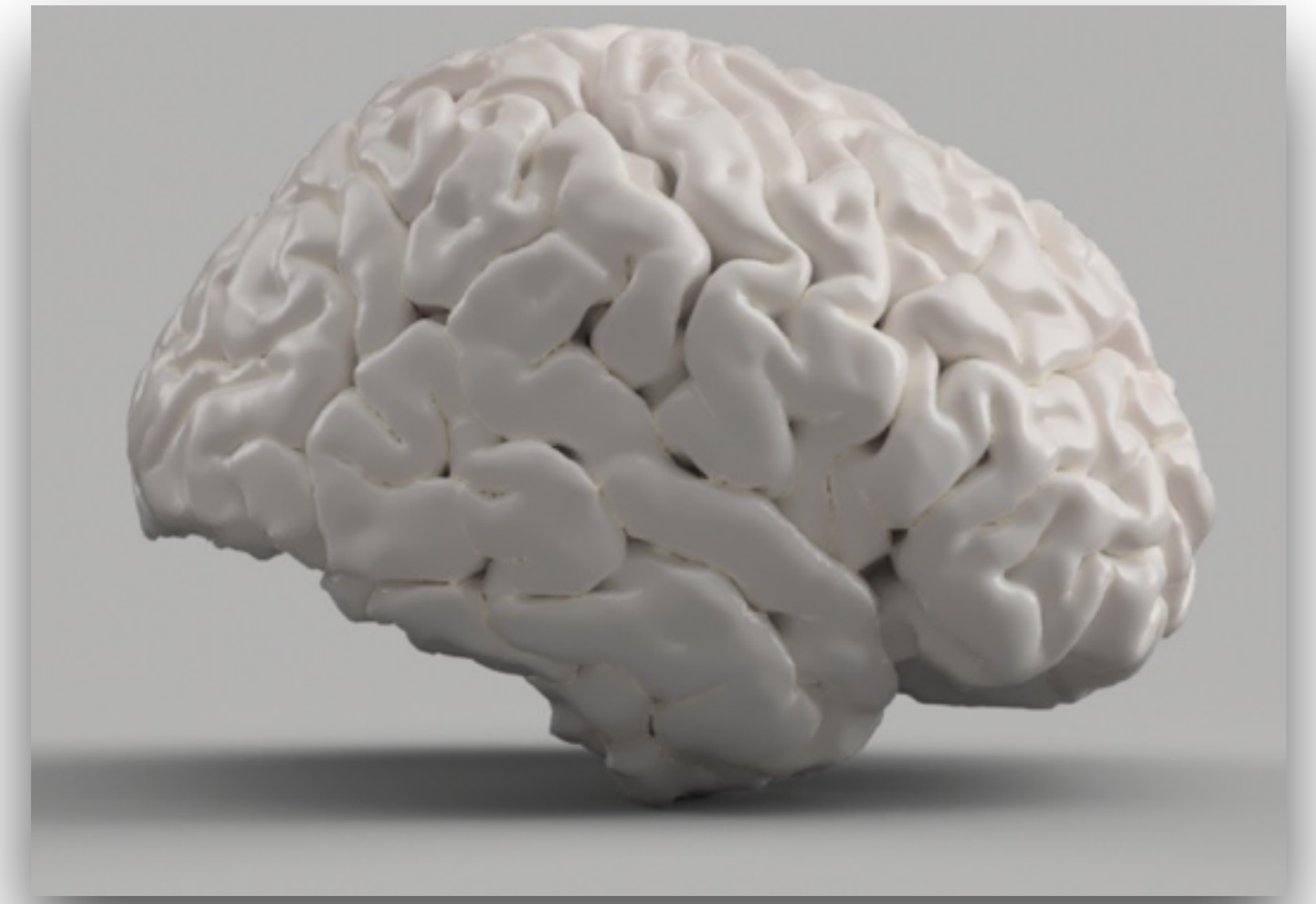
COGS109: Lecture 9



Hypothesis testing

July 20, 2023

Modeling and Data Analysis

Summer Session 1, 2023

C. Alex Simpkins Jr., Ph.D.

RDPRobotics LLC | Dept. of CogSci, UCSD

Plan for today

- Announcements
- Inference using central tendency and variability concepts
- Hypothesis testing
- Confounds

Plan for today II

- Hypothesis testing
- Introduction to models and the modeling process
- Colormaps - custom

Announcements

- Q2 posted
- D4 released
- Github repos
- In process on feedback and A1

Update: the big picture

- Where we are
  - 5 parts of the course
    - We discussed data
      - What is it, how do we manipulate it, import it, python and some matlab implementation
      - Filtering
      - Computing basic statistics
    - We discussed basic visualization
      - Plotting data (2d, histograms, scatterplots, etc)

Update: the big picture (II)

- Where we’re going
  - We will now cover
    - Modeling
      - what is modeling?
      - interpolation, approximation, extrapolation
    - Error analysis
      - How good is your model?

Update: the big picture (III)

- Where we're going (continued)
  - **What we're going to cover**
    - Basic models
      - **Linear fits, nonlinear fits**
      - **Regression**
      - **Relationship to machine learning**
      - **Interpolation/extrapolation (also data analysis methods)**
    - Advanced models and modeling methods
      - **Fitting models with optimization methods**
      - **Artificial neural networks**
      - **AI**
    - Communicating results
      - **This has been integrated and will continue to be integrated**
      - **Proper forms of inserting figures and tables in scientific communications**
      - **Format in homeworks is designed to teach proper communication methodology**

Extending central tendency and variability to inference and hypothesis testing

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

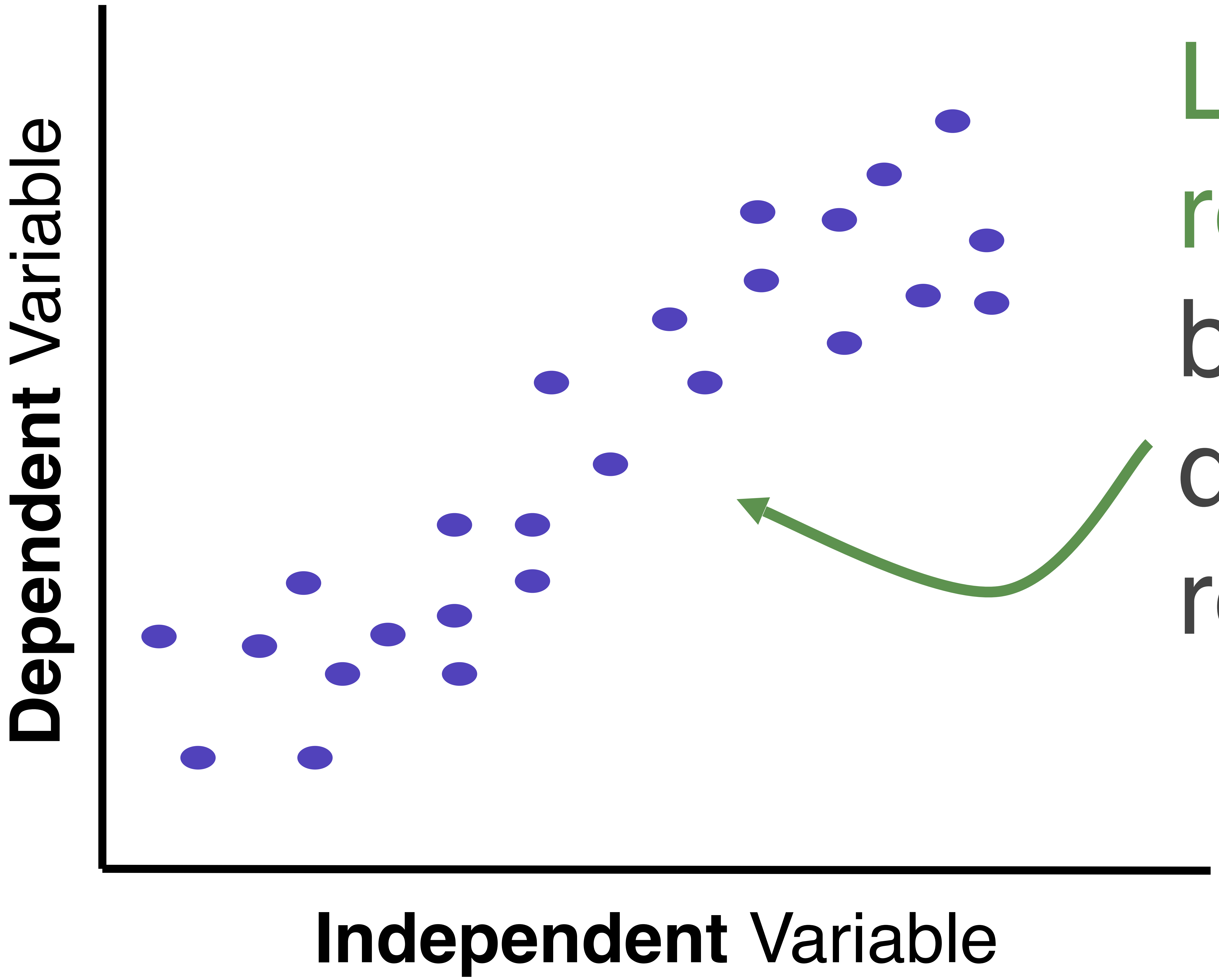
DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

