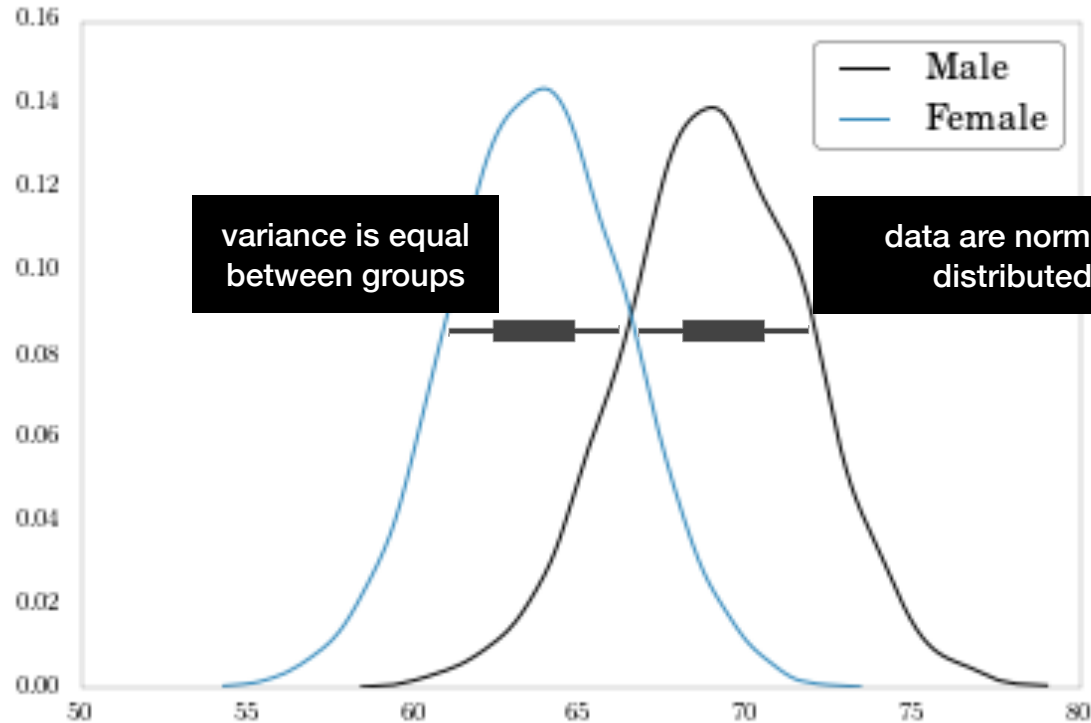


N=10,000

t-test Assumptions

1. Data are continuous
2. Normally distributed
3. Large enough sample size
4. Equal variance b/w groups

Do the heights between males and females differ?

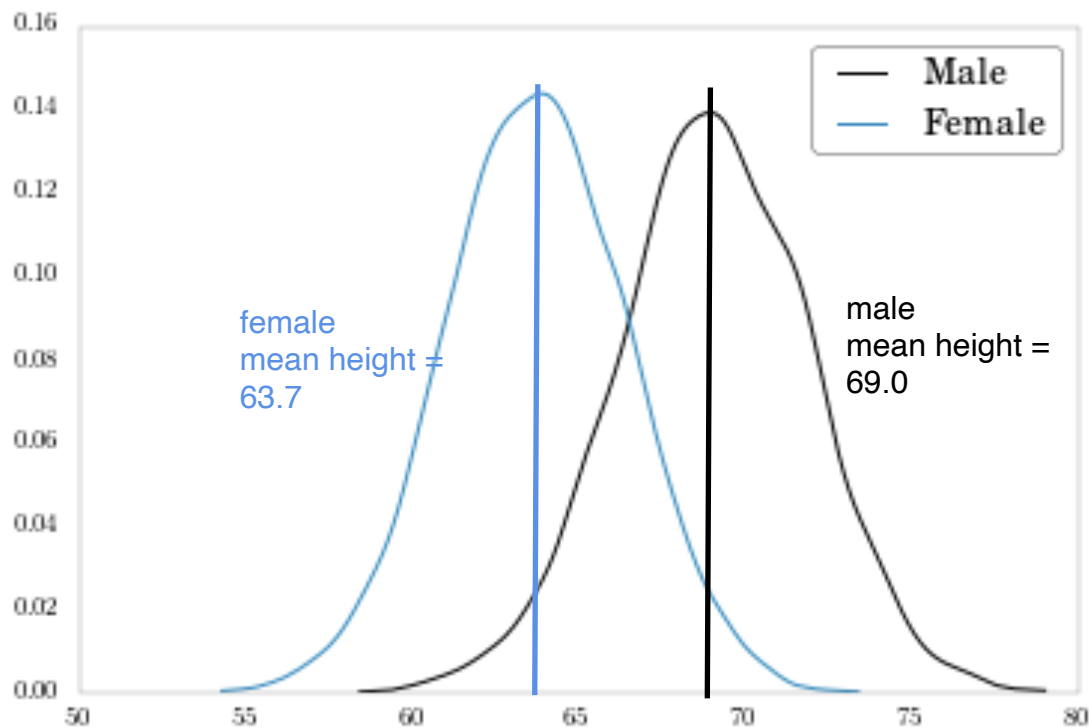


sample size is
large enough

$N=10,000$

data are continuous

Do the heights between males and females differ?



N=10,000

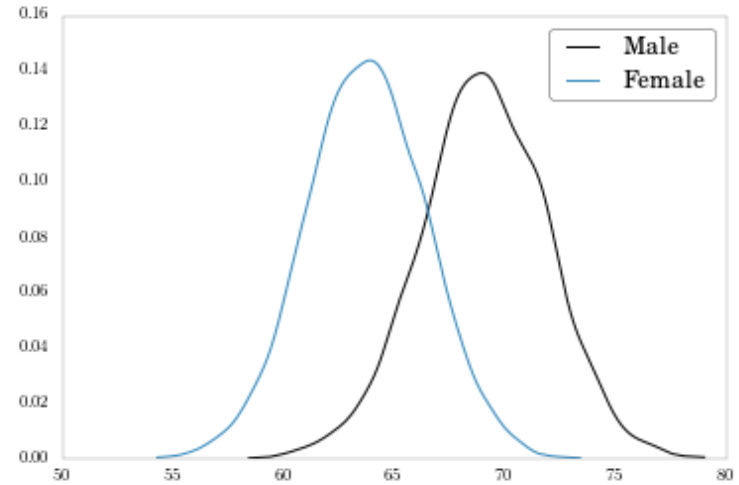
Do the heights between males and females differ?

t-statistic: -95.6

p-value $\ll 0.001$

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

Yes.



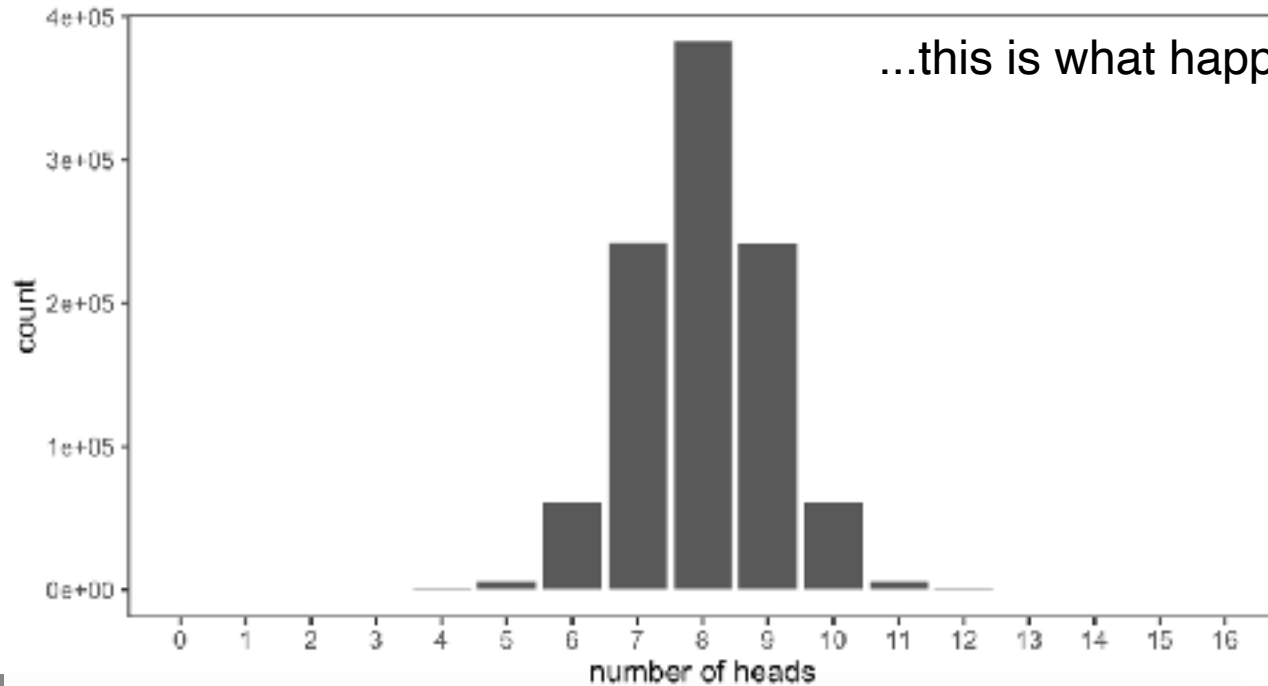
p-value : the probability of getting the observed results (or results more extreme) by chance alone

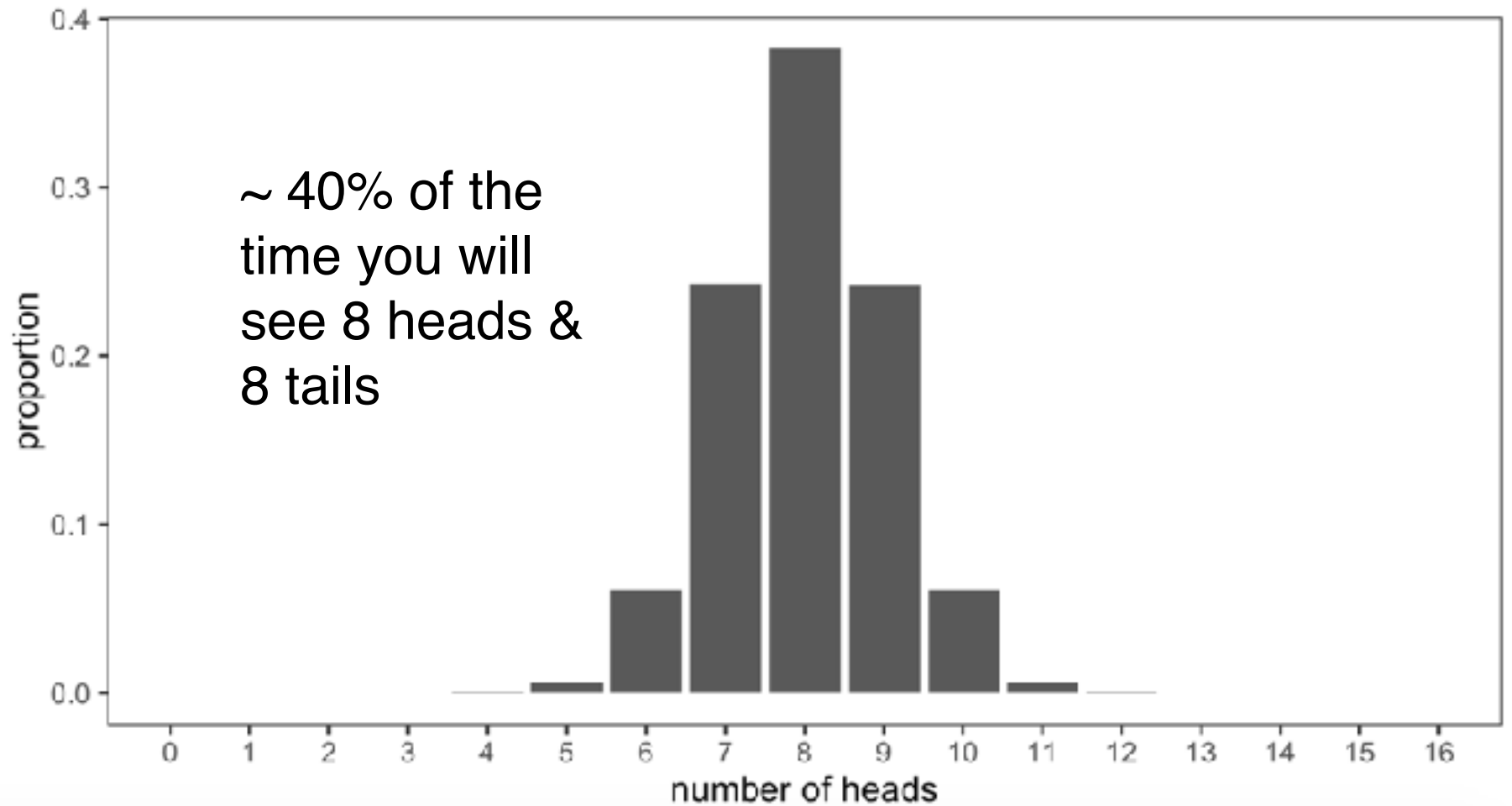


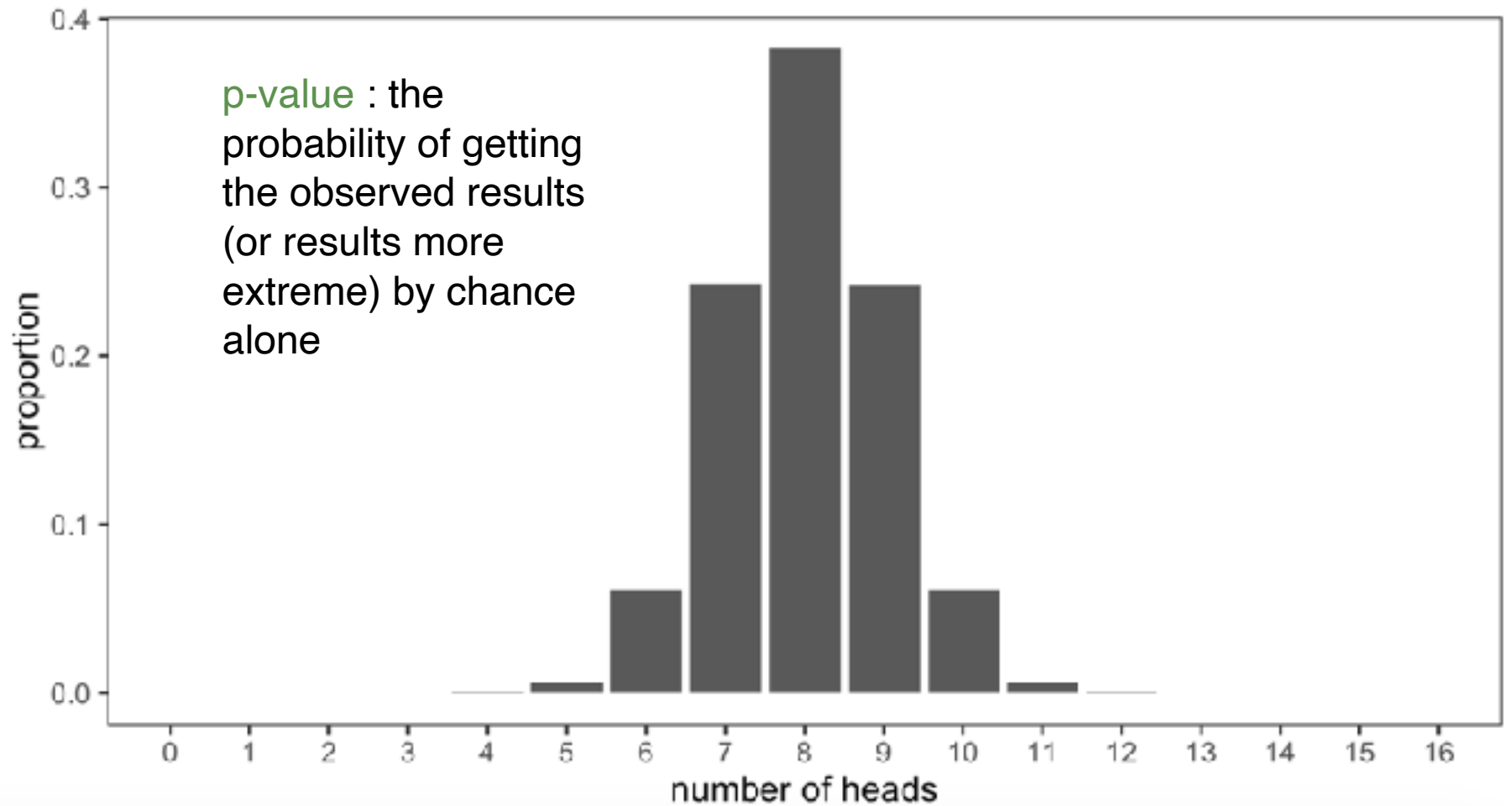
[https://forms.gle/
6MCyp7qFsaHgGKi5A](https://forms.gle/6MCyp7qFsaHgGKi5A)

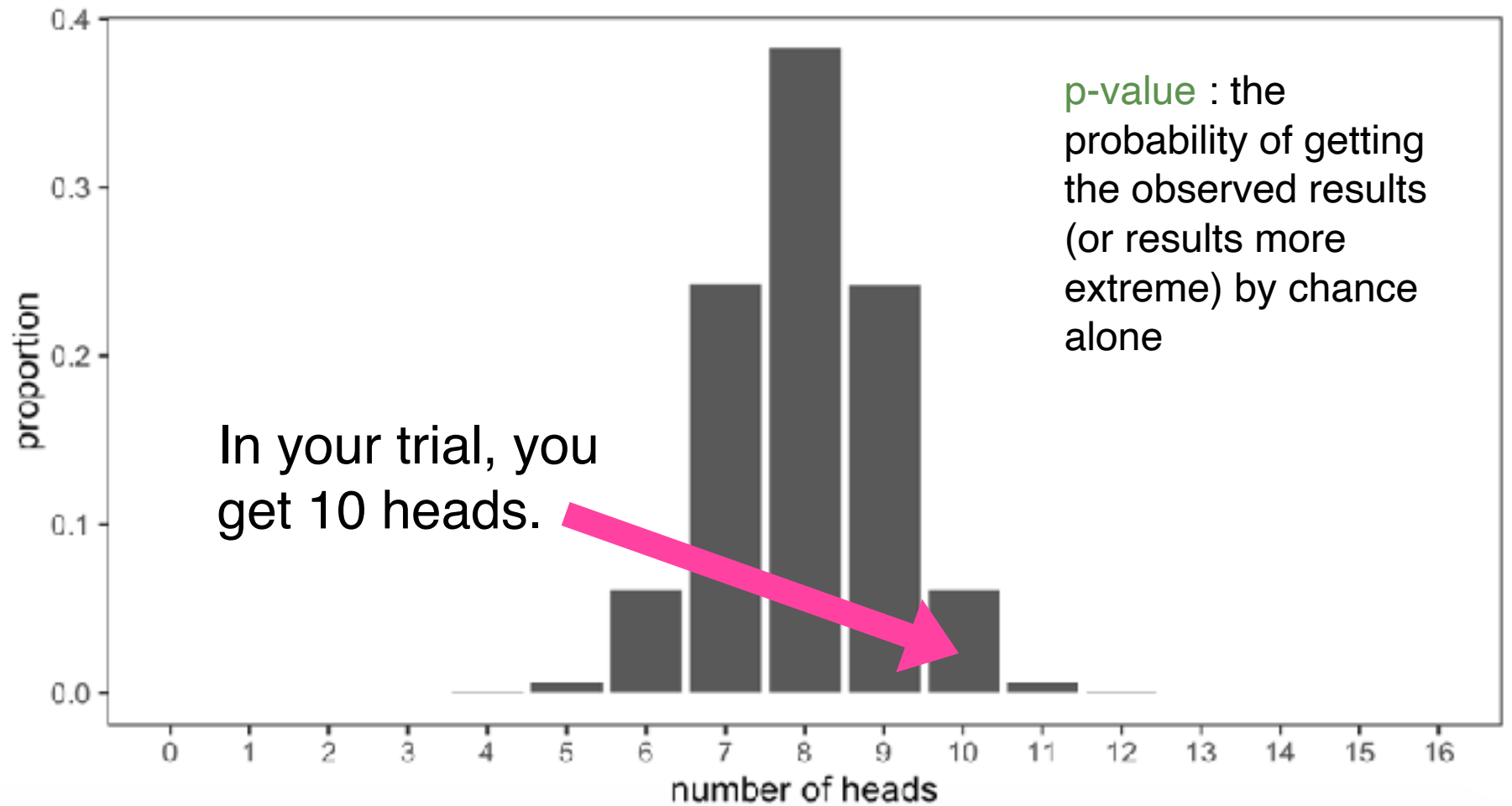
If we flip a coin 16 times and
record the number of heads....
....and then do that 1M times

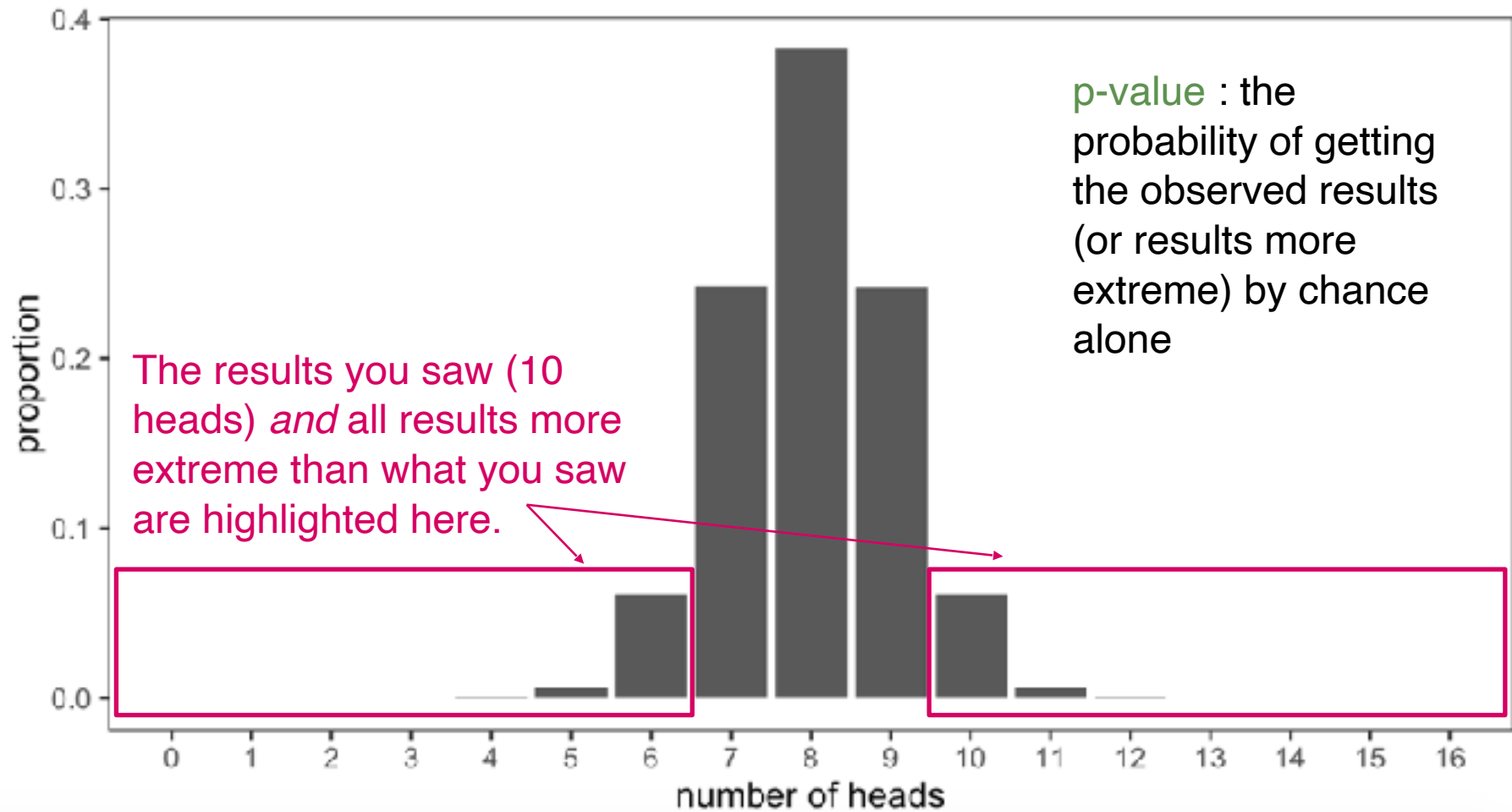
...this is what happens by chance alone.