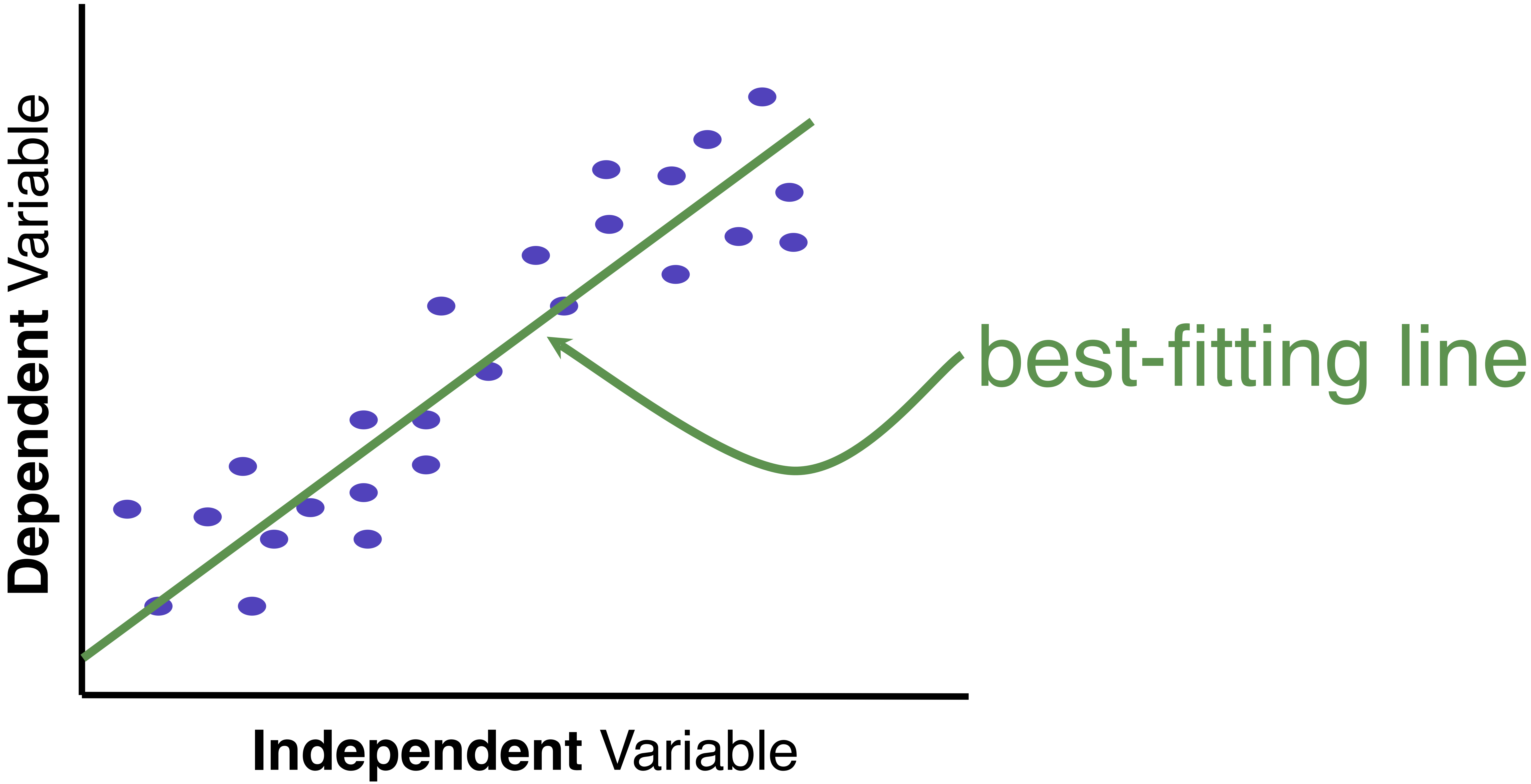
NON-PARAMETRIC TESTS

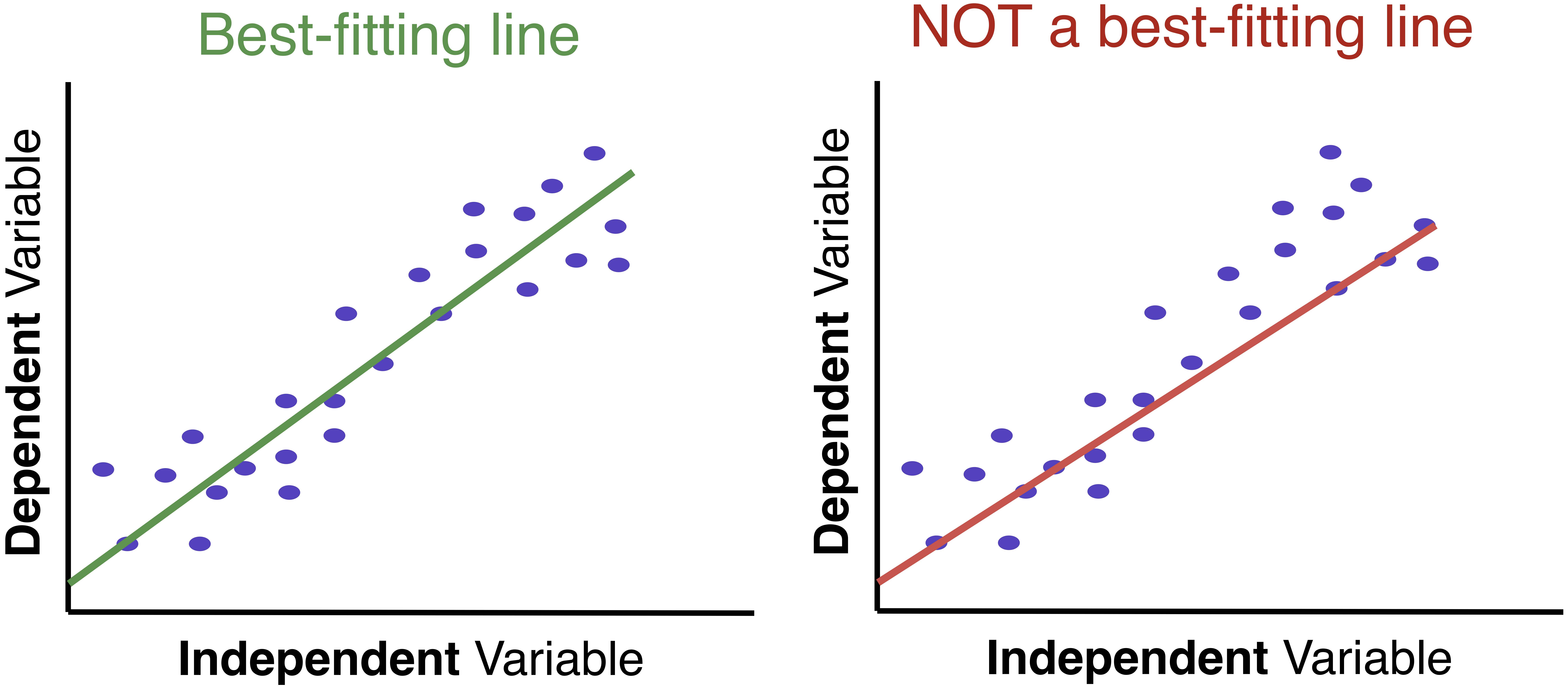
FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

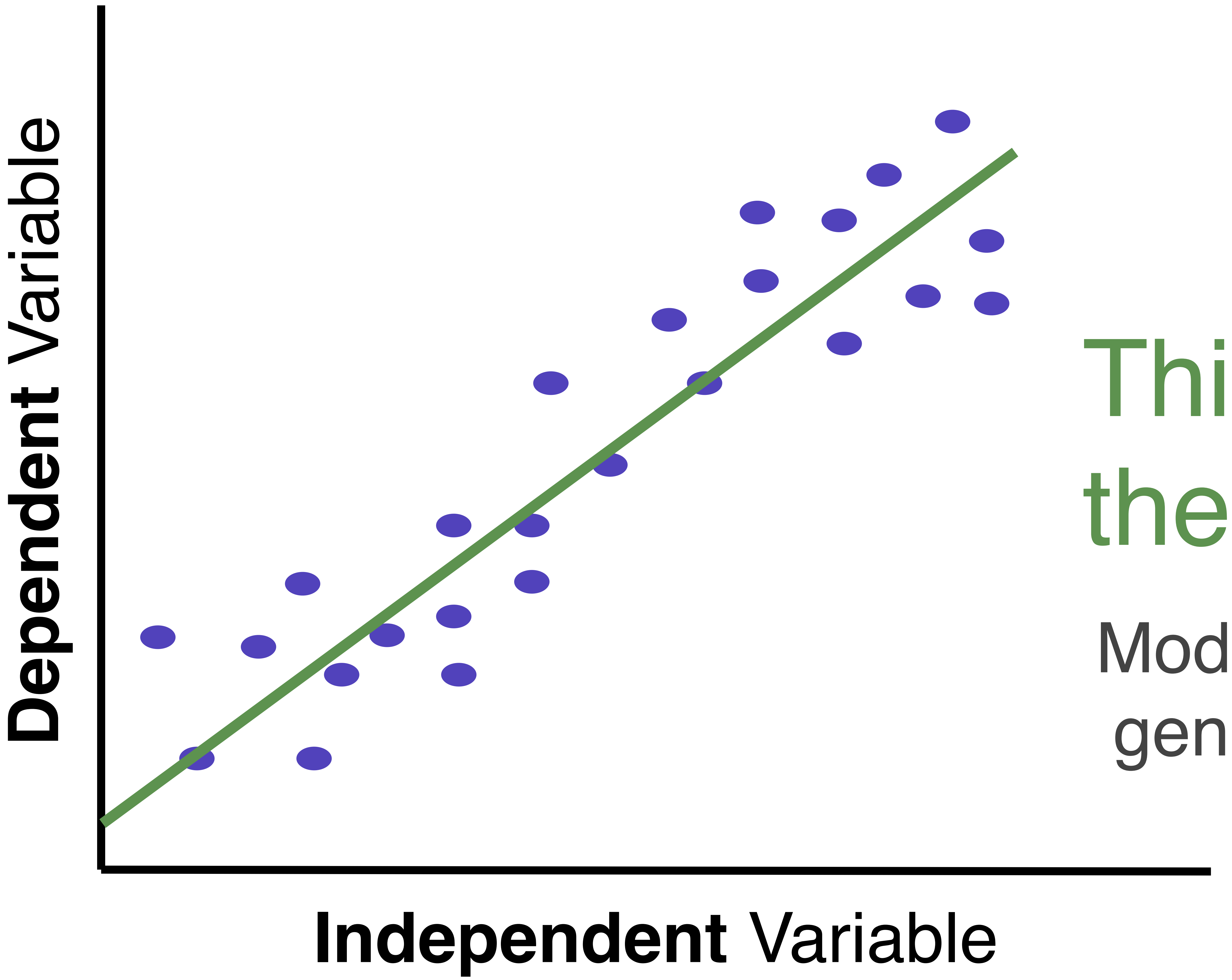
I.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test



Linear regression can be used to describe this relationship

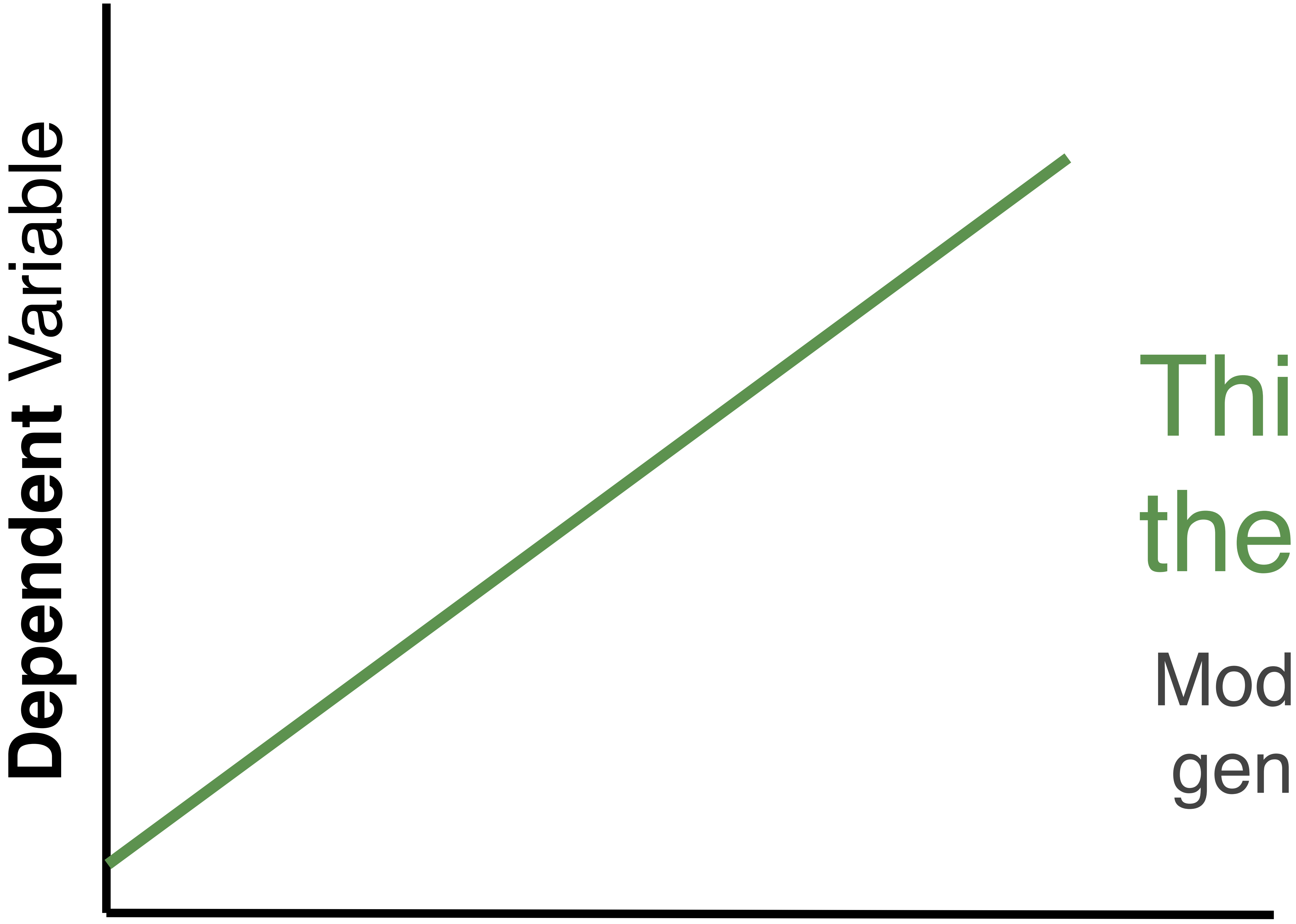

Dependent Variable
best-fitting line
Independent Variable


Best-fitting line
Dependent Variable
Independent Variable
NOT a best-fitting line
Dependent Variable
Independent Variable