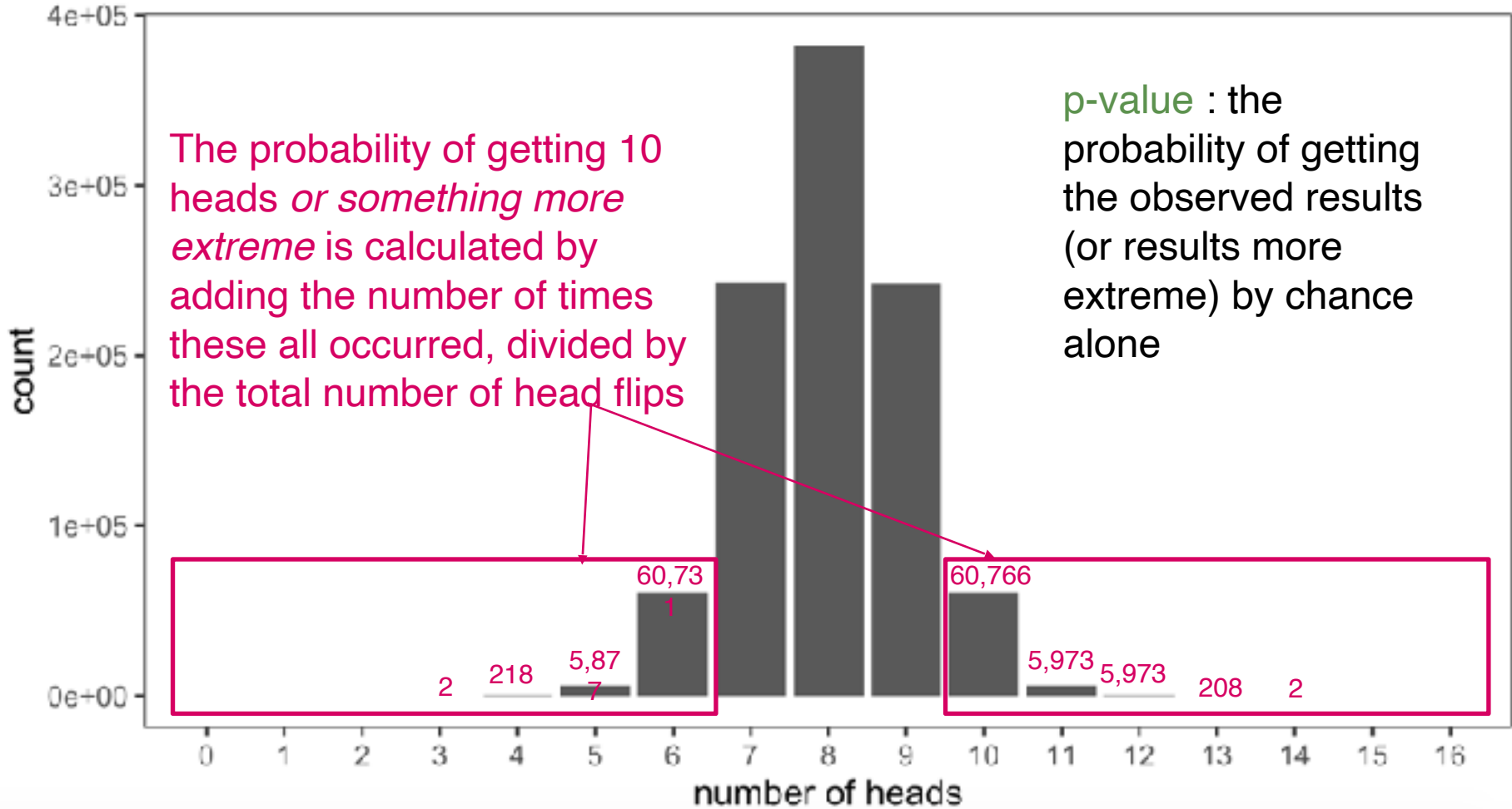


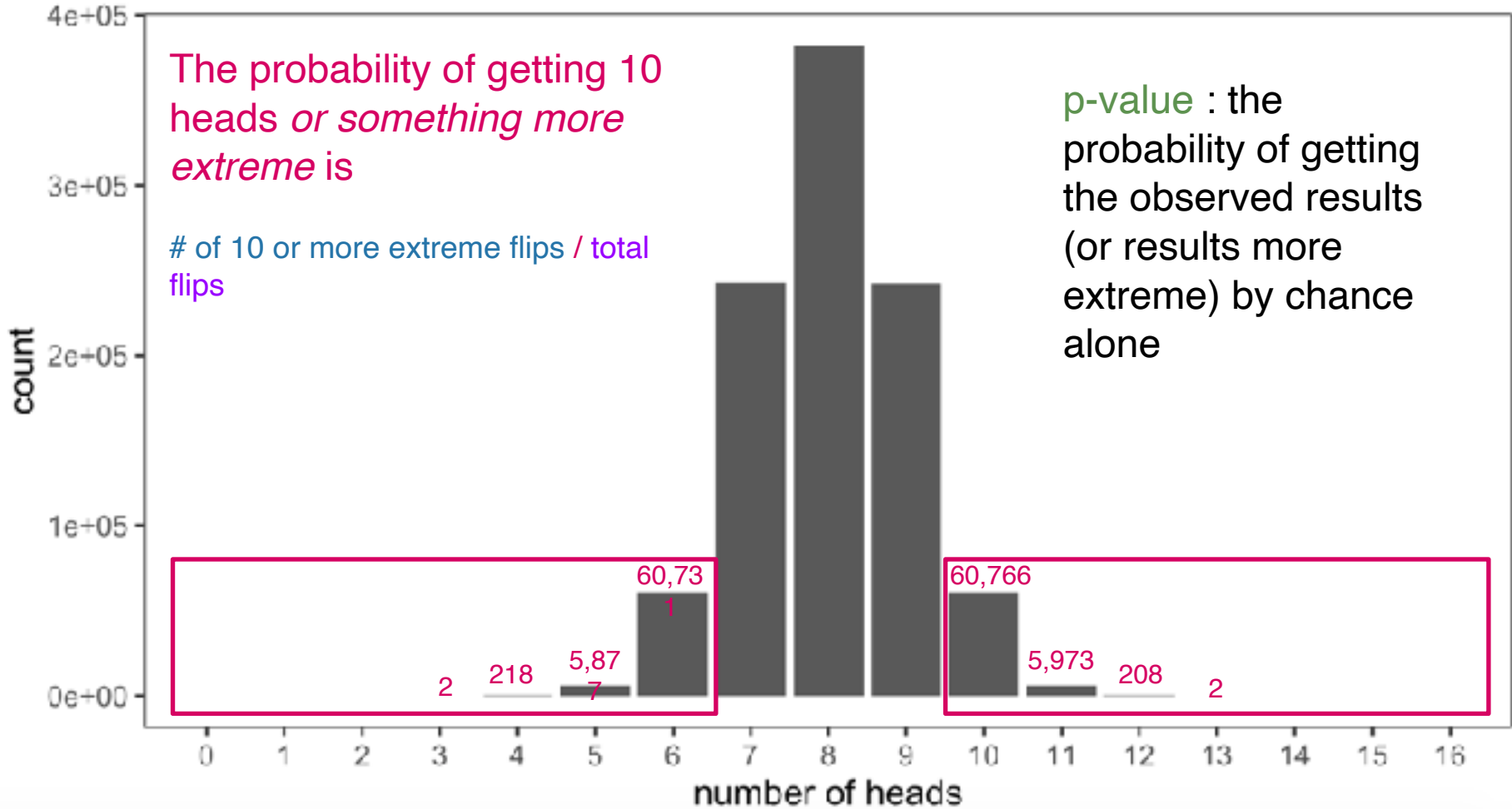


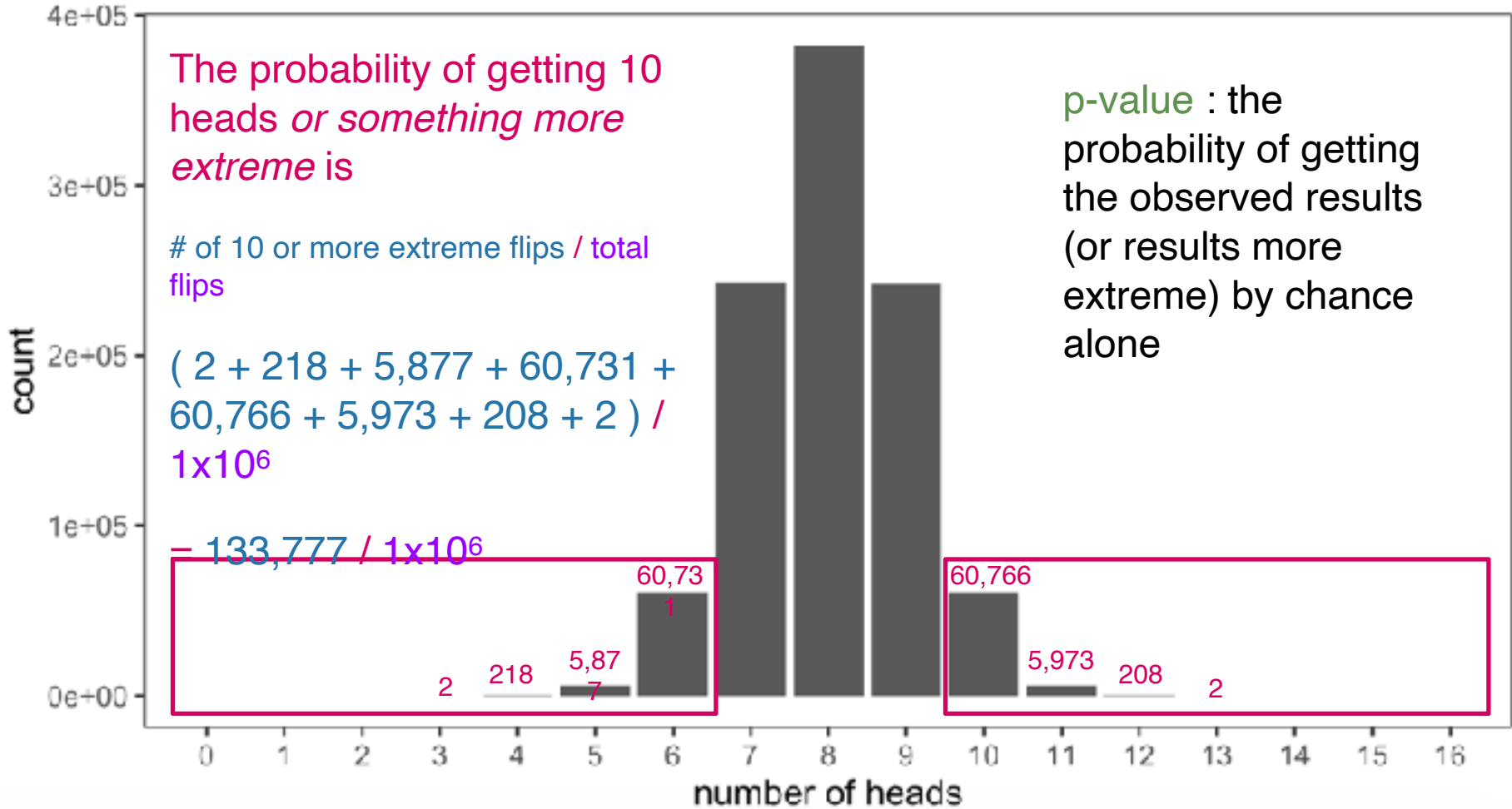


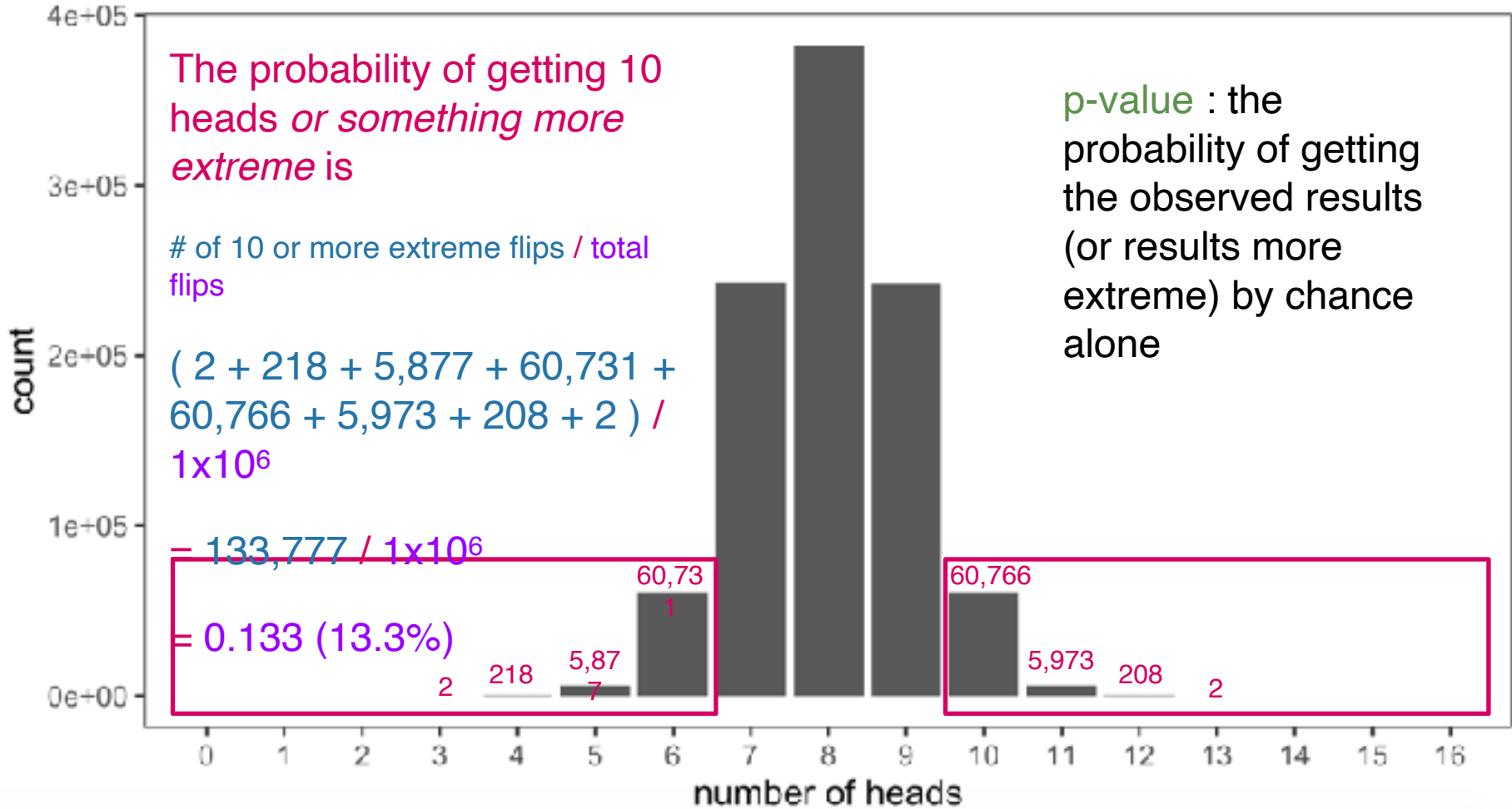


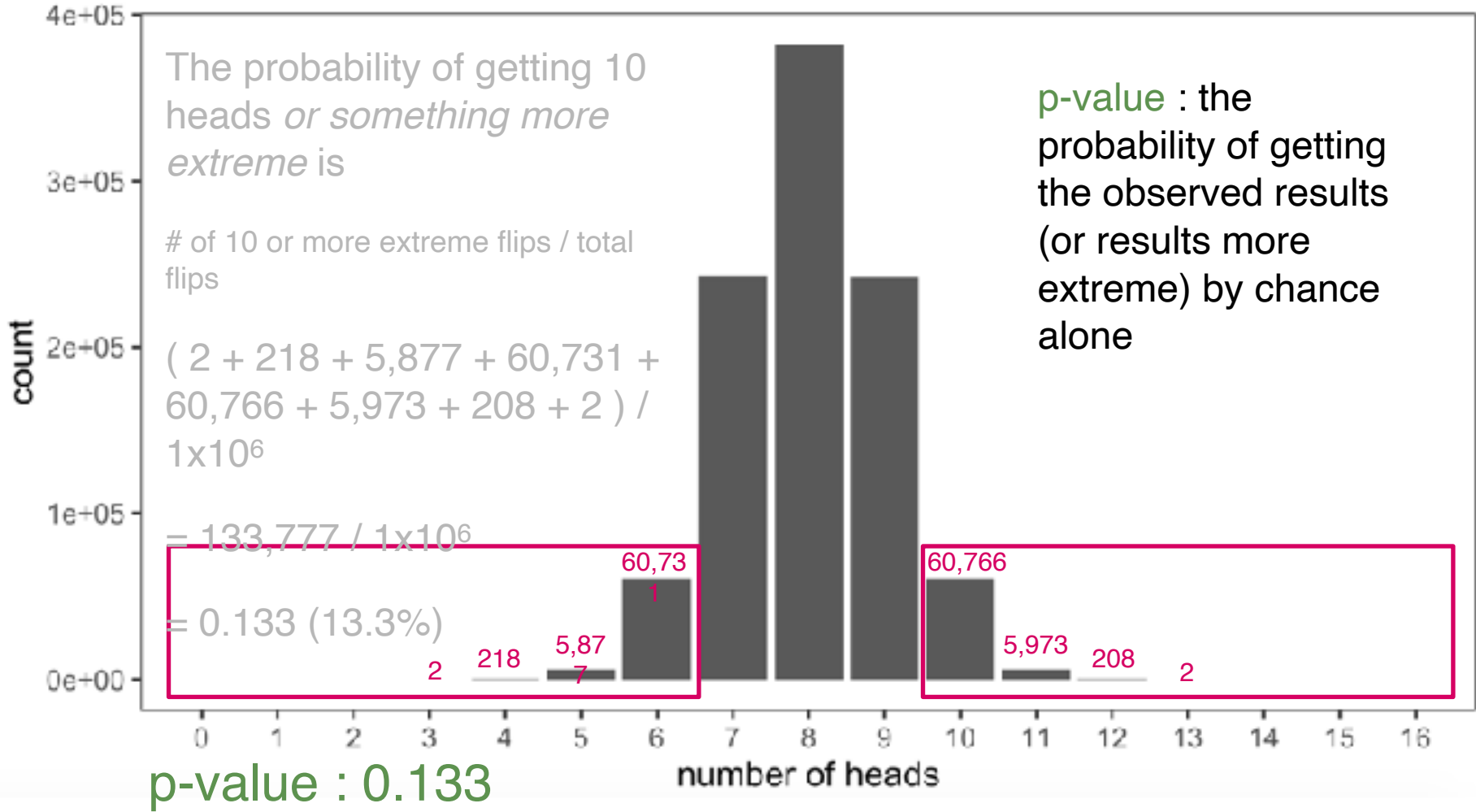


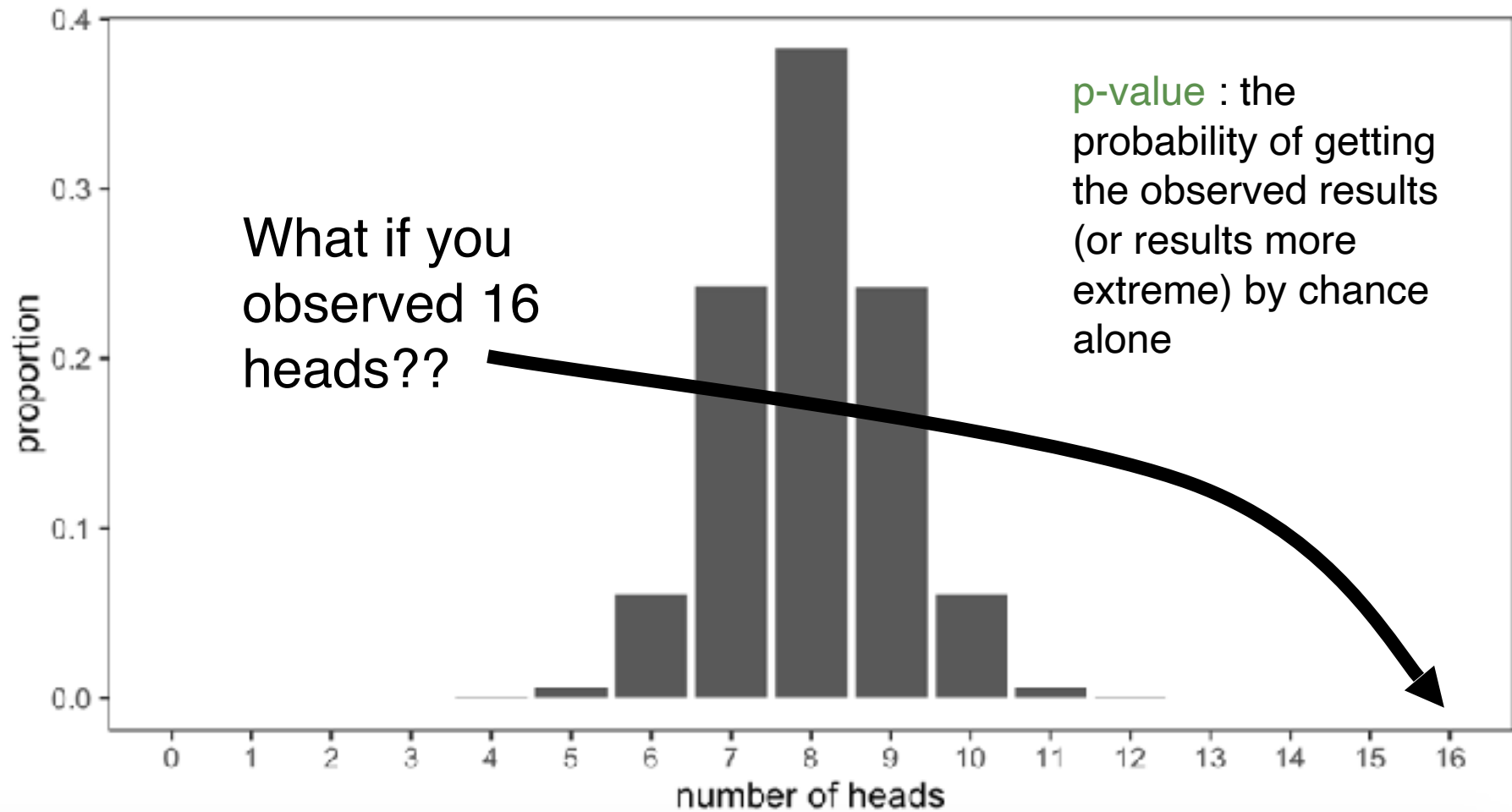


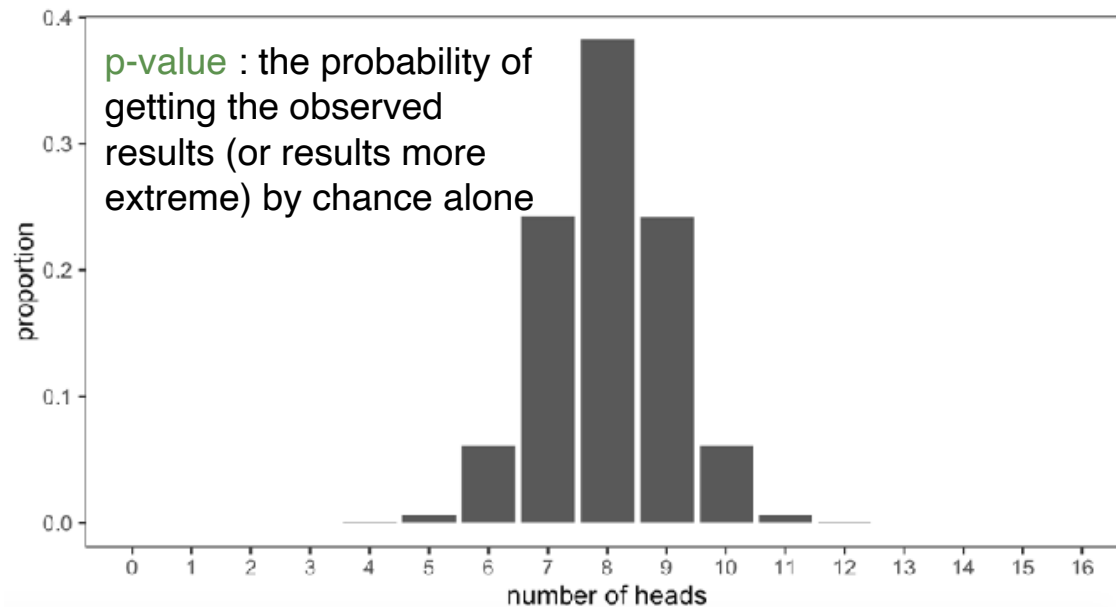












What would be the p-value of you flipping 16 heads?



Do the heights between males and females differ?

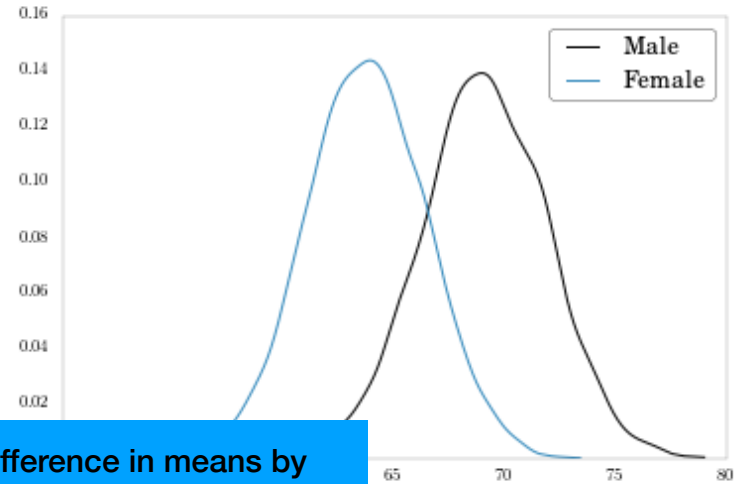
t-statistic: -95.6

p-value $\ll 0.001$

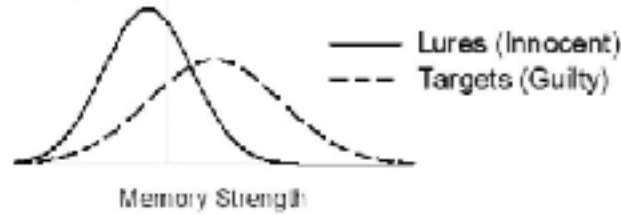
The probability of seeing this difference in means by random chance alone is much less than 1 in 1000

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?

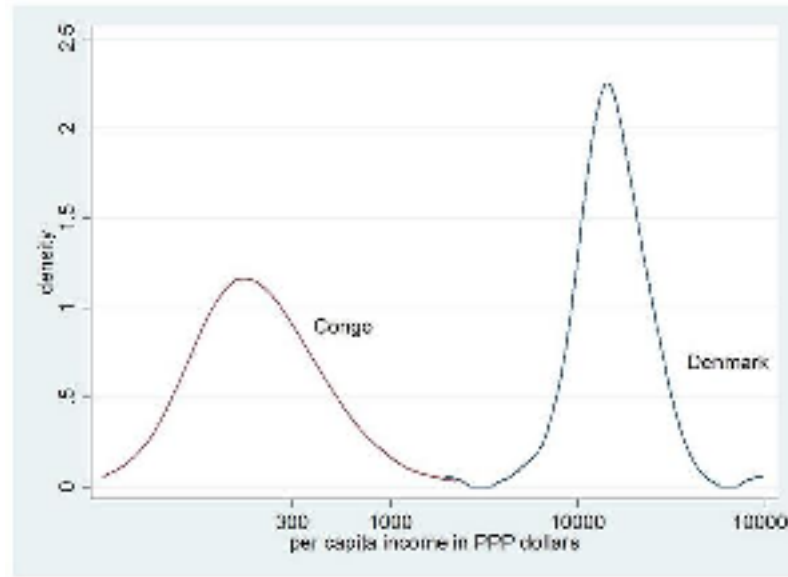


A
Not normally
distributed

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

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

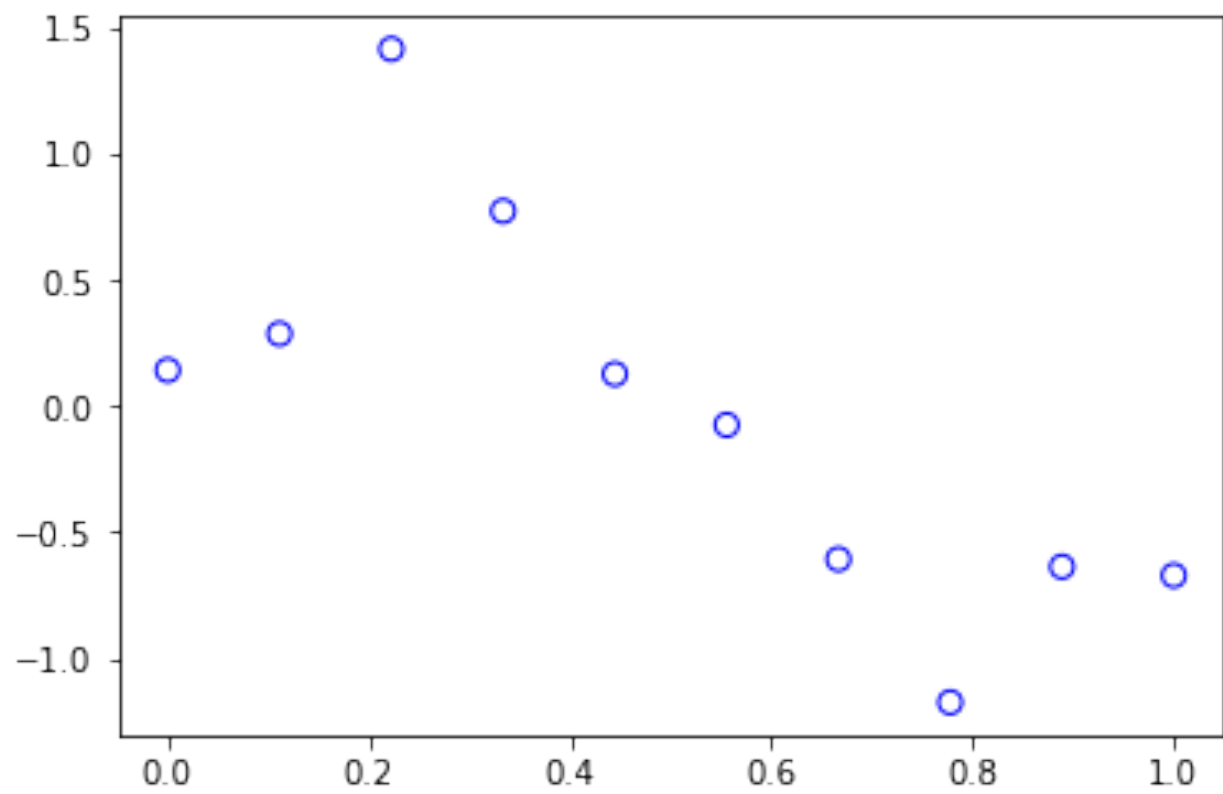
DOES CHANGE IN ONE
VARIABLE MEAN CHANGE
IN ANOTHER?

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



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?

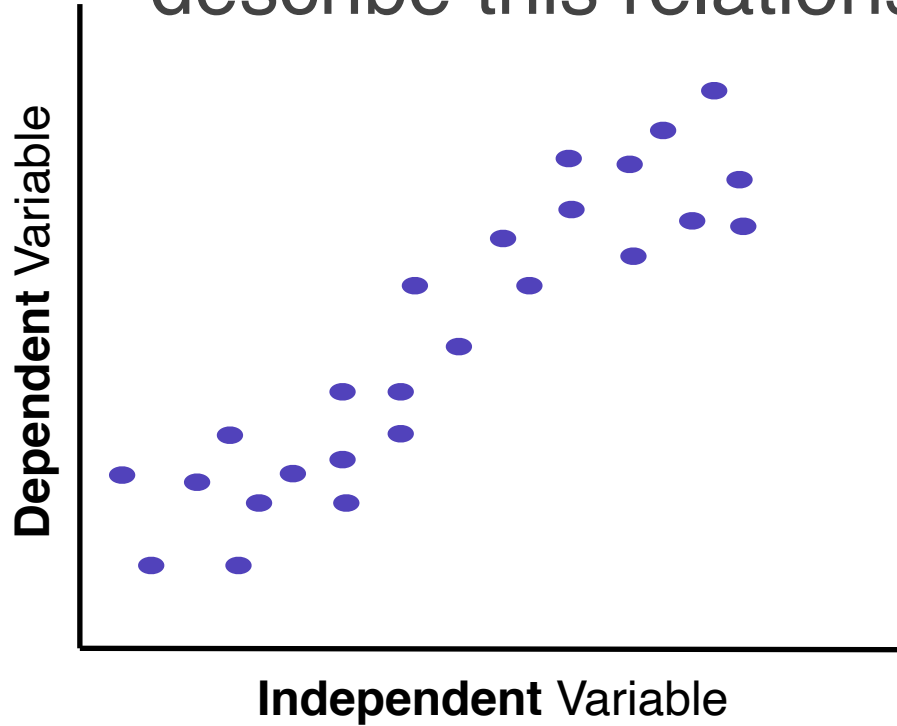
I.e. simple
regression, multiple
regression

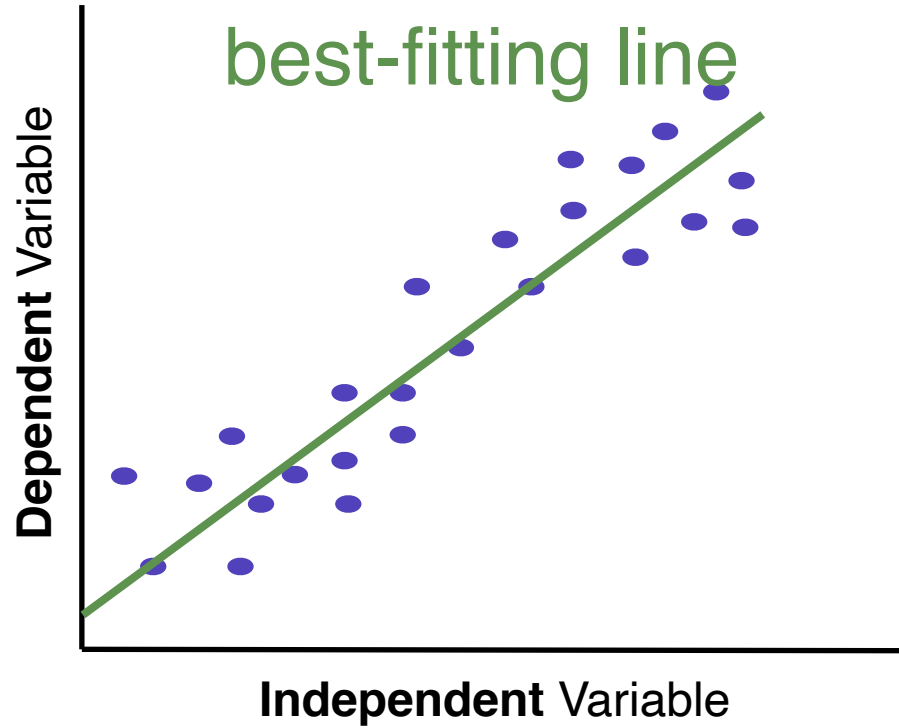
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS
IN THESE OTHER 3
CATEGORIES ARE NOT
MET

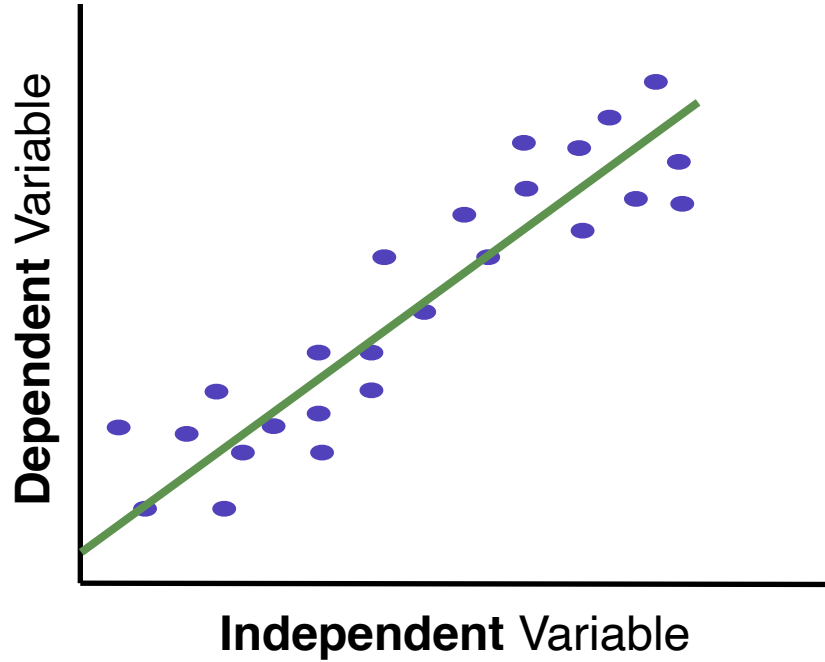
i.e. Wilcoxon rank-
sum test, Wilcoxon
sign-rank test, sign
test