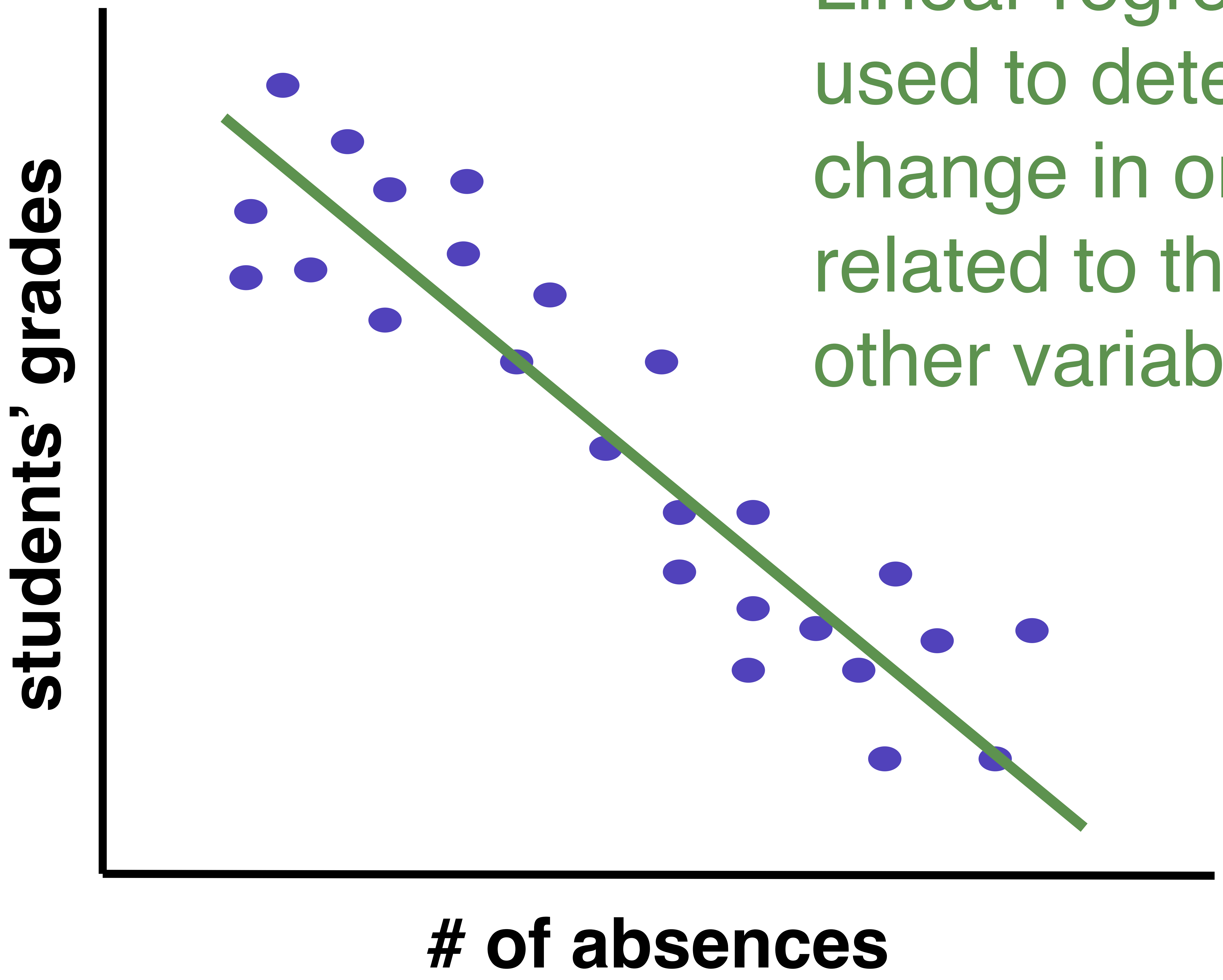
This line is a model of the data

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

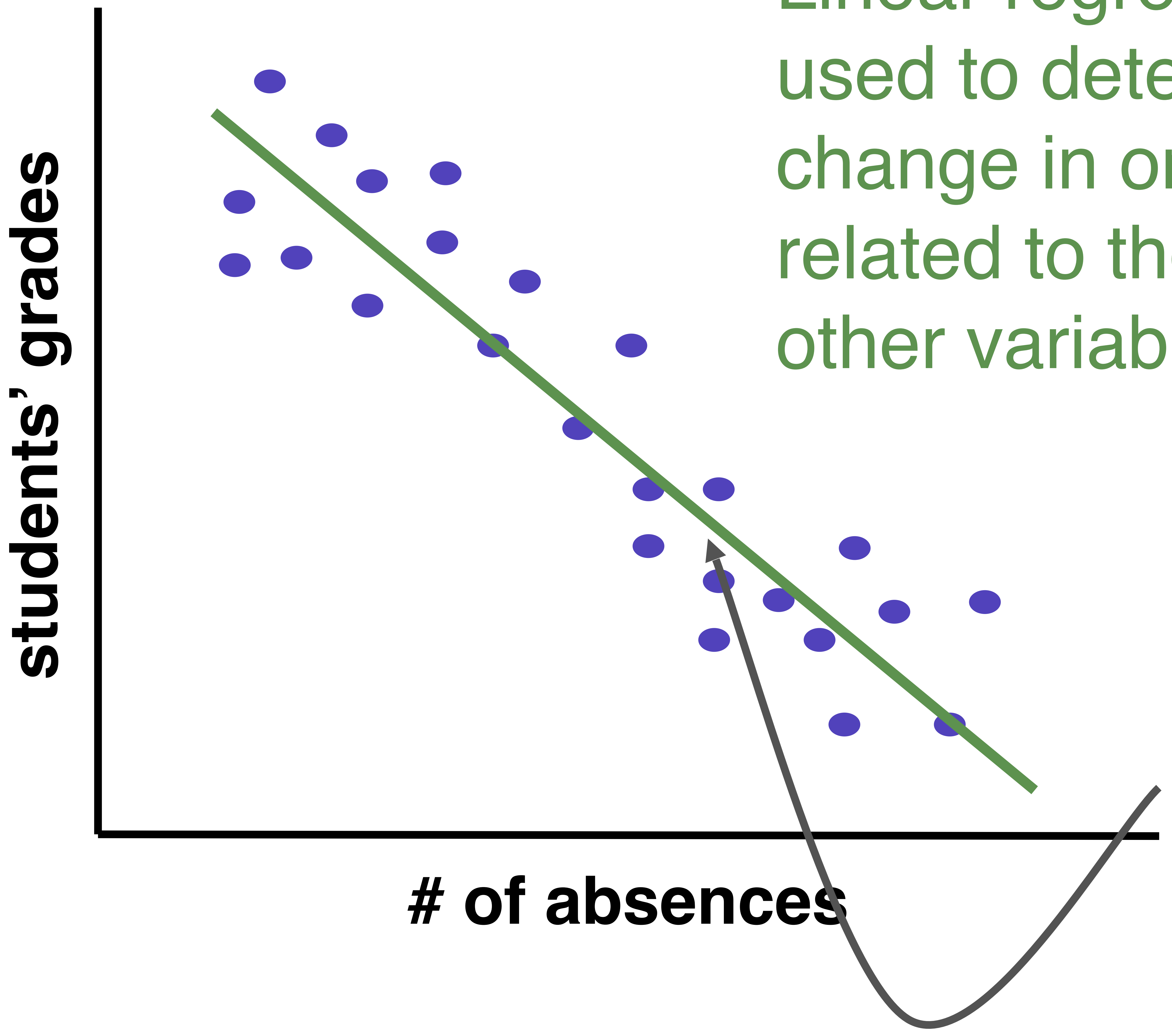


This line is a model of the data

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



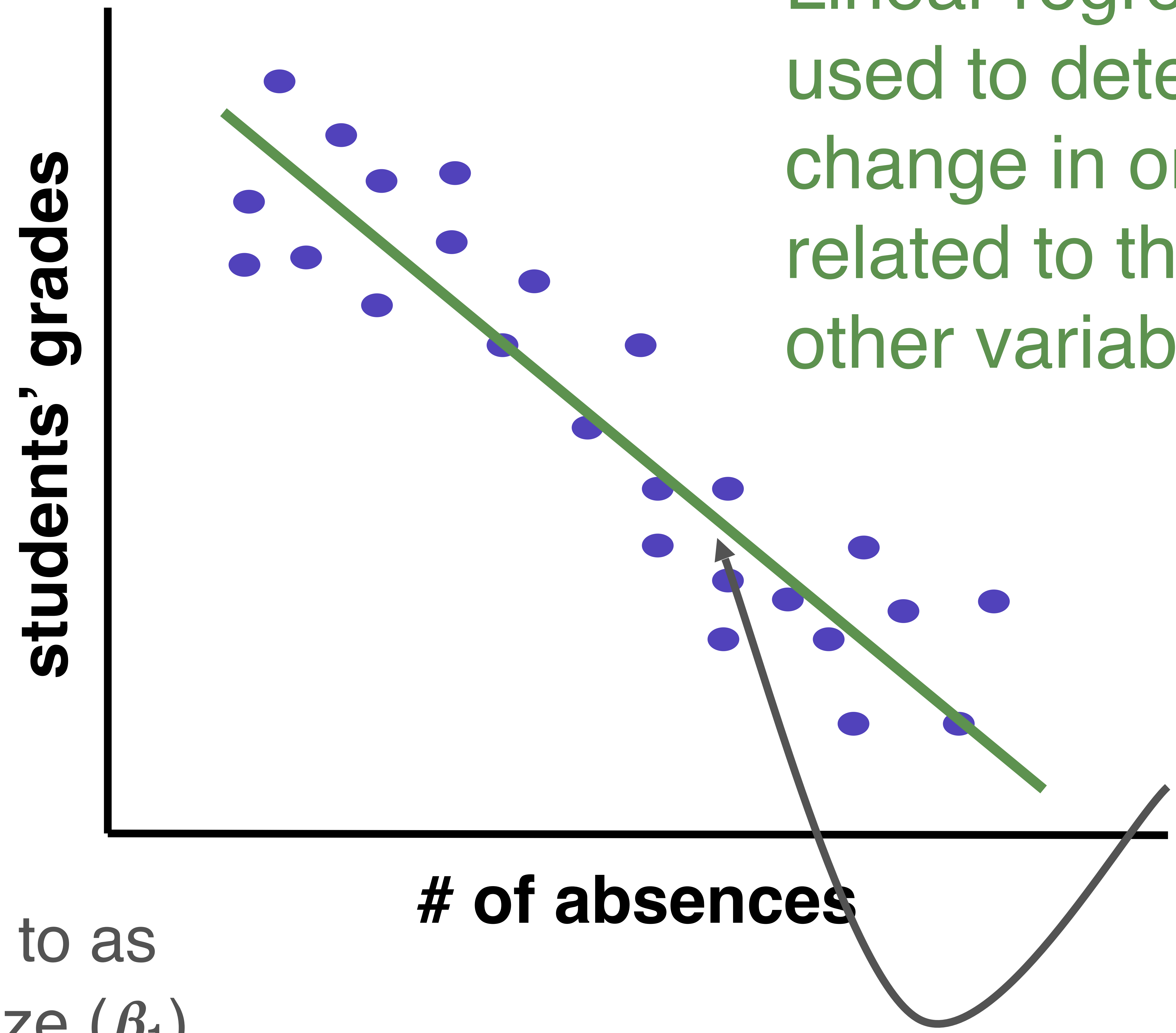
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



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

The magnitude of the relationship is measured by the slope of the line

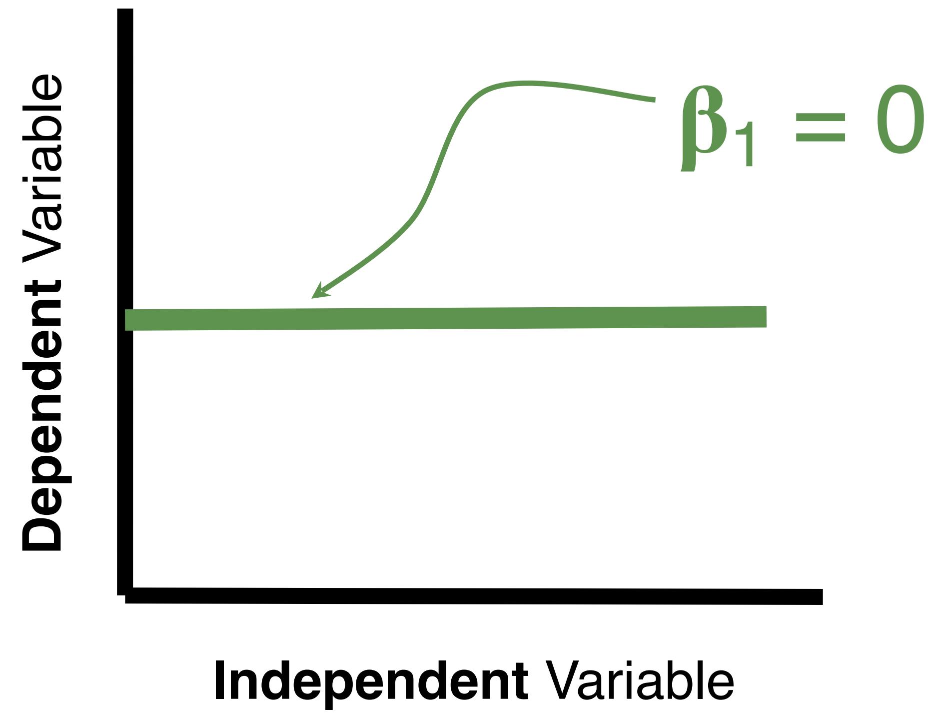
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