Linear regression can be used to describe this relationship

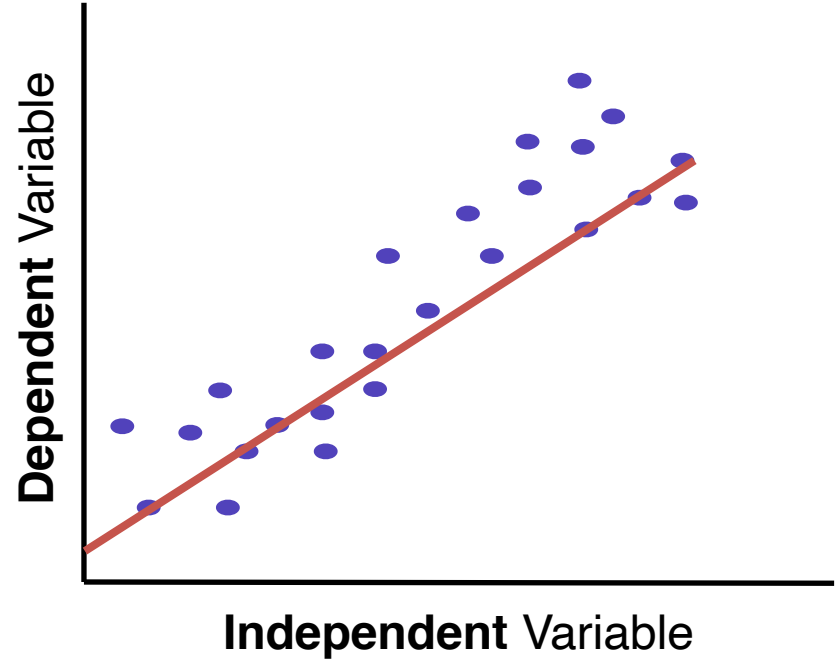




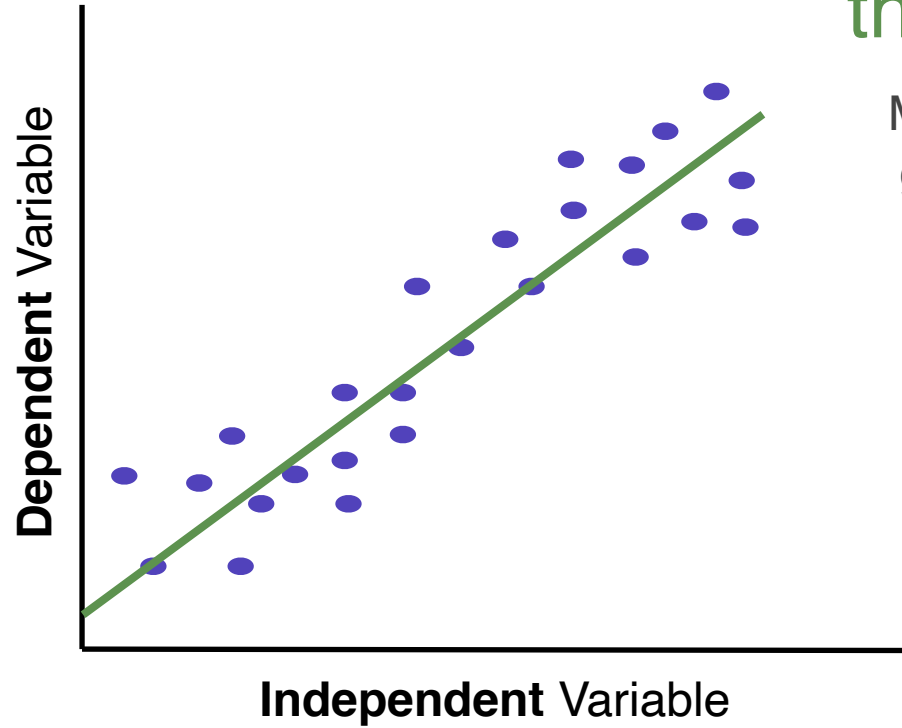
Best-fitting line



NOT a best-fitting line

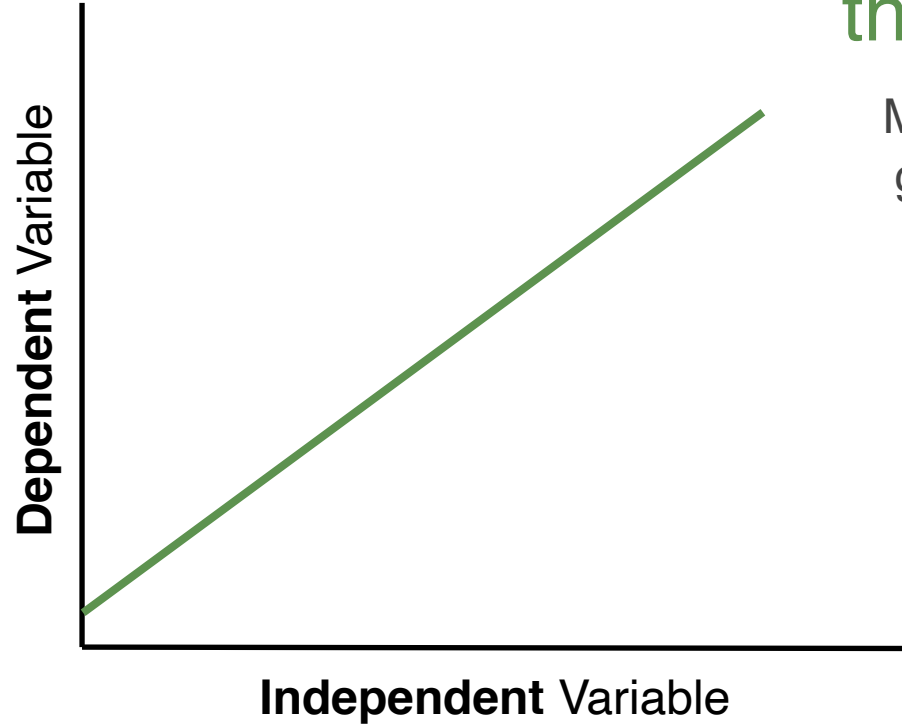


This line is a model of the data



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

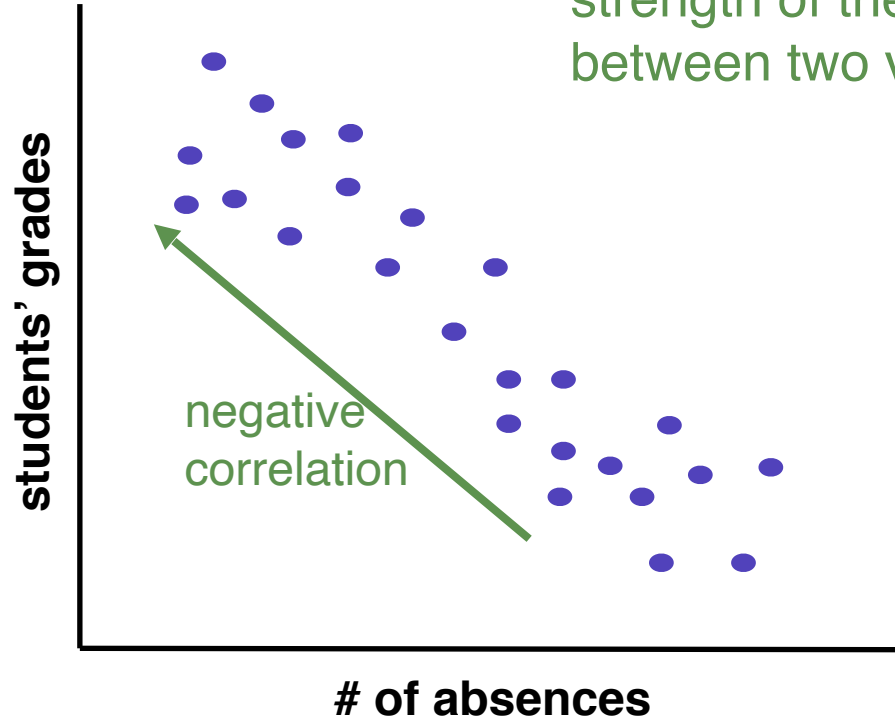
2.3 Parsimony

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.

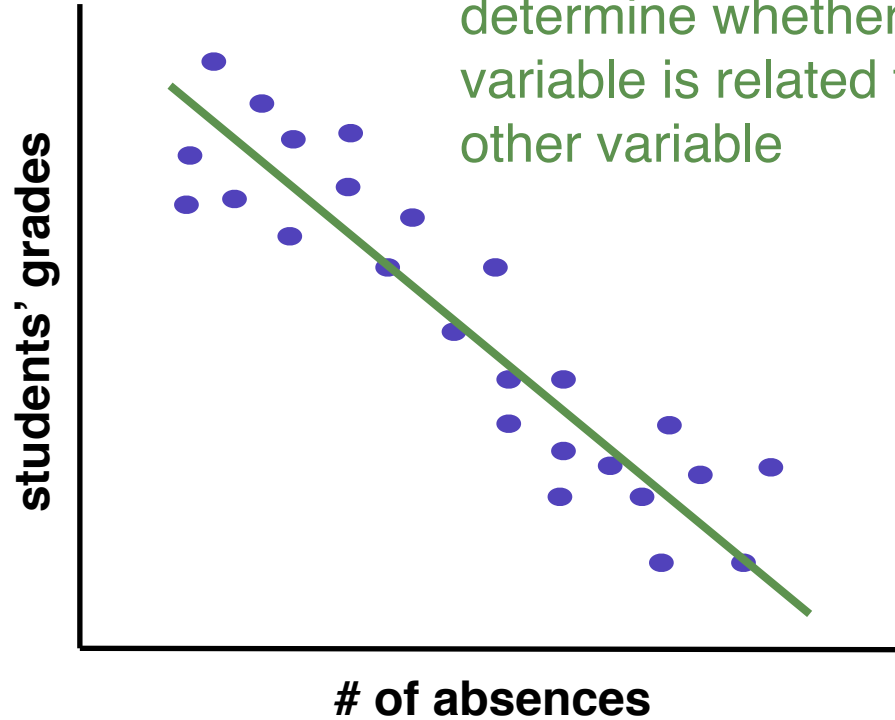
2.4 Worrying Selectively

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.

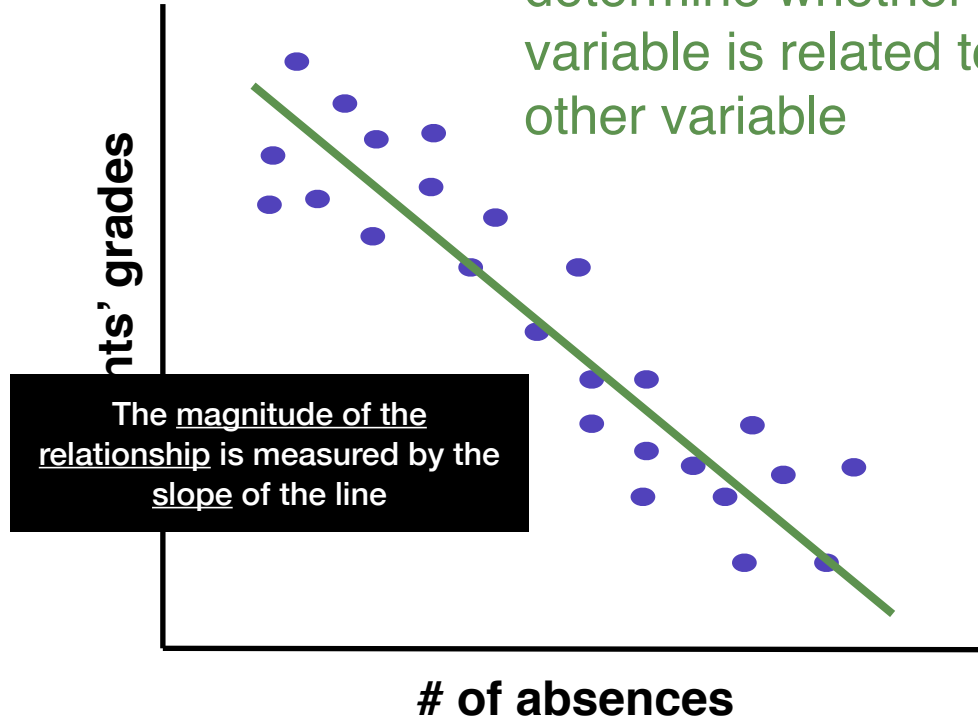
Correlation measures the strength of the linear relationship between two variables



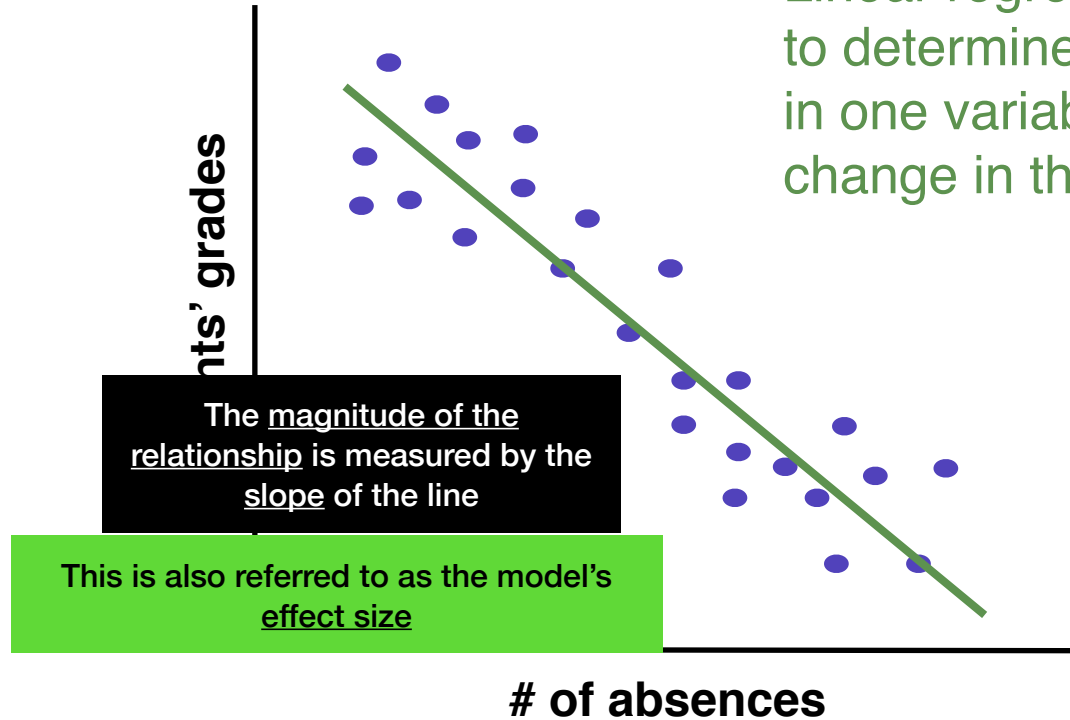
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



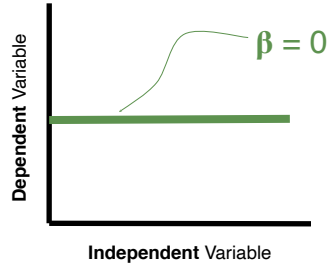
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



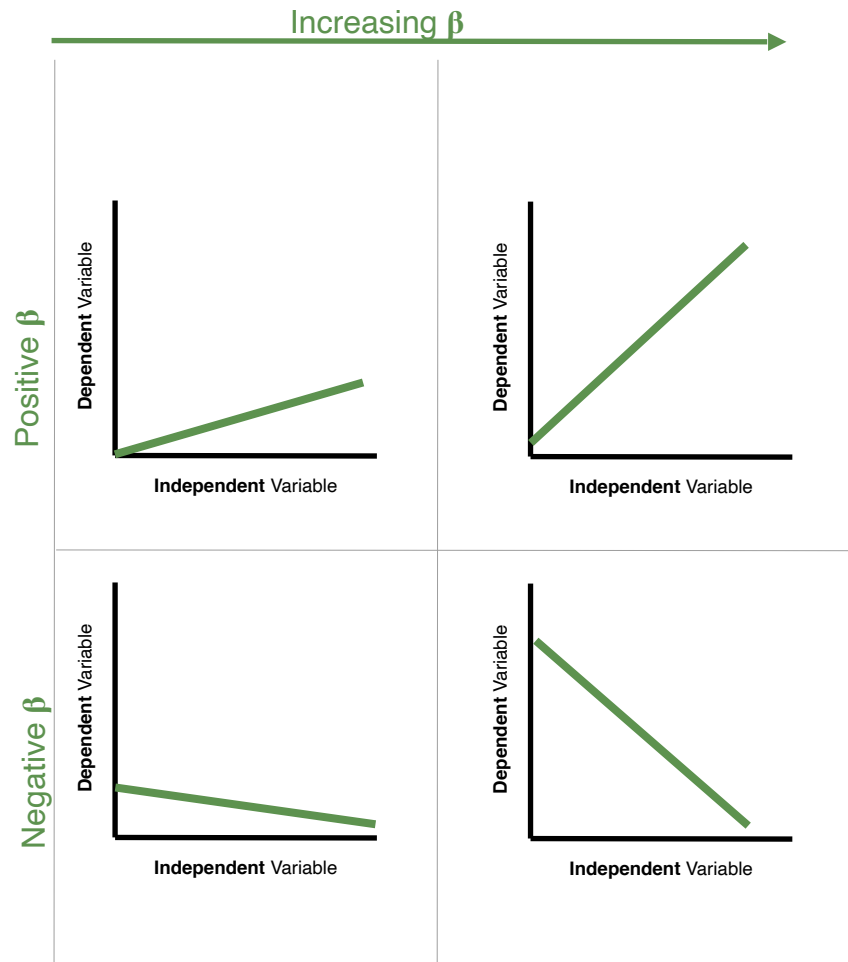
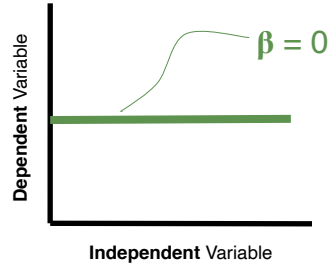
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



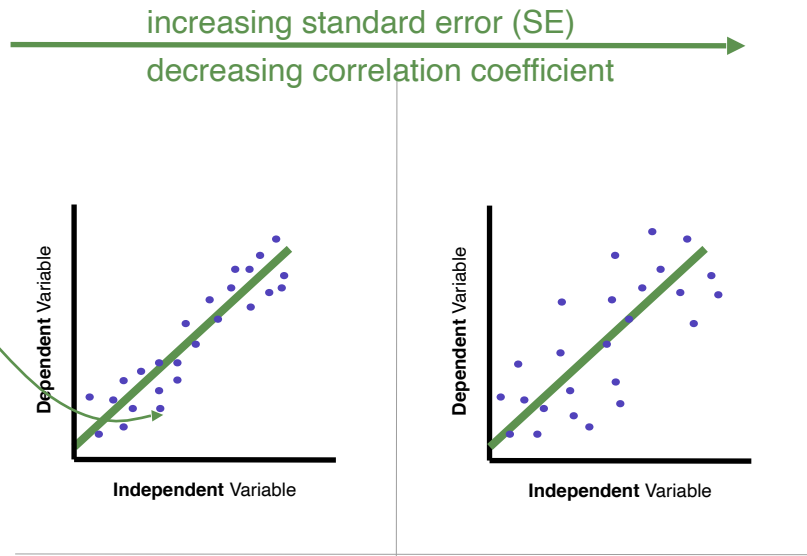
Effect size (β) can
be estimated using
the slope of the line