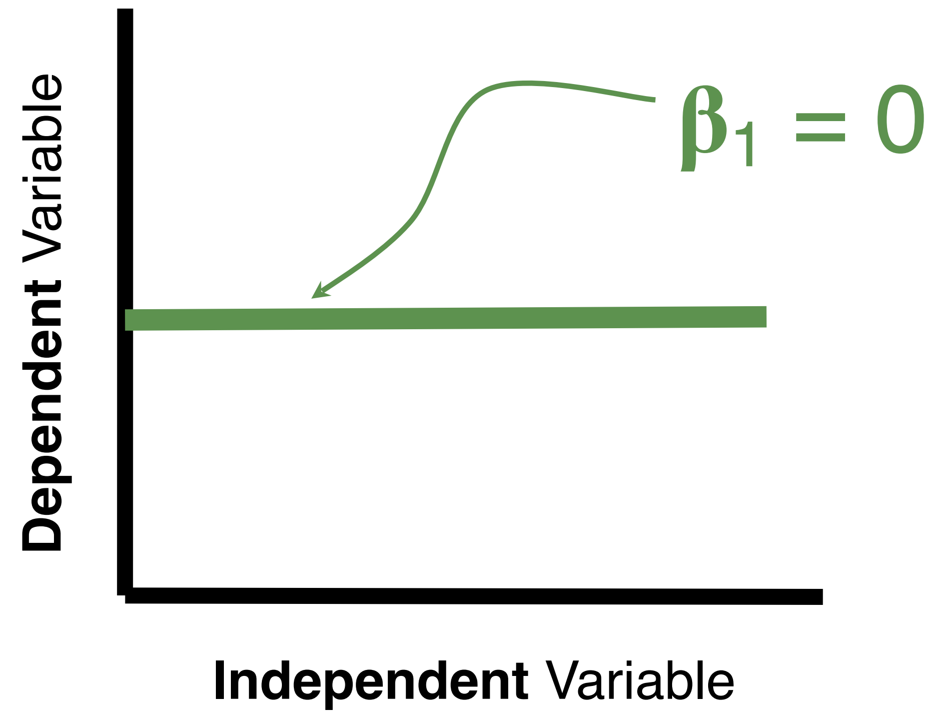
The magnitude of the relationship is measured by the slope of the line

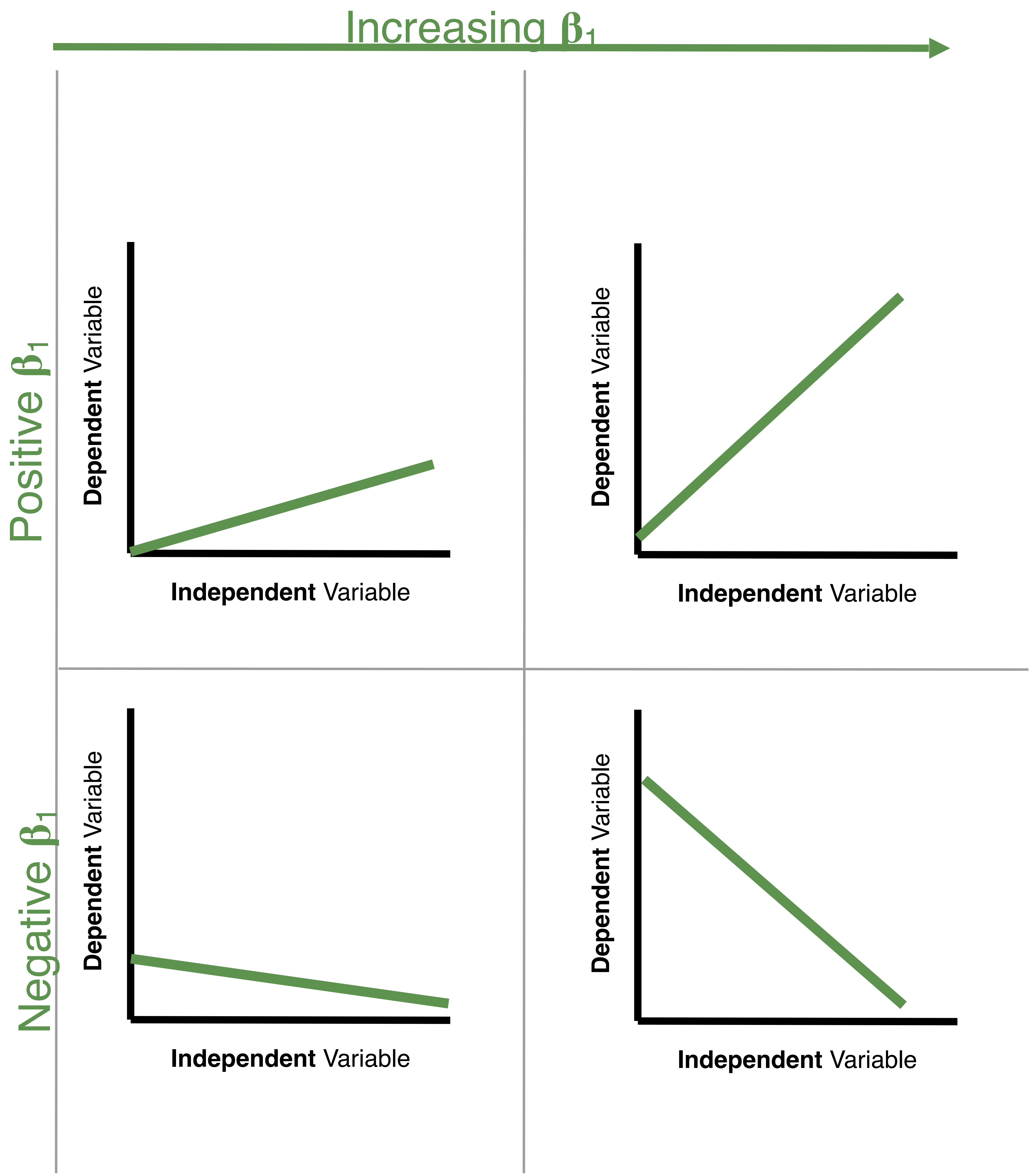
This is also referred to as the model's effect size (β_1)

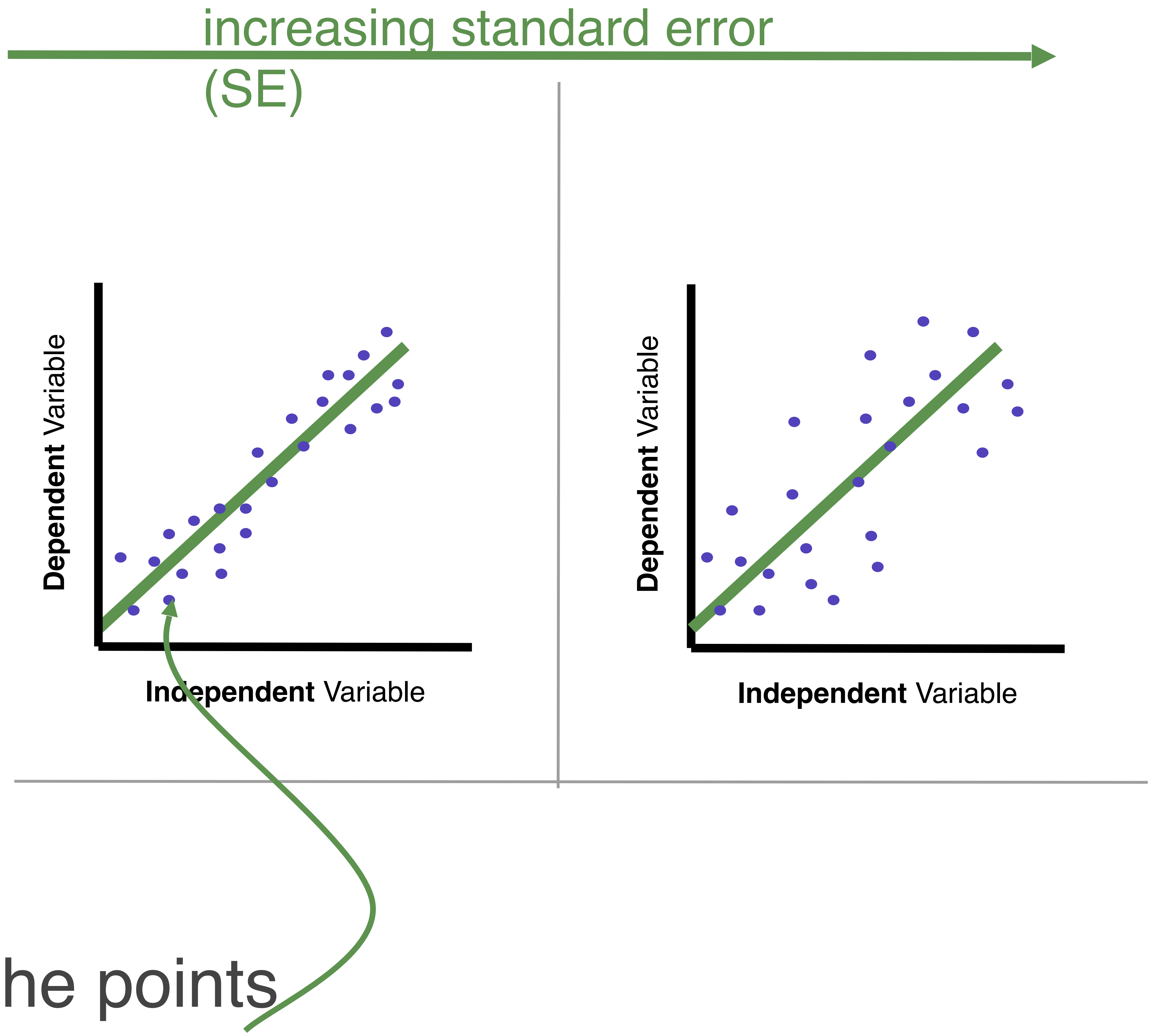
Effect size (β_1) can be estimated using the slope of the line



Effect size (β_1) can be estimated using the slope of the line





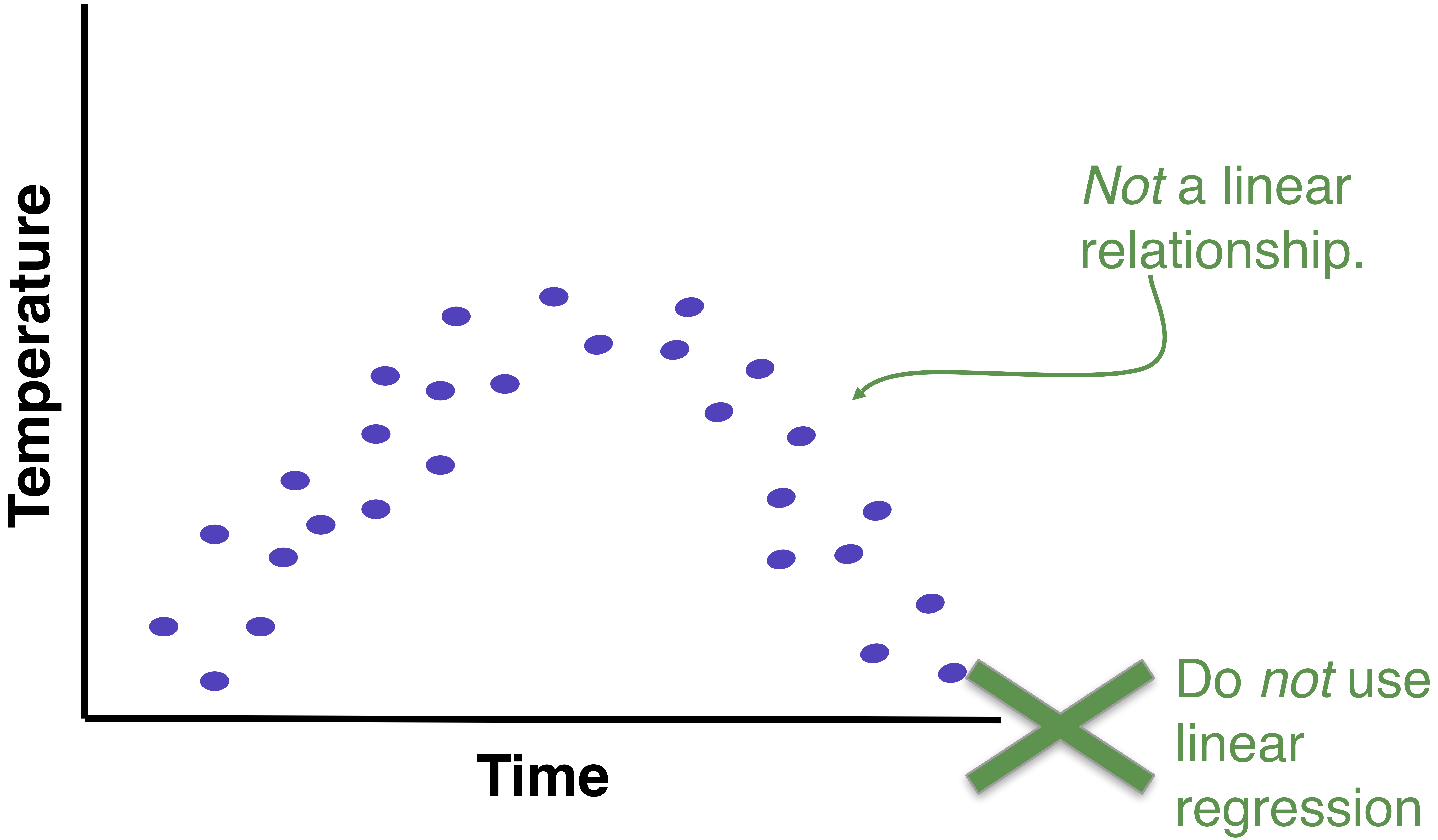


The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

Assumptions of linear regression

1. Linear relationship
2. No multicollinearity
3. No auto-correlation
4. Homoscedasticity

Linearity


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

Multicollinearity

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

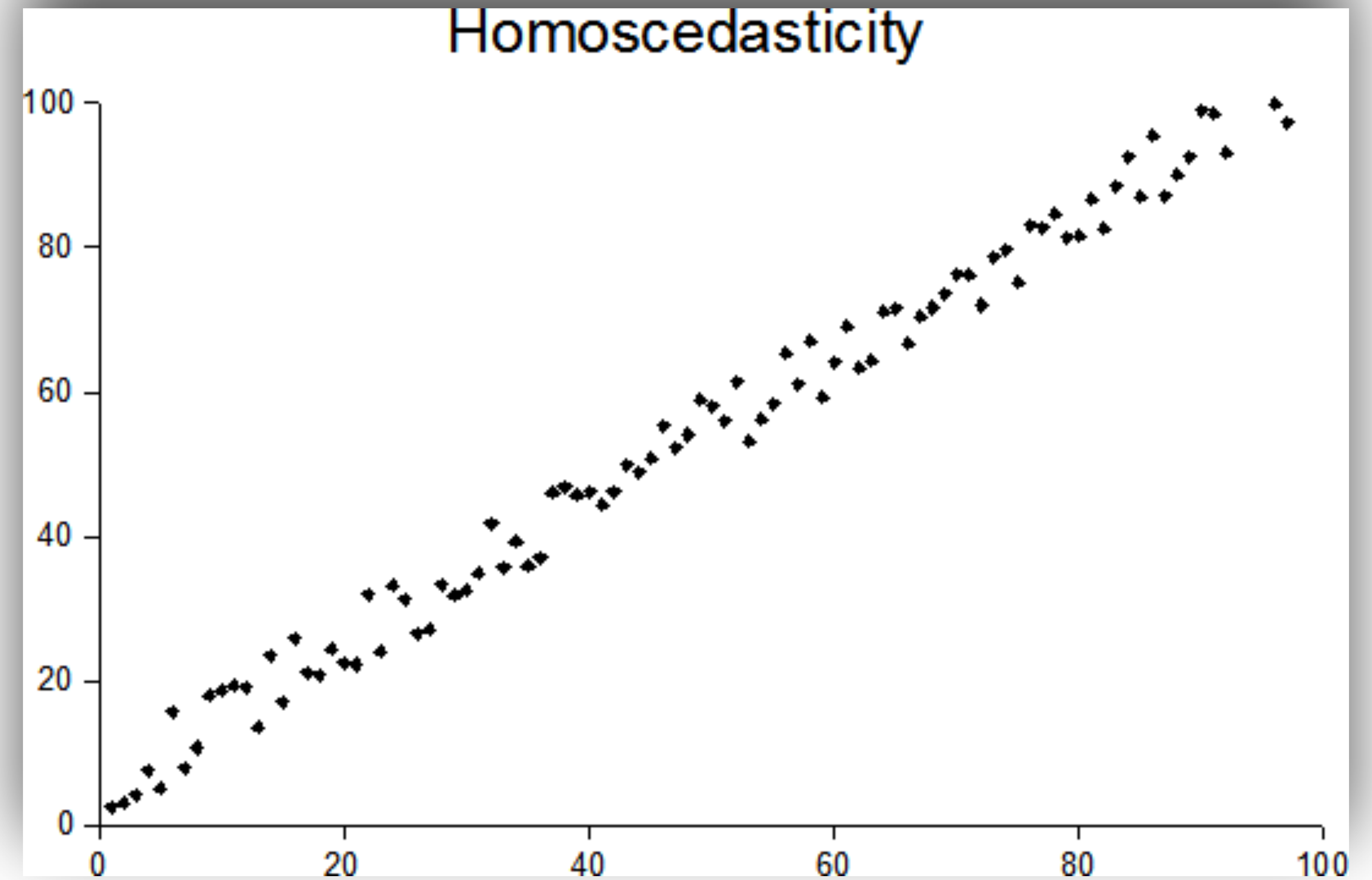
- 2 variables are perfectly correlated if they have a correlationn coefficient of 1.0

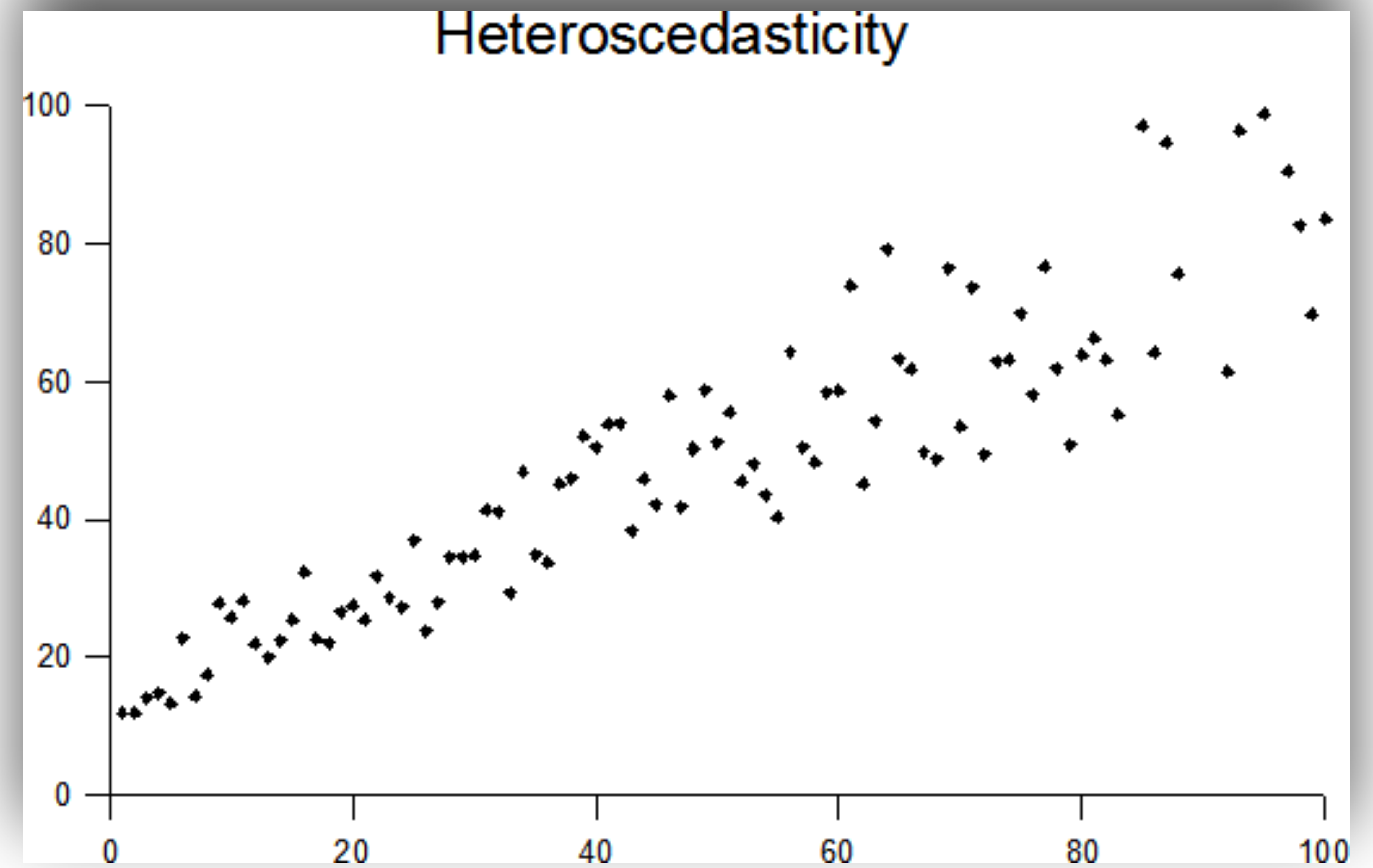
Autocorrelation

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

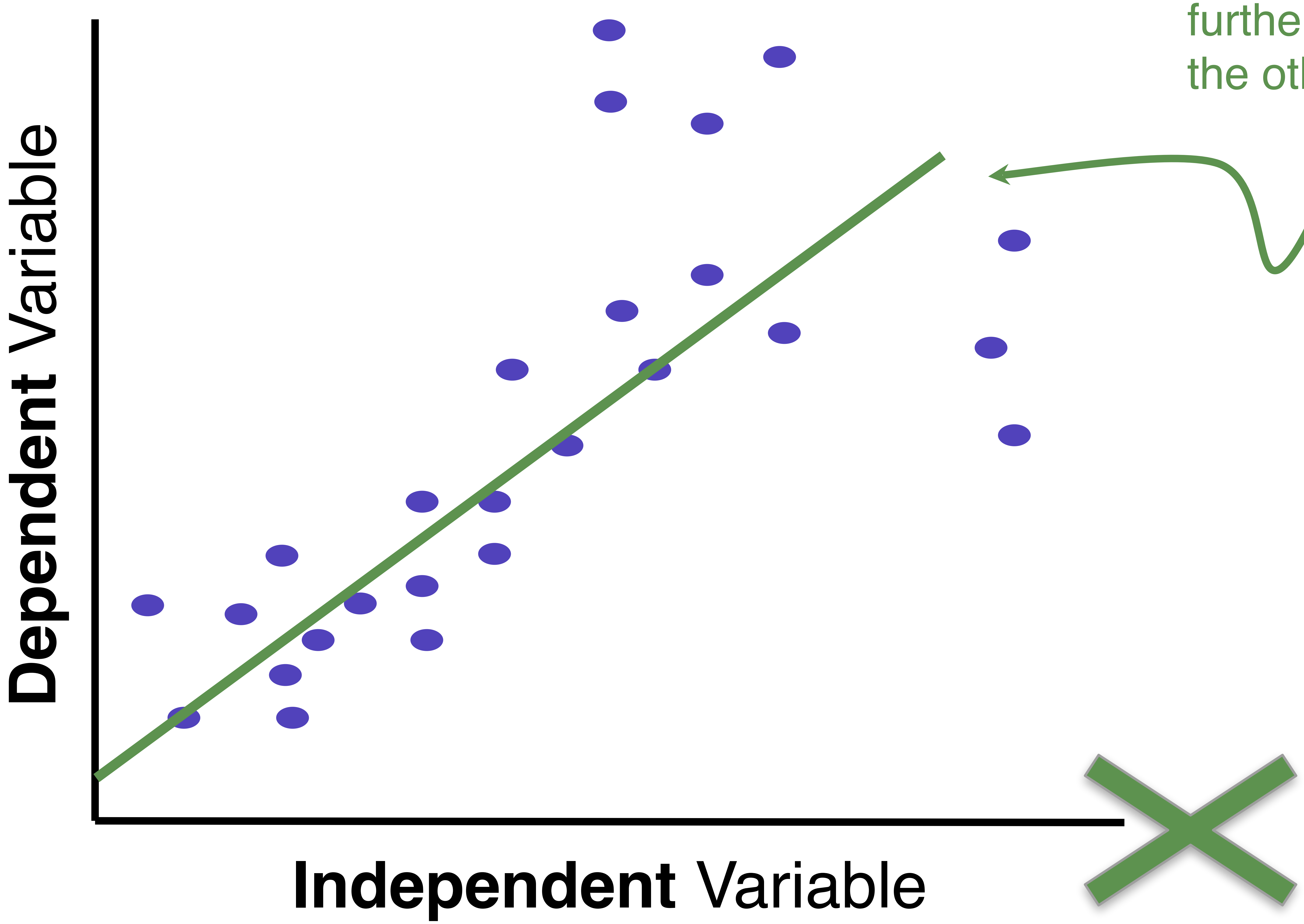
https://www.statisticssolutions.com/assumptions-of-linear-regression/