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

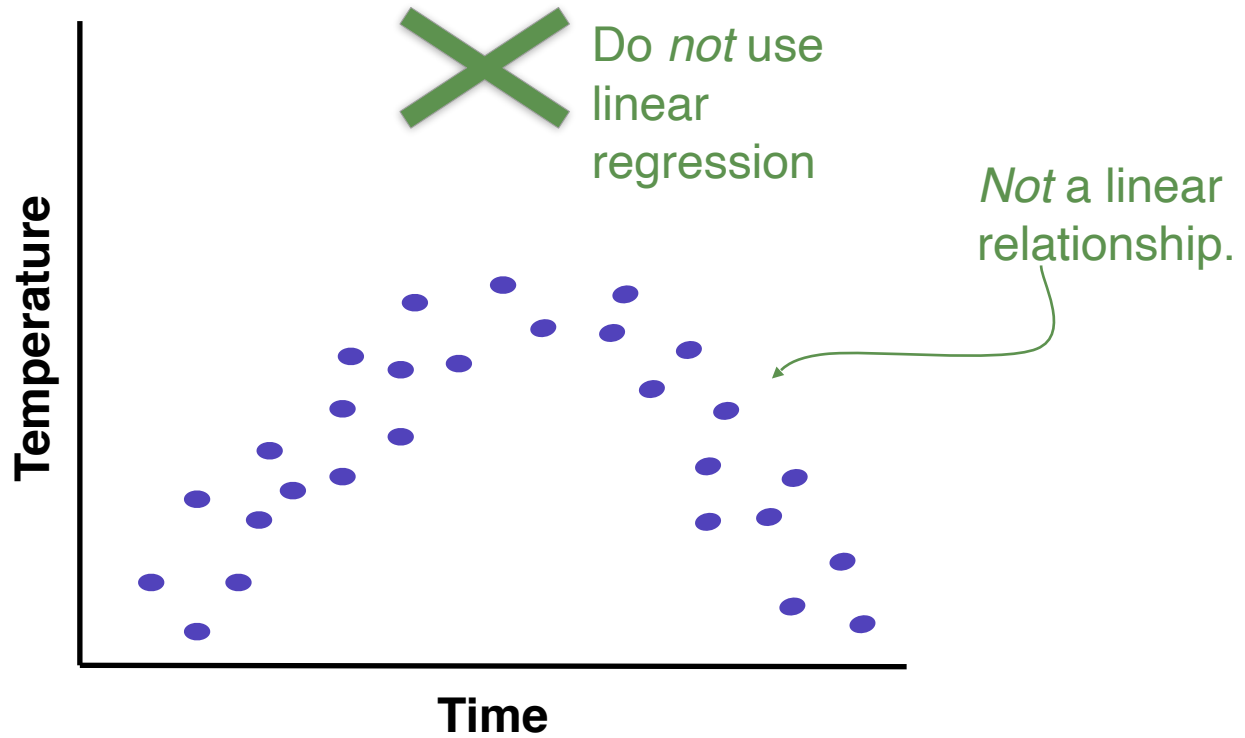


The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

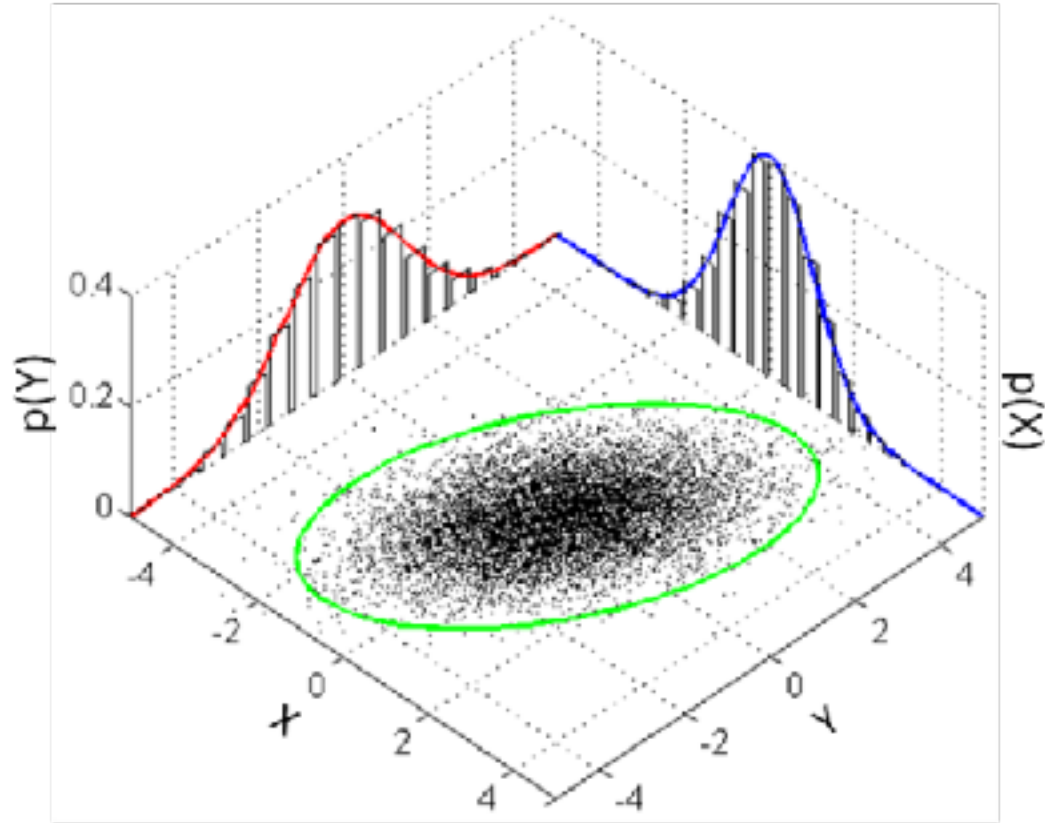


Assumptions of linear regression

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



A multivariate normal probability distribution (joint normal)

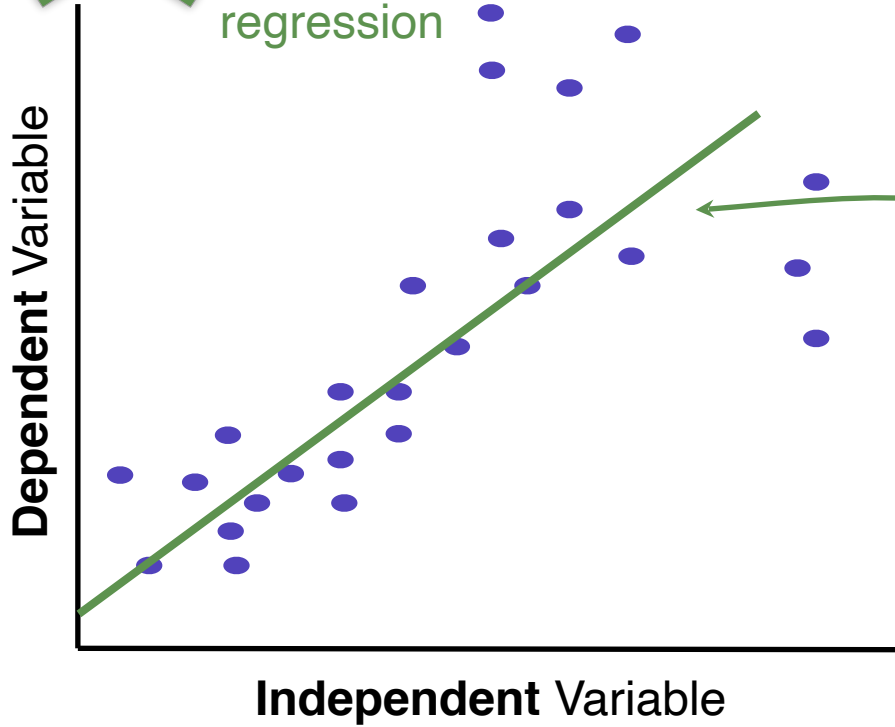


Linear regression assumes no multicollinearity. **Multicollinearity** occurs when the independent variables (in multiple linear regression) are too highly correlated with each other.



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

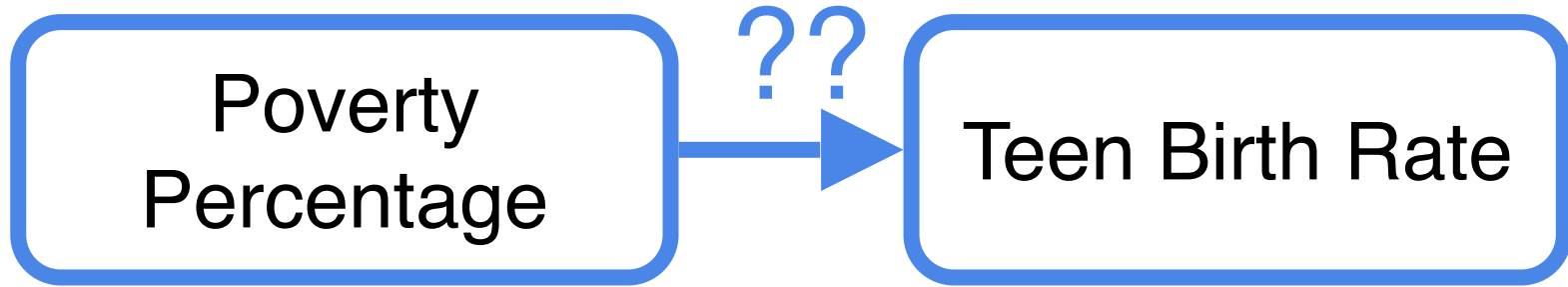
 Do *not* use
linear
regression



Not
homoscedastic:
points at this end are much
further from the line than at
the other end



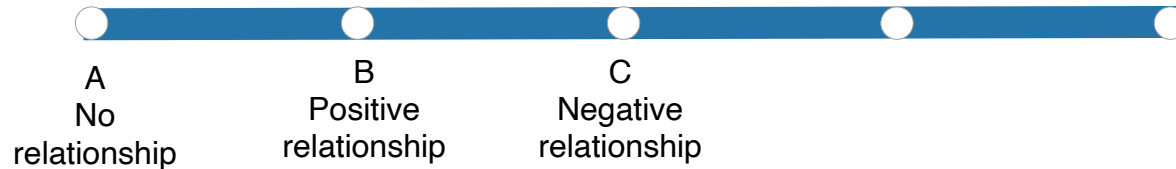
Does Poverty
Percentage affect Teen
Birth Rate?





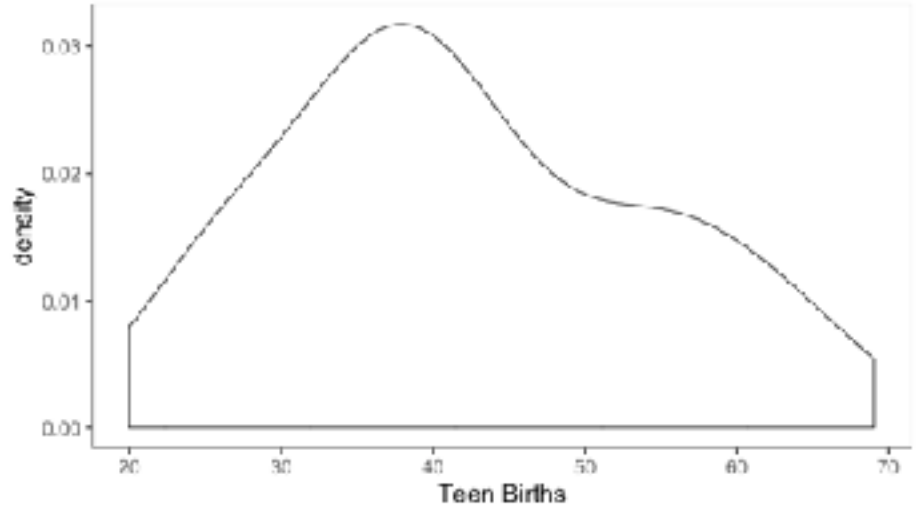
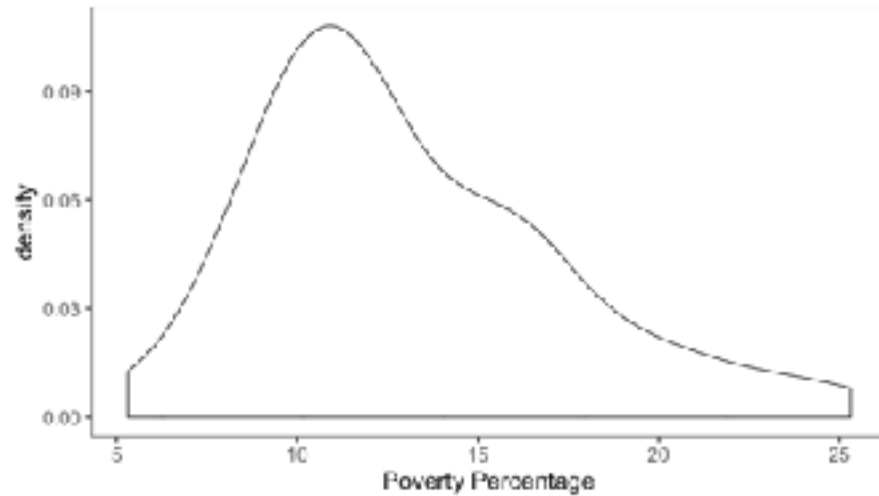
What is the relationship between Poverty Percentage & Teen Birth Rate?

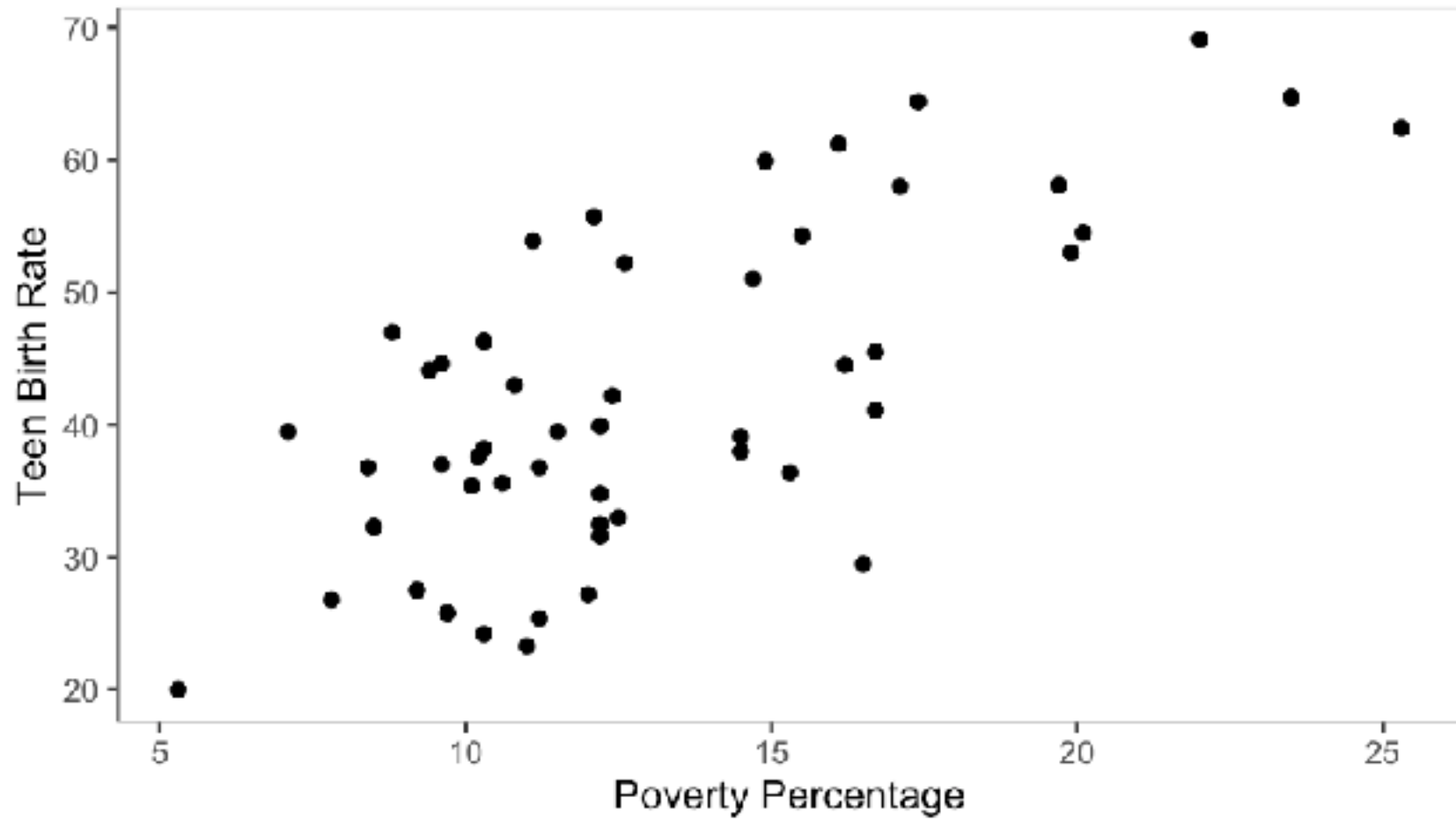
What's your hypothesis?