Homoscedasticity - a reminder of what that is





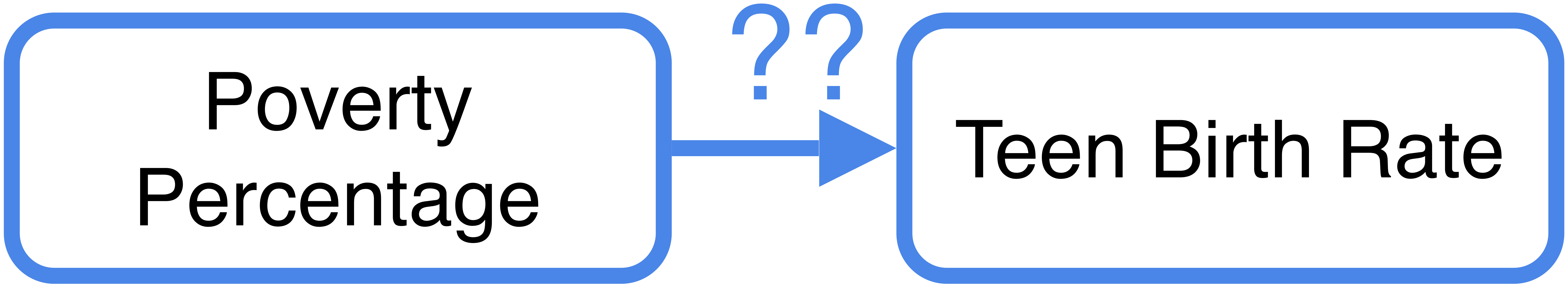
Homoscedasticity



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

Do *not* use linear regression

Does Poverty Percentage affect Teen Birth Rate?



<u>Null Hypothesis</u>:

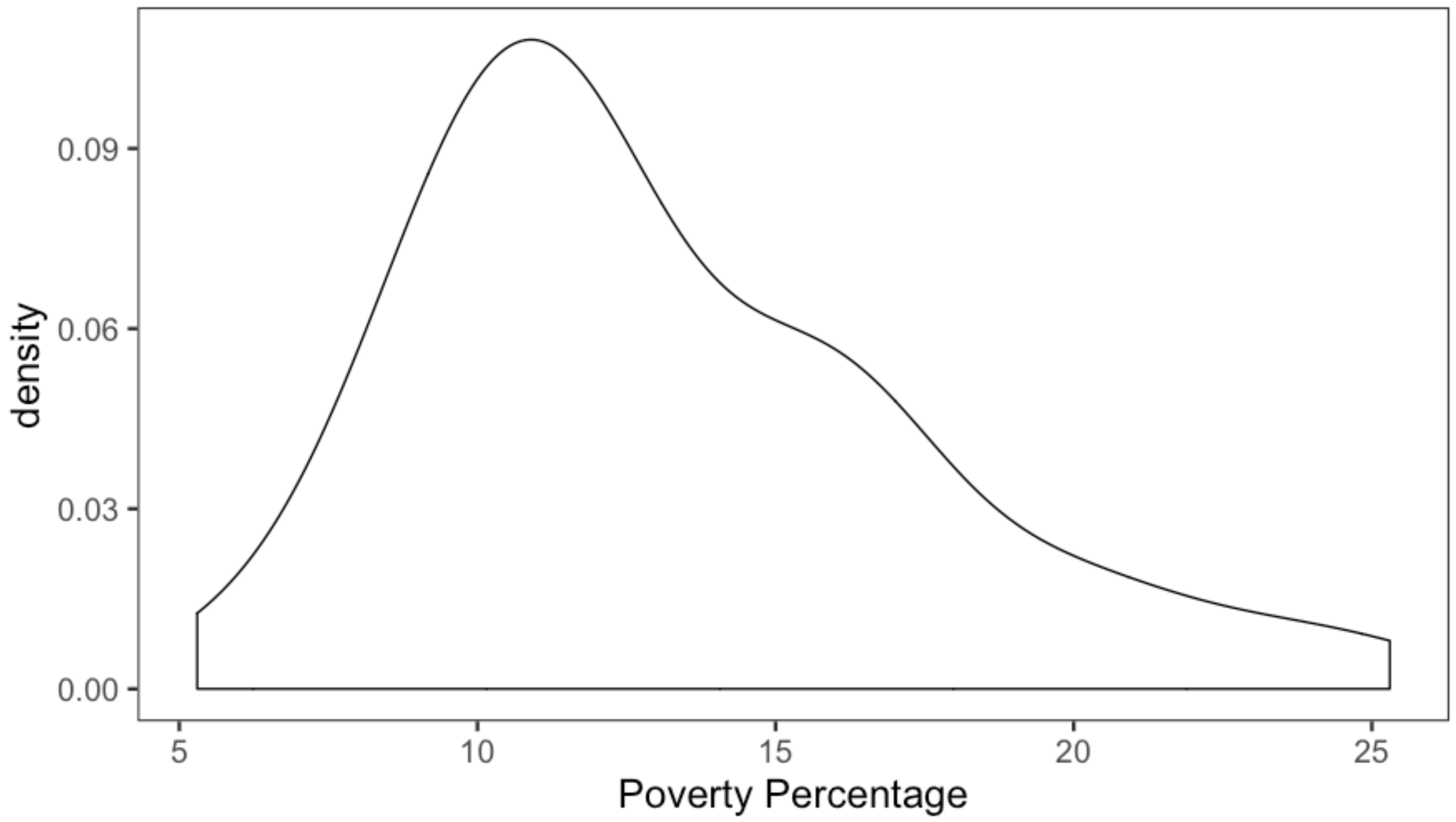
H_0: Poverty Rate does not affect Teen Birth Rate (β_1=0)

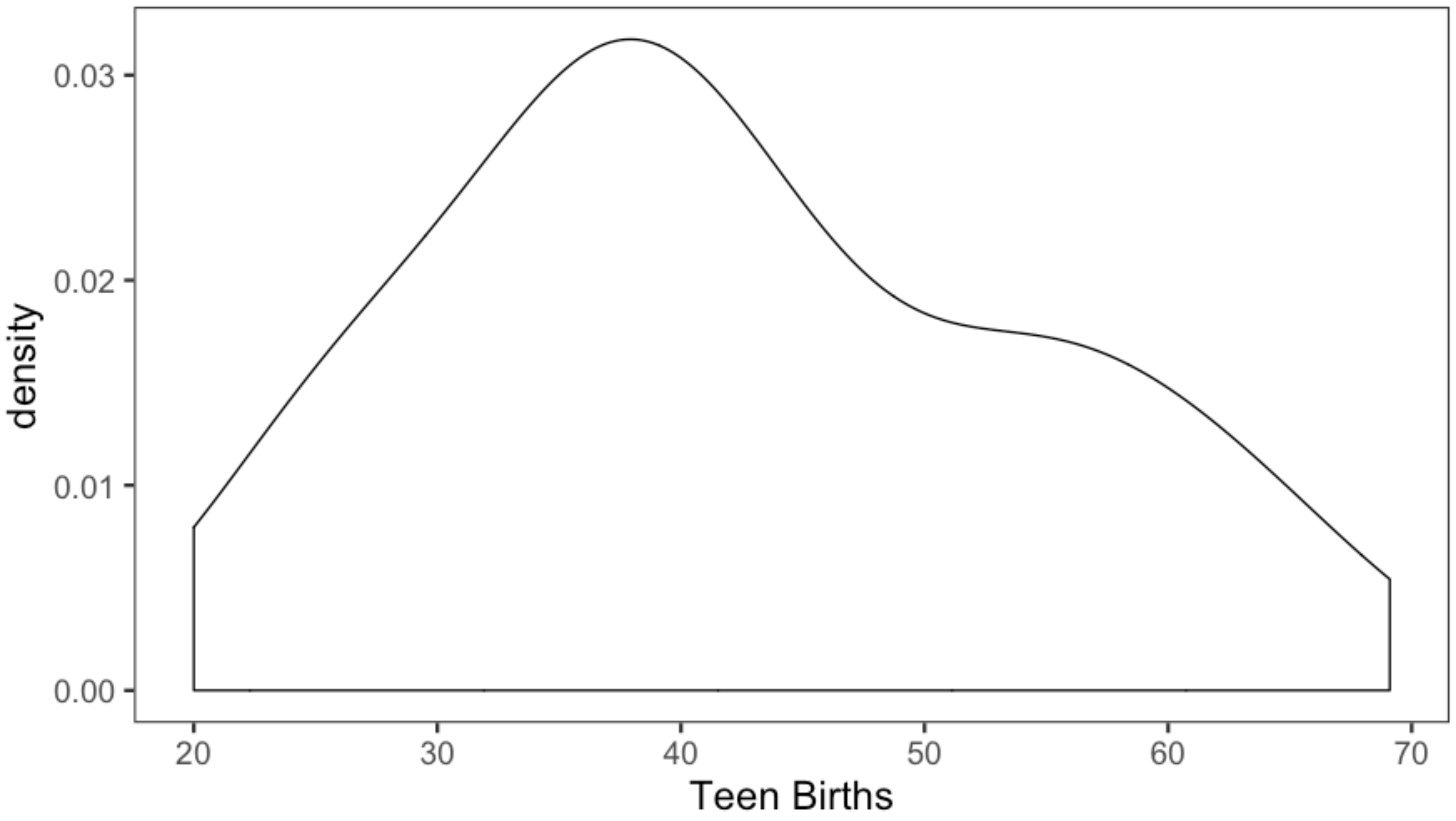
<u>Alternative Hypothesis</u>:

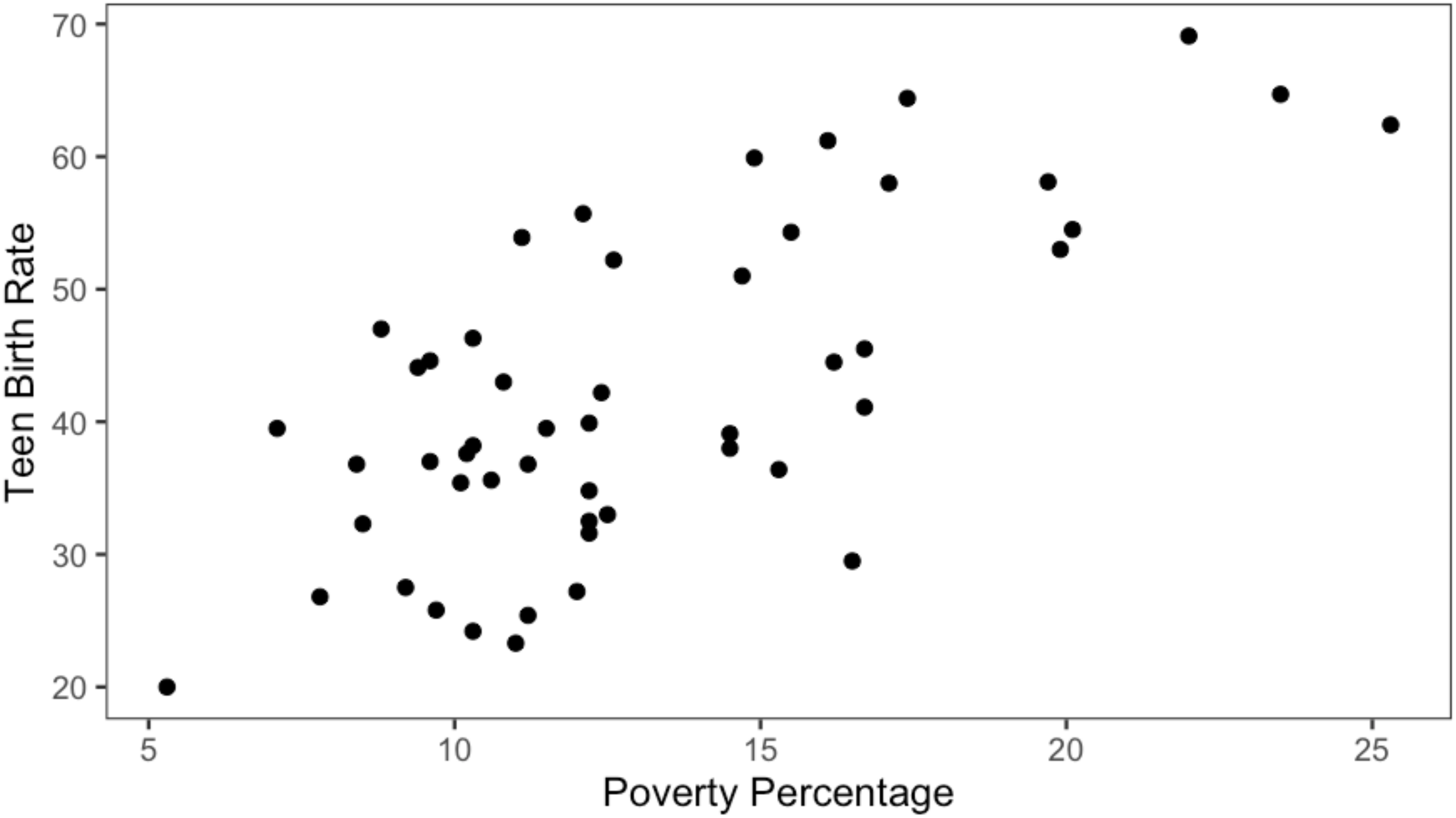
H_a: Poverty Rate affects Teen Birth Rate ($\beta_1 \neq 0$)

	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

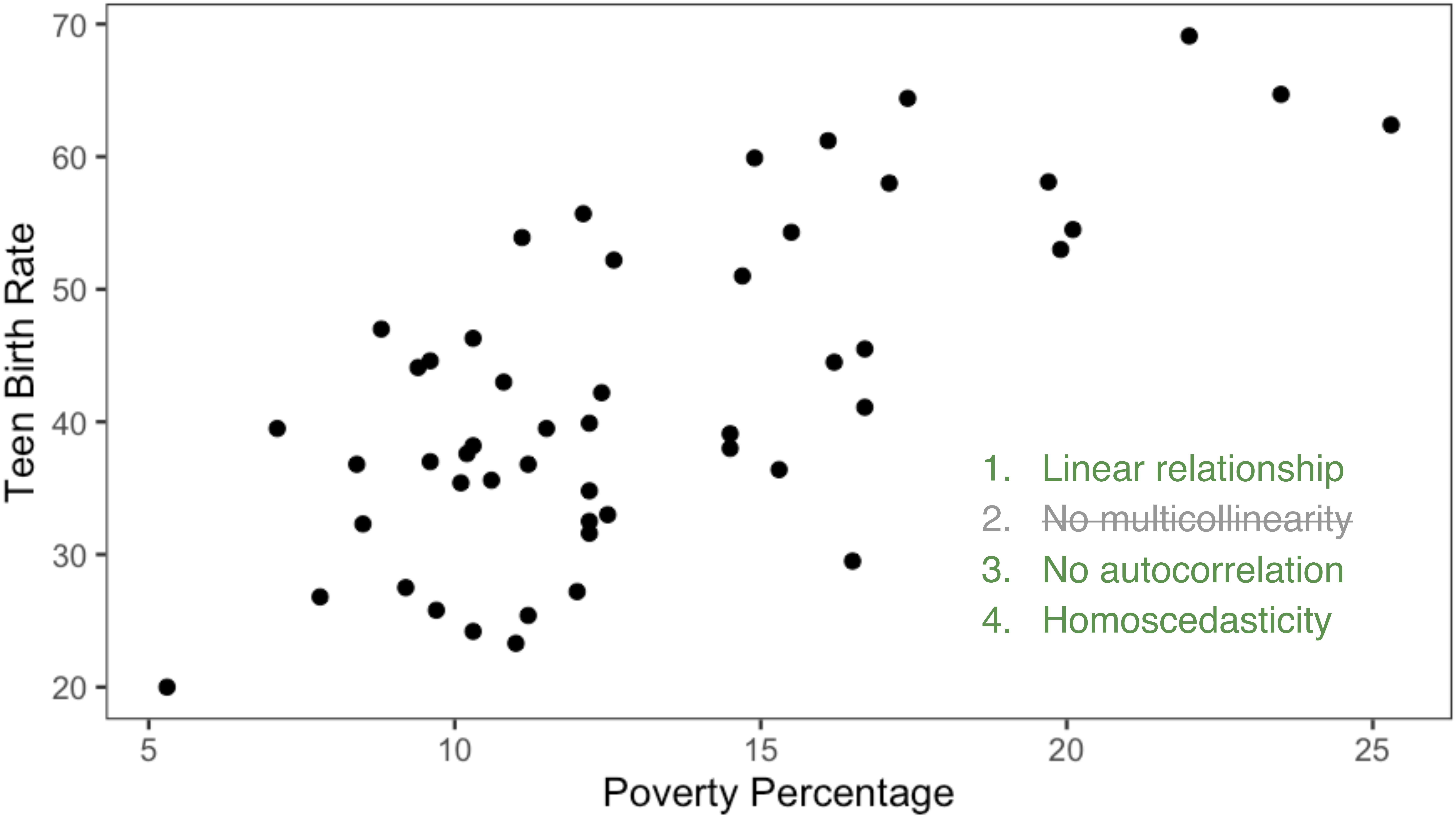
EDA: distributions



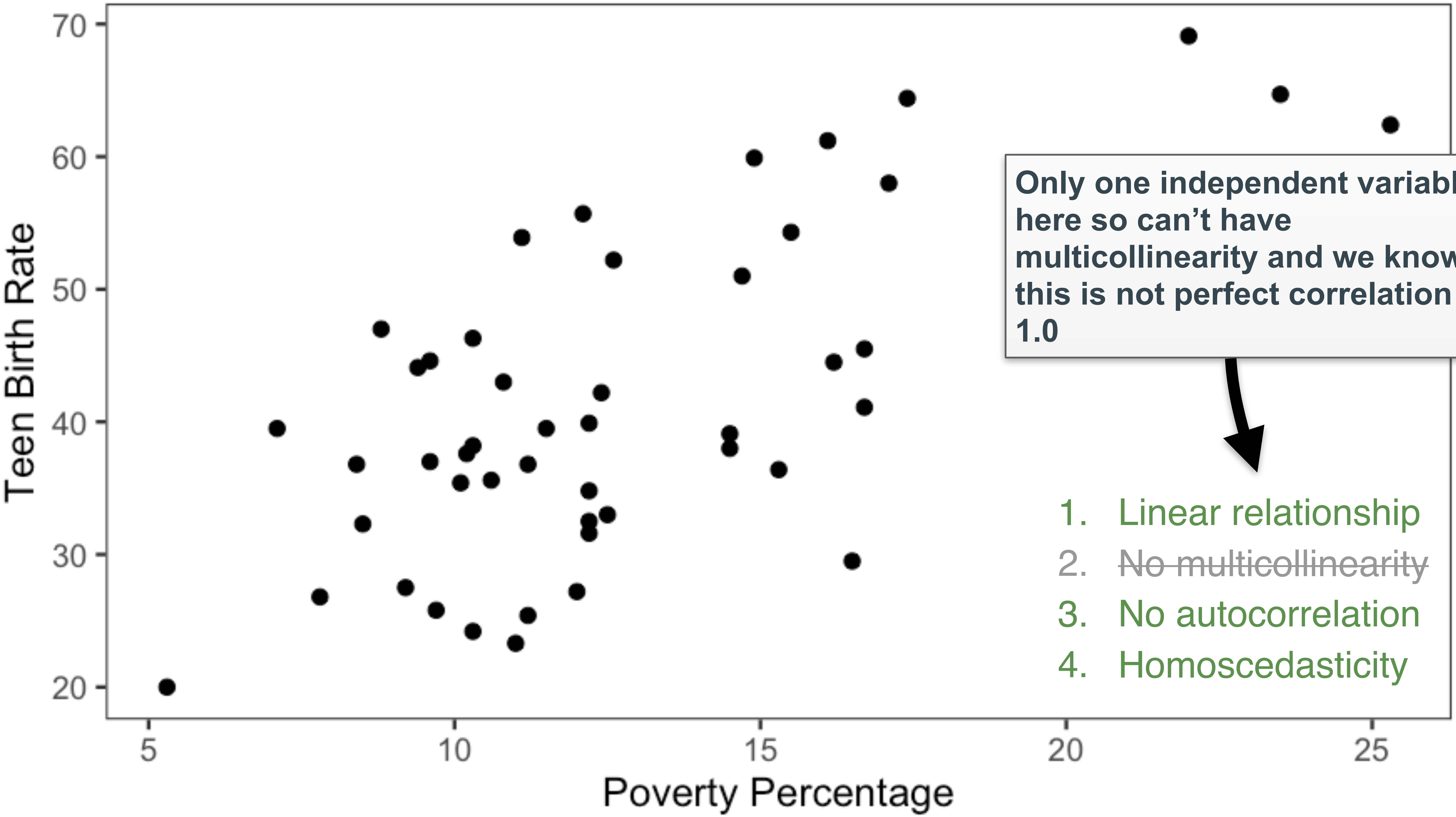




Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.

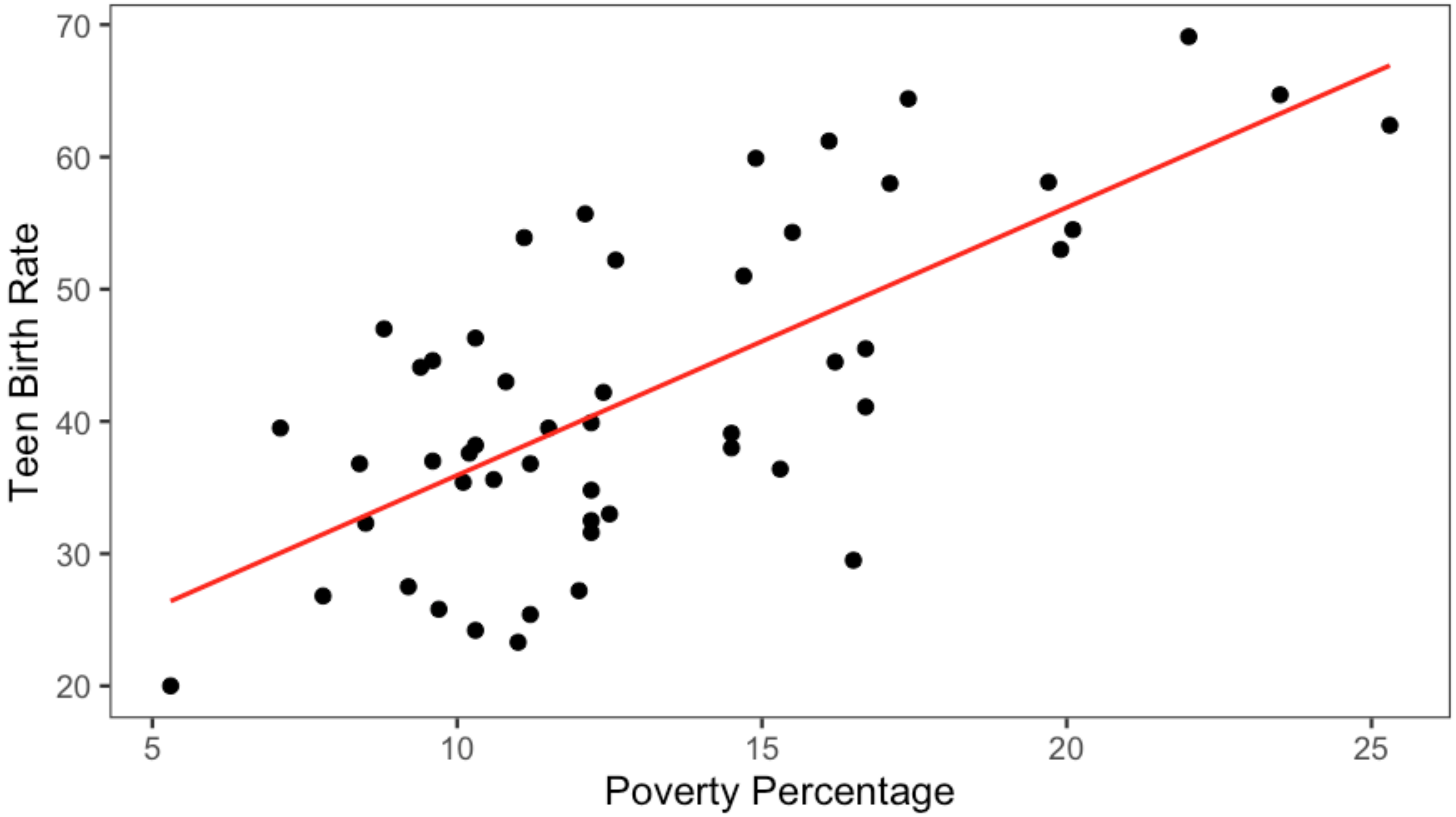


Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.