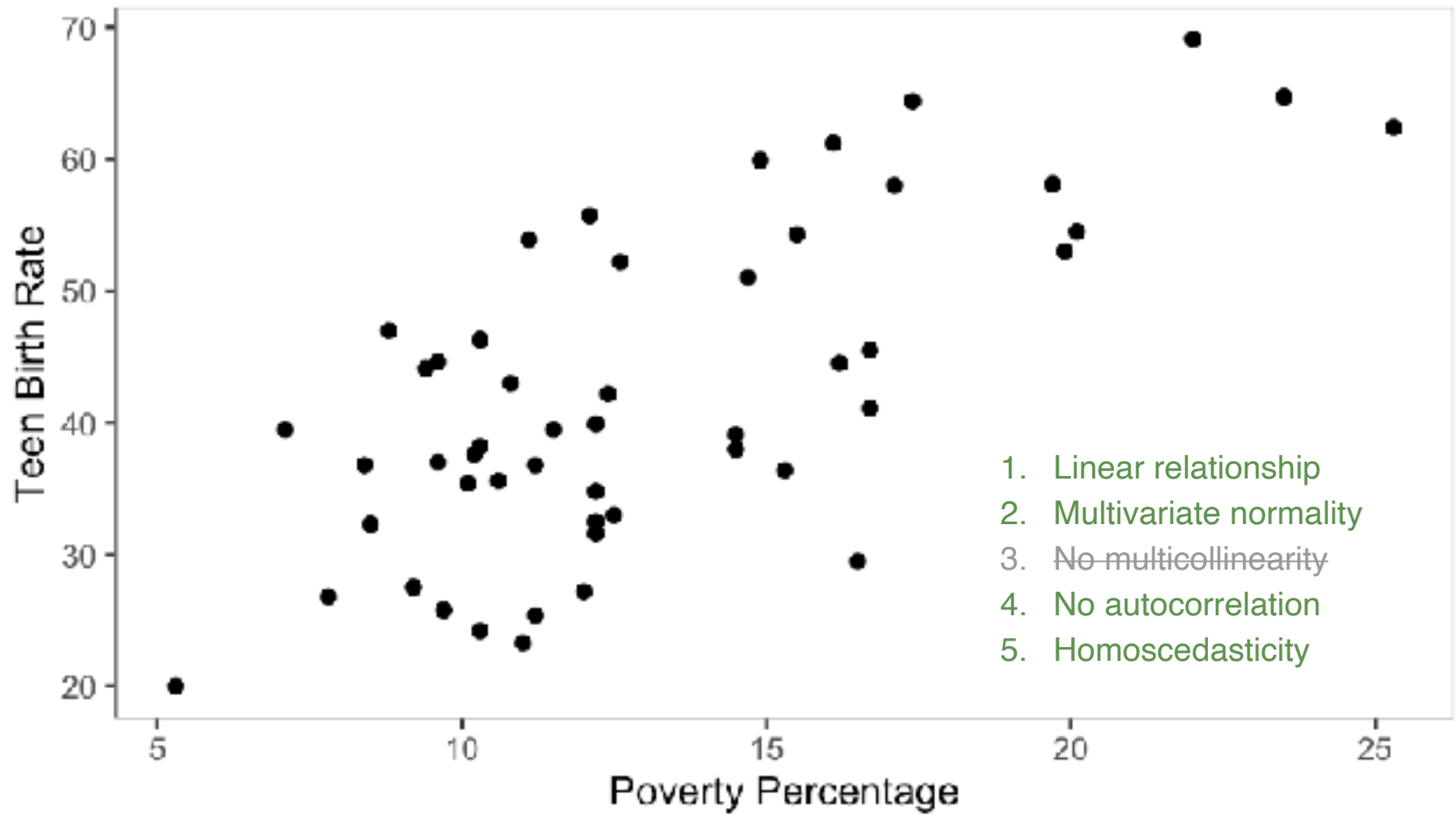


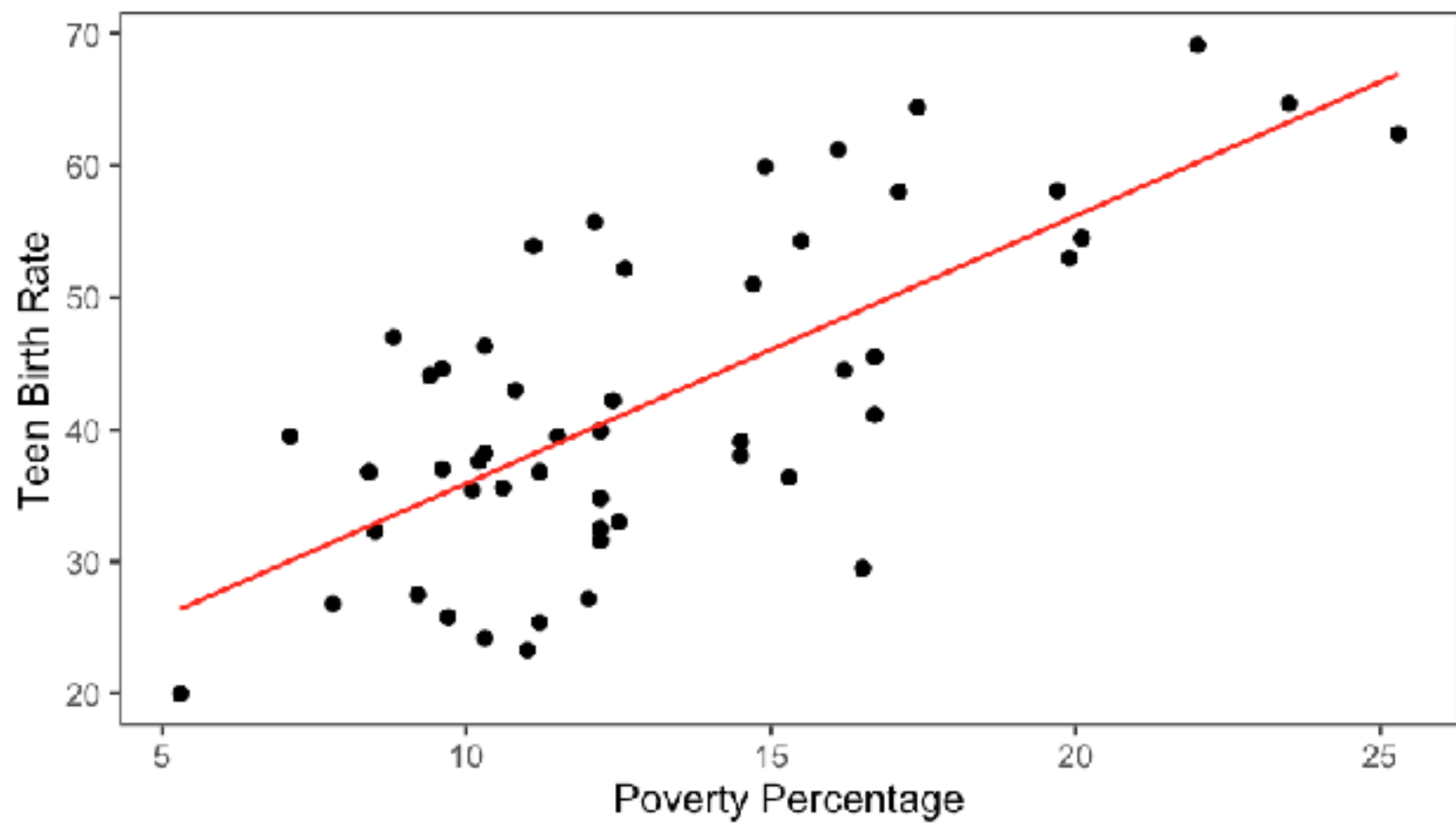
	Location	PovPct	Brth15to17	Brth18to19	ViolCrime	TeenBrth
1	Alabama	20.1	31.5	88.7	11.2	54.5
2	Alaska	7.1	18.9	73.7	9.1	39.5
3	Arizona	16.1	35.0	102.5	10.4	61.2
4	Arkansas	14.9	31.6	101.7	10.4	59.9
5	California	16.7	22.6	69.1	11.2	41.1
6	Colorado	8.8	26.2	79.1	5.8	47.0
7	Connecticut	9.7	14.1	45.1	4.6	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

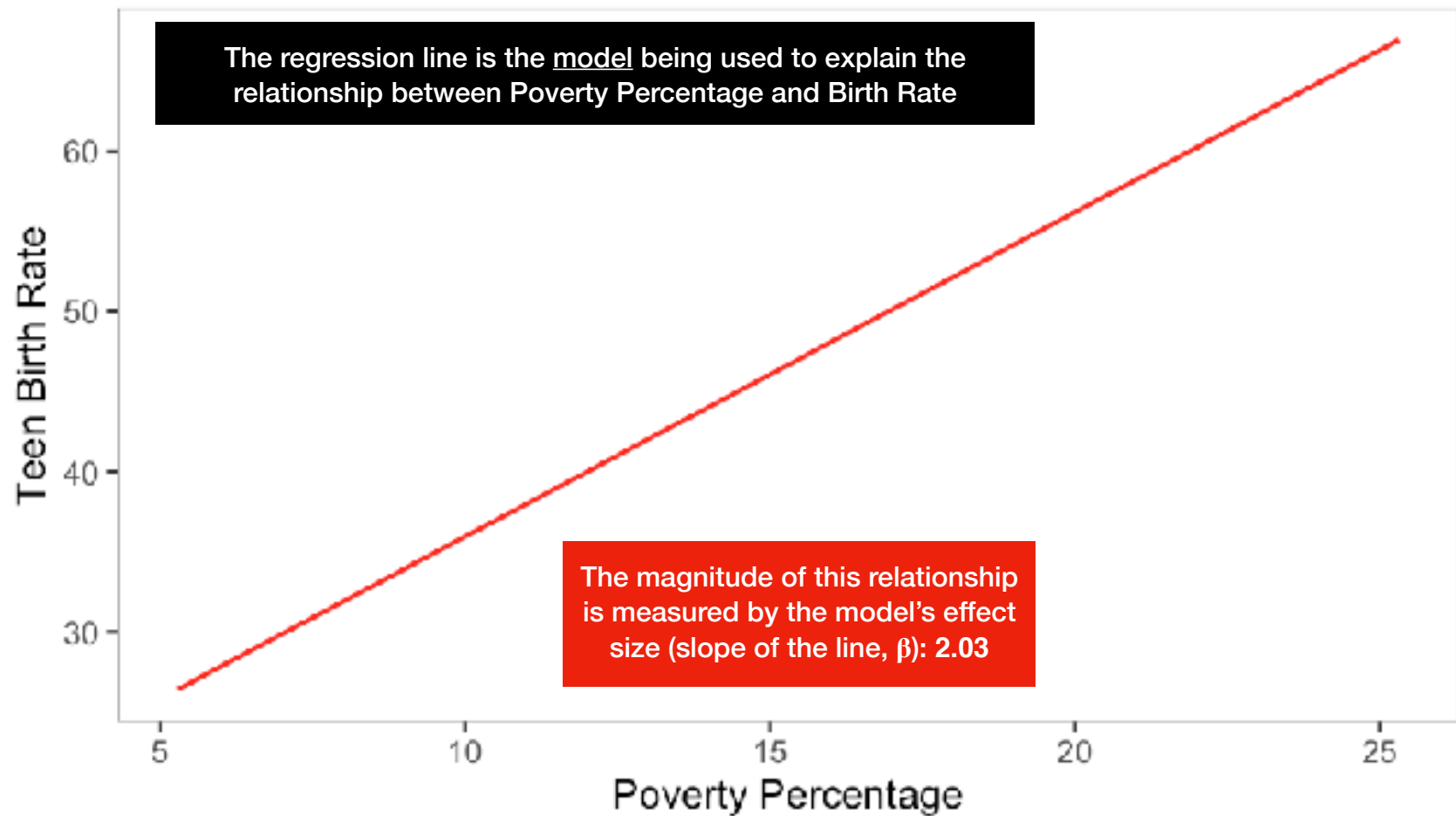
Normal(ish) distributions

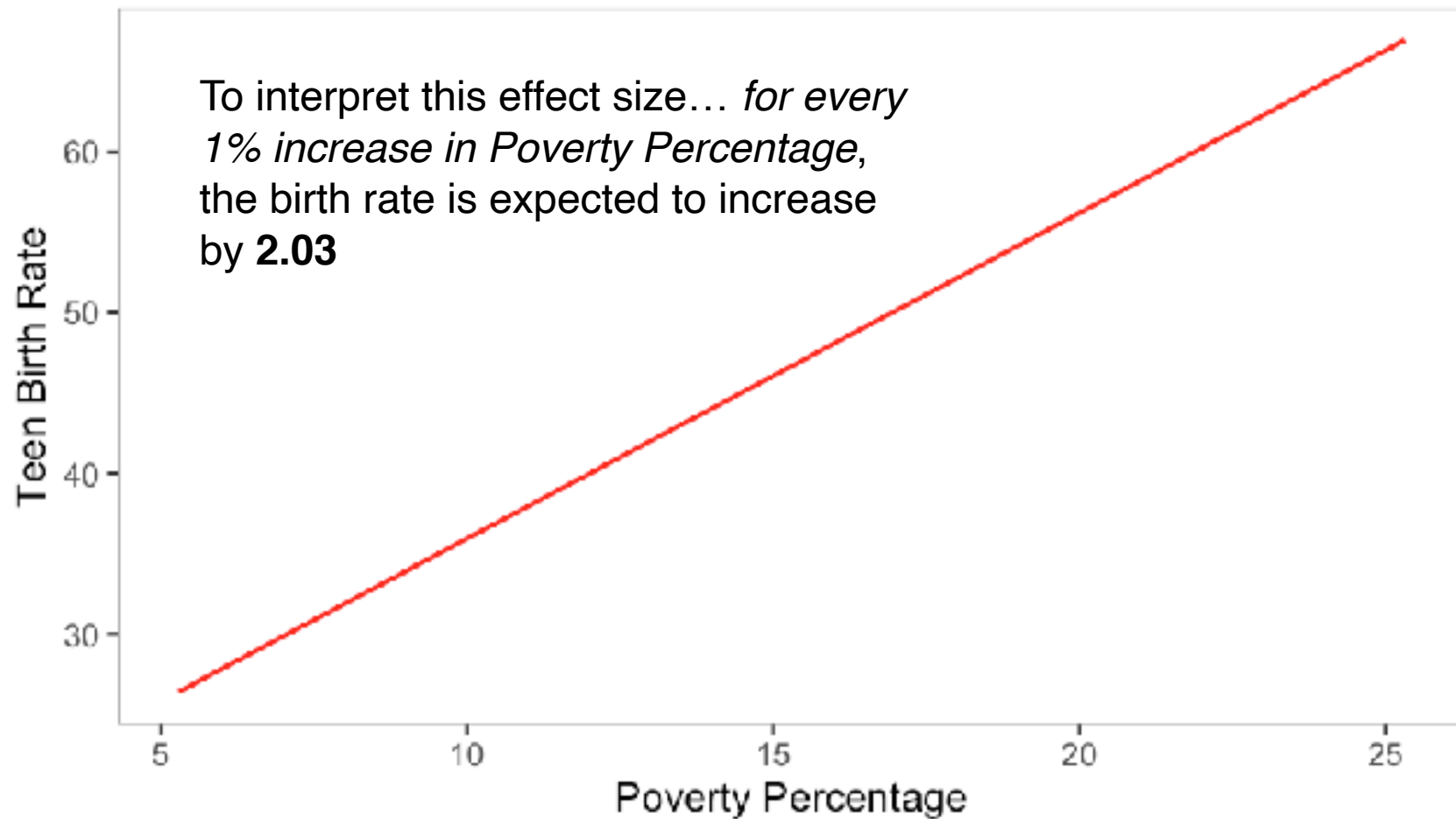






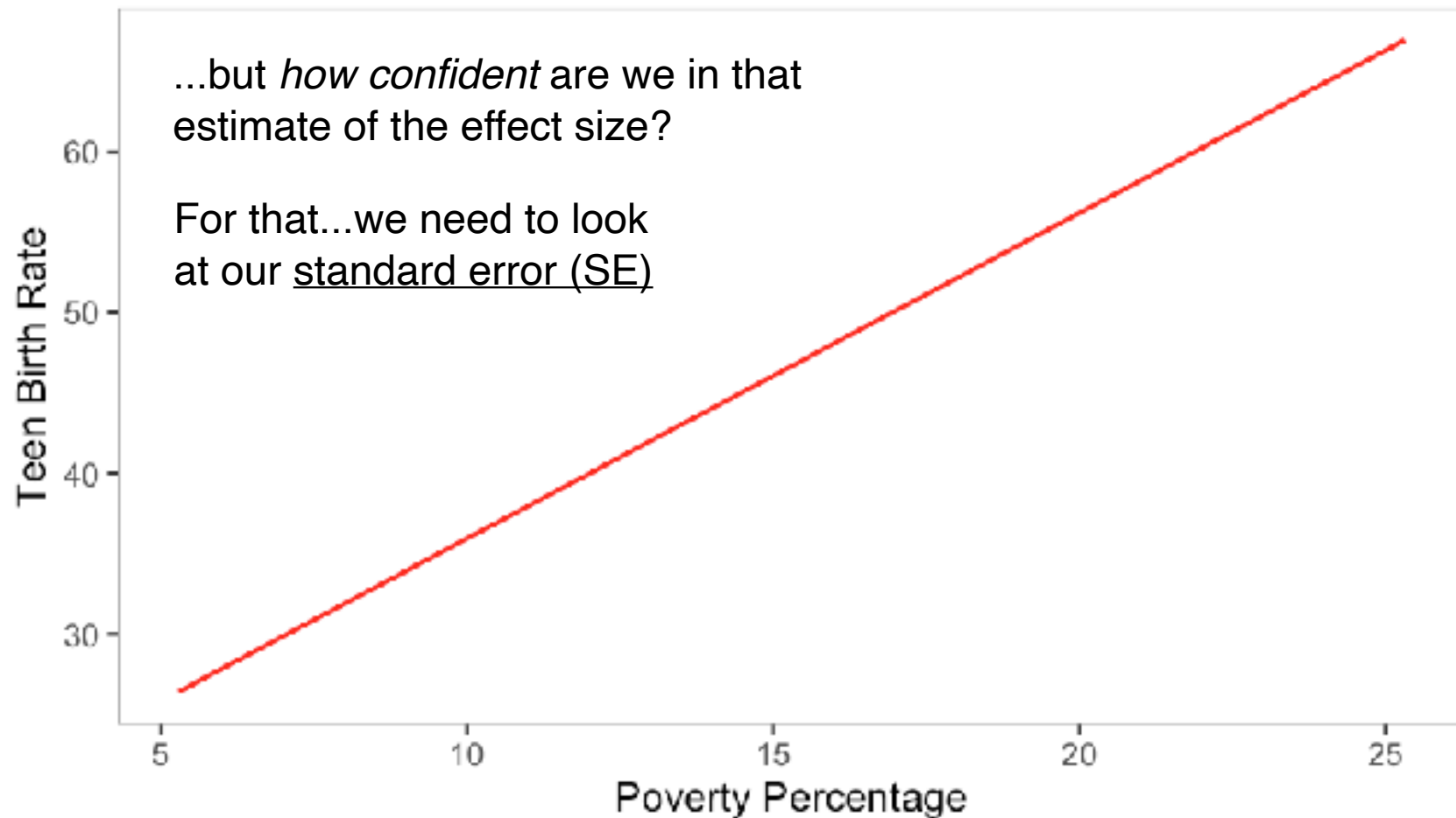


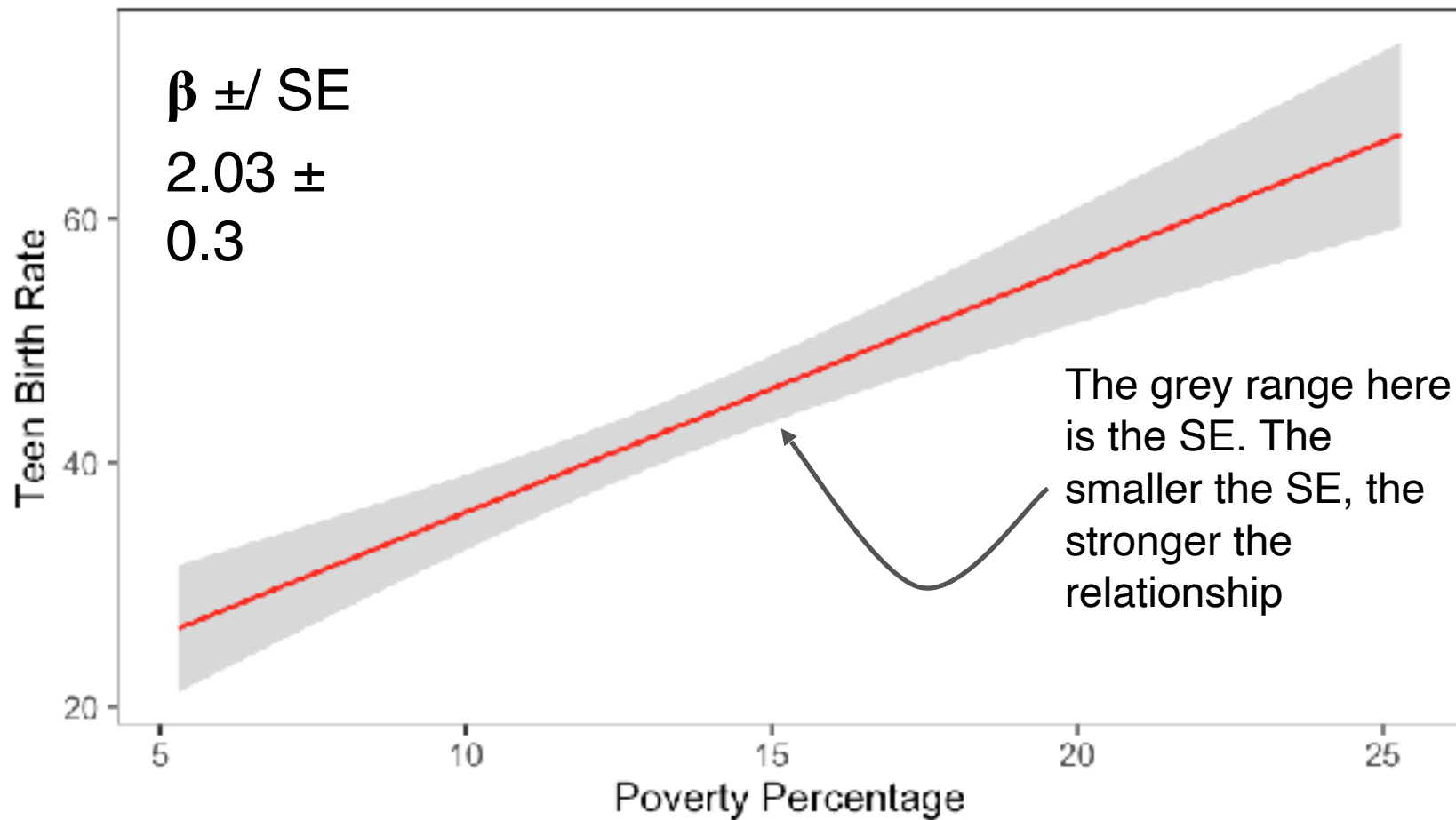


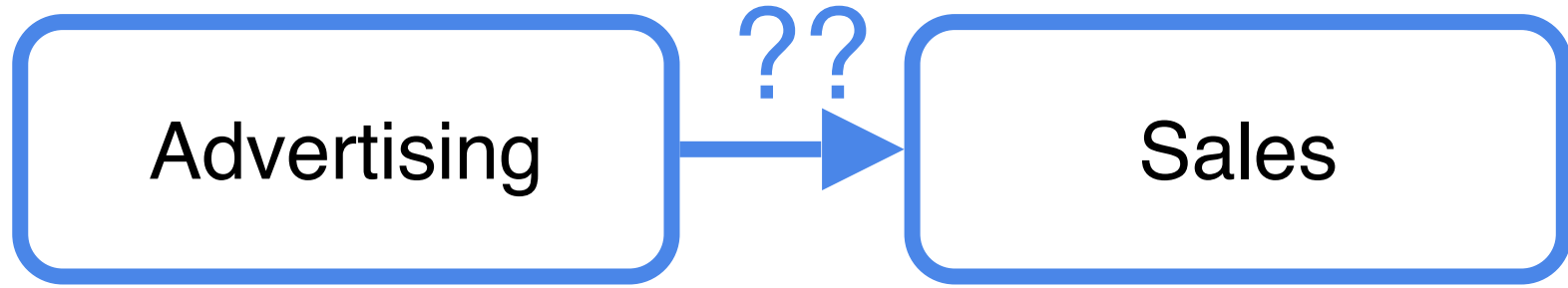


...but *how confident* are we in that estimate of the effect size?

For that...we need to look at our standard error (SE)

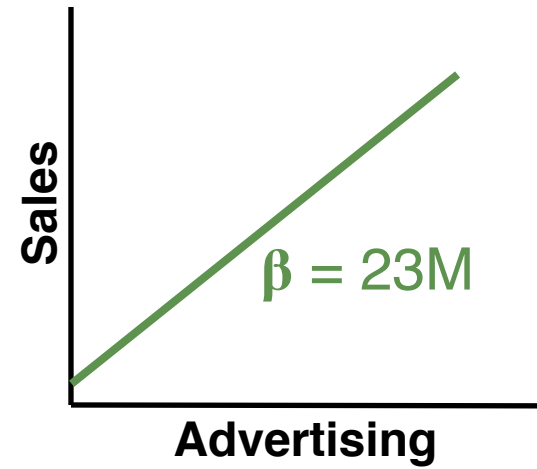






Effect size interpretation

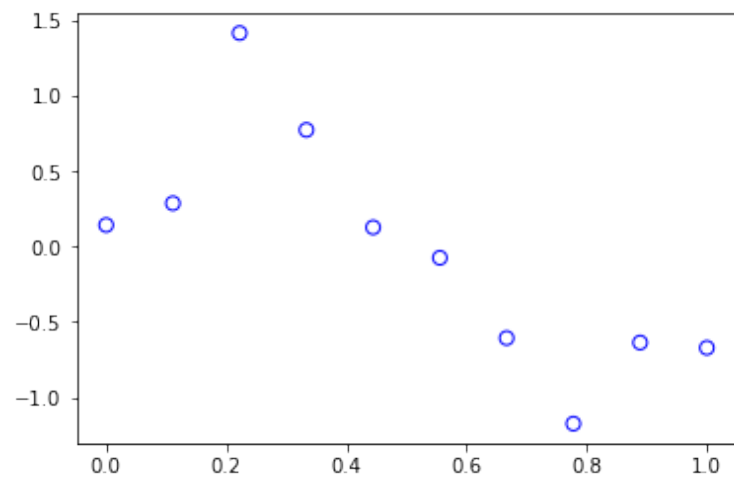
Sales (Million Euro)	Advertising (Million Euro)
651	23
762	26
856	30
1,063	34
1,190	43
1,298	48
1,421	52
1,440	57
1,518	58

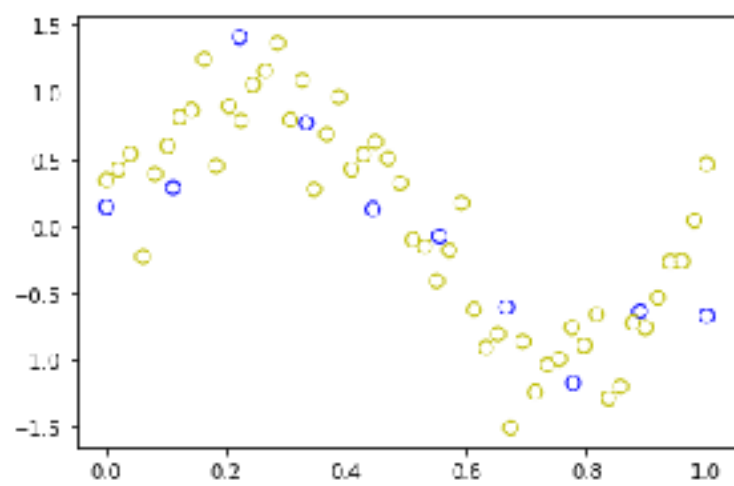


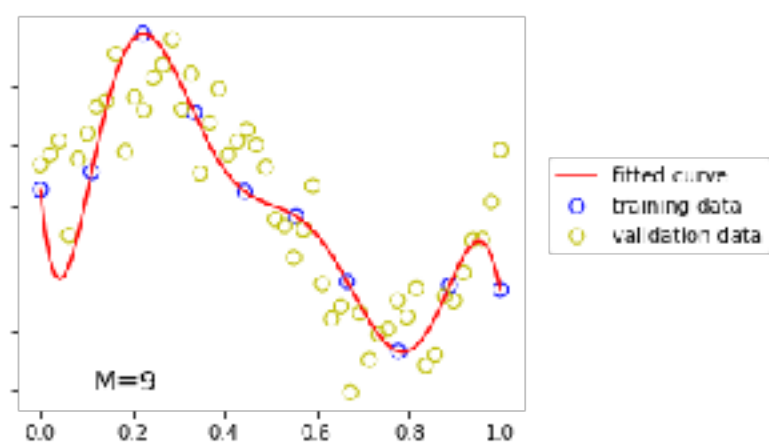
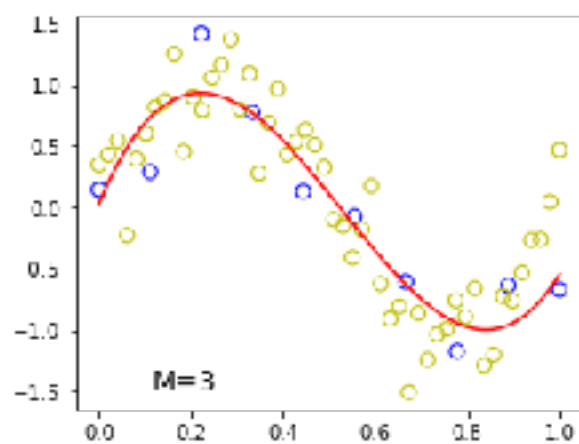
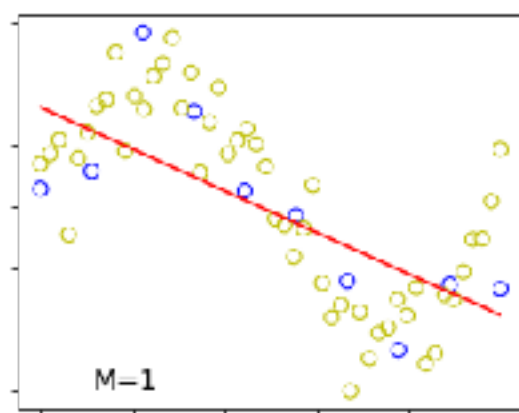
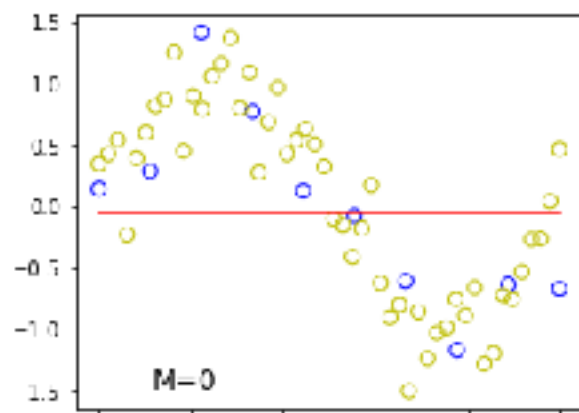
The effect size (β) between the advertising and sales is 23M. What does this mean?



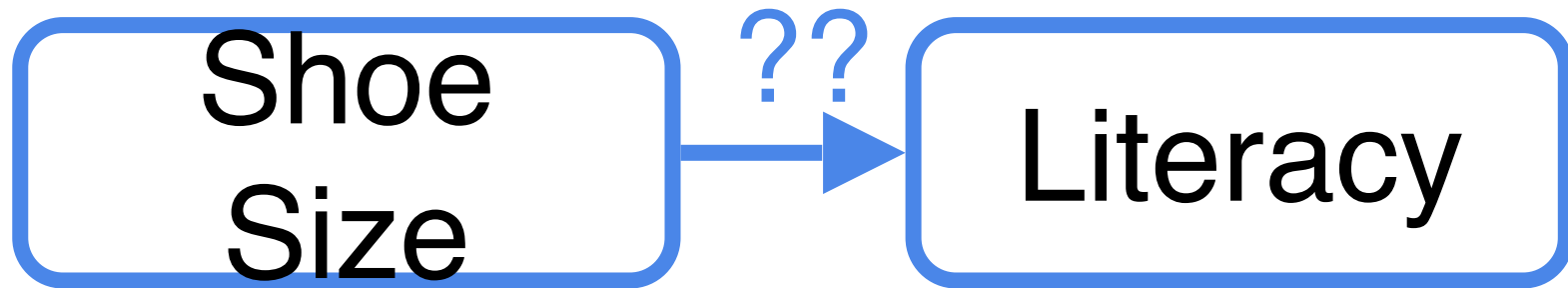
- A For every 1M Euro spent on advertising, the company sees 23M more in sales
- B For every 1M Euro spent in sales, the company spends 23M more in advertising
- C For every 1M Euro spent on advertising, the company sees 24M less in sales
- D For every 1M Euro spent in sales, the company spends 23M less in advertising







Confounding





Small shoes
Not literate
Child

Big shoes
Literate
Adult

Shoe
Size

Literacy

Age

```
graph TD; A[Shoe Size] --> C[Age]; B[Literacy] --> C;
```

The diagram illustrates a relationship where 'Shoe Size' and 'Literacy' are factors that point to 'Age'. 'Shoe Size' and 'Literacy' are contained within solid blue rounded rectangles, while 'Age' is contained within a dashed blue rounded rectangle. Two blue arrows originate from the bottom of the 'Shoe Size' box and the bottom of the 'Literacy' box, converging at the top of the 'Age' box.

Variable1

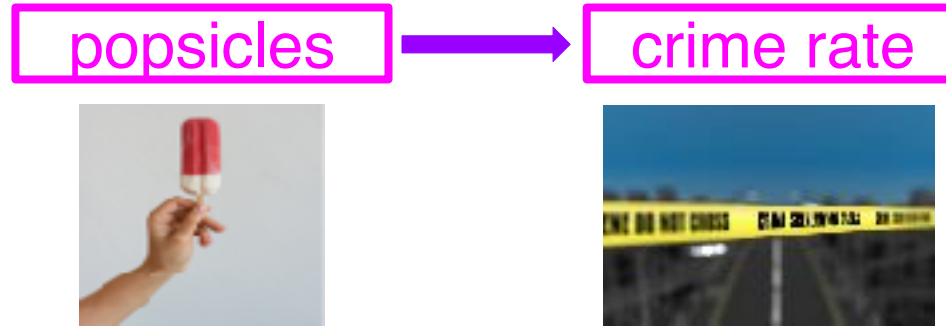
Variable2

Confounder

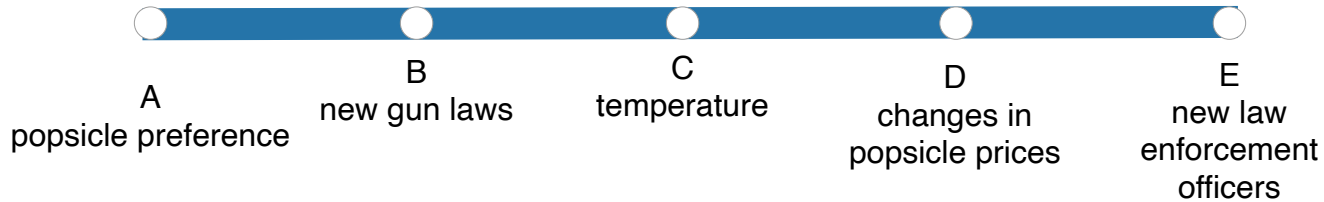
```
graph BT; C[Confounder] --> V1[Variable1]; C --> V2[Variable2];
```

The diagram illustrates a causal relationship where a confounder, represented by a dashed blue box at the bottom, influences two separate variables. Two solid blue arrows point upwards from the top of the dashed box to the bottom of the 'Variable1' and 'Variable2' boxes, which are outlined in solid blue. The entire diagram is set against a white background with a grey horizontal bar at the bottom.

Confounding



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

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

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

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

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