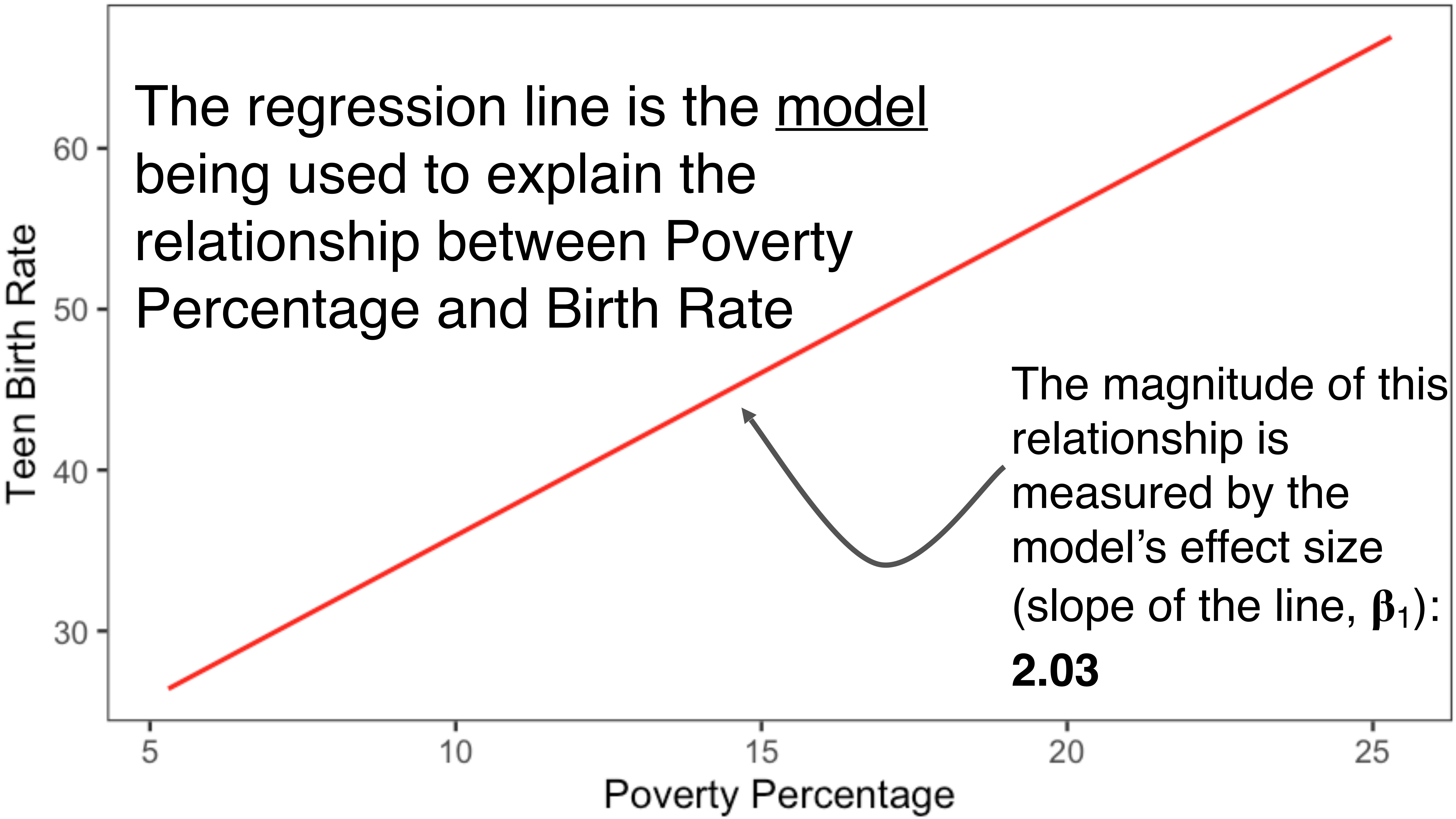


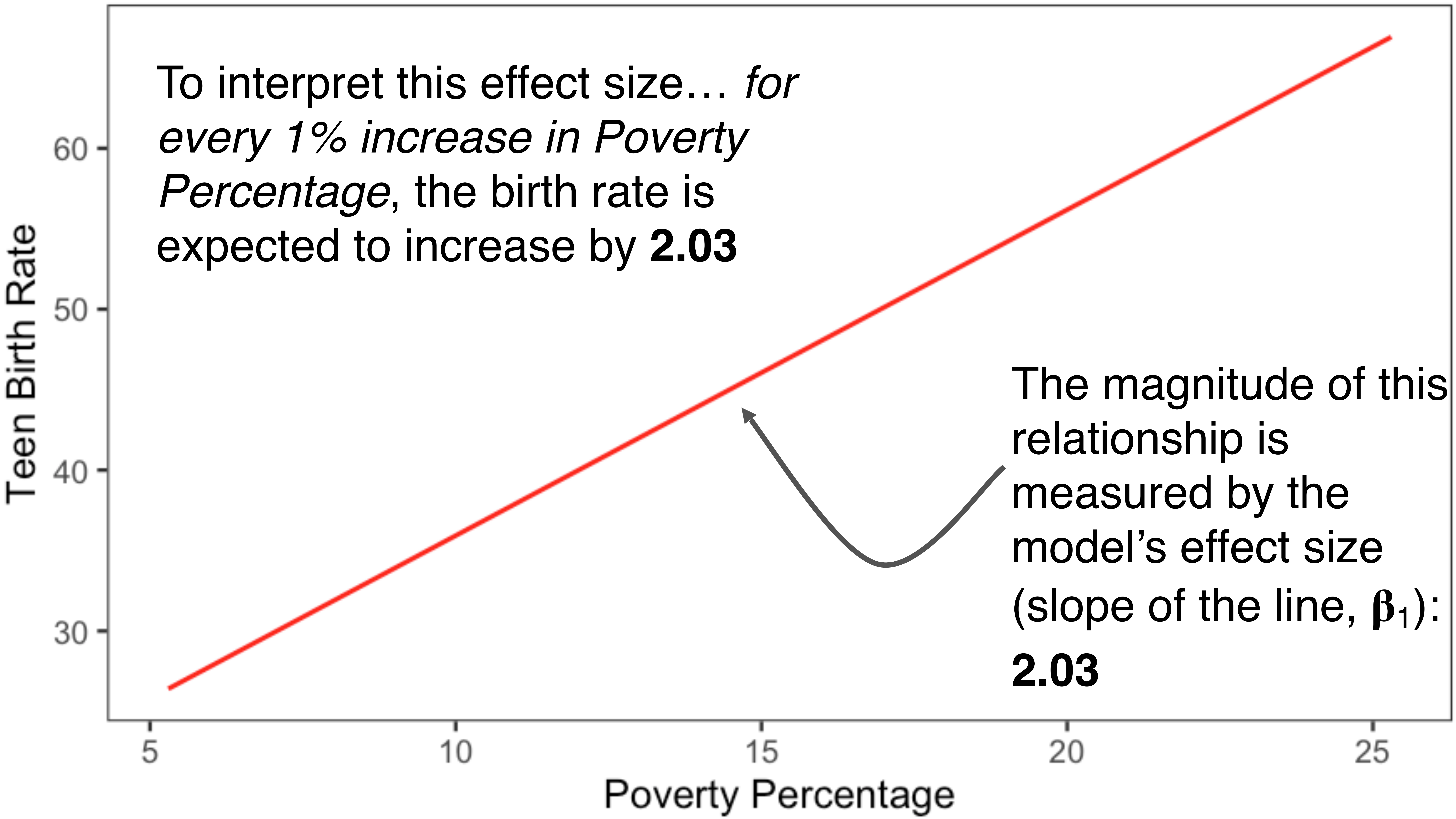
Only one independent variable here so can't have multicollinearity and we know this is not perfect correlation of 1.0

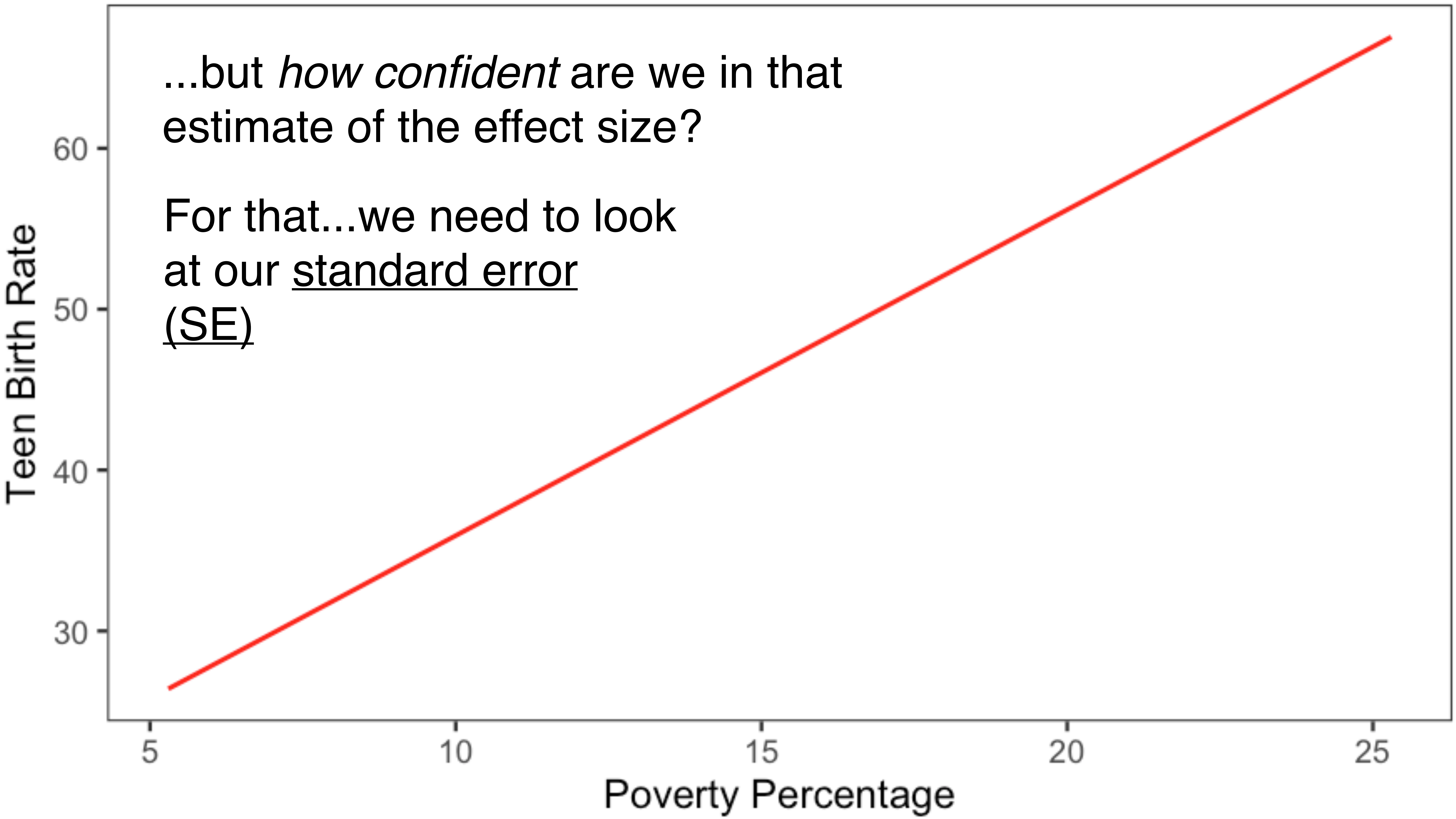
1. Linear relationship
2. ~~No multicollinearity~~
3. No autocorrelation
4. Homoscedasticity



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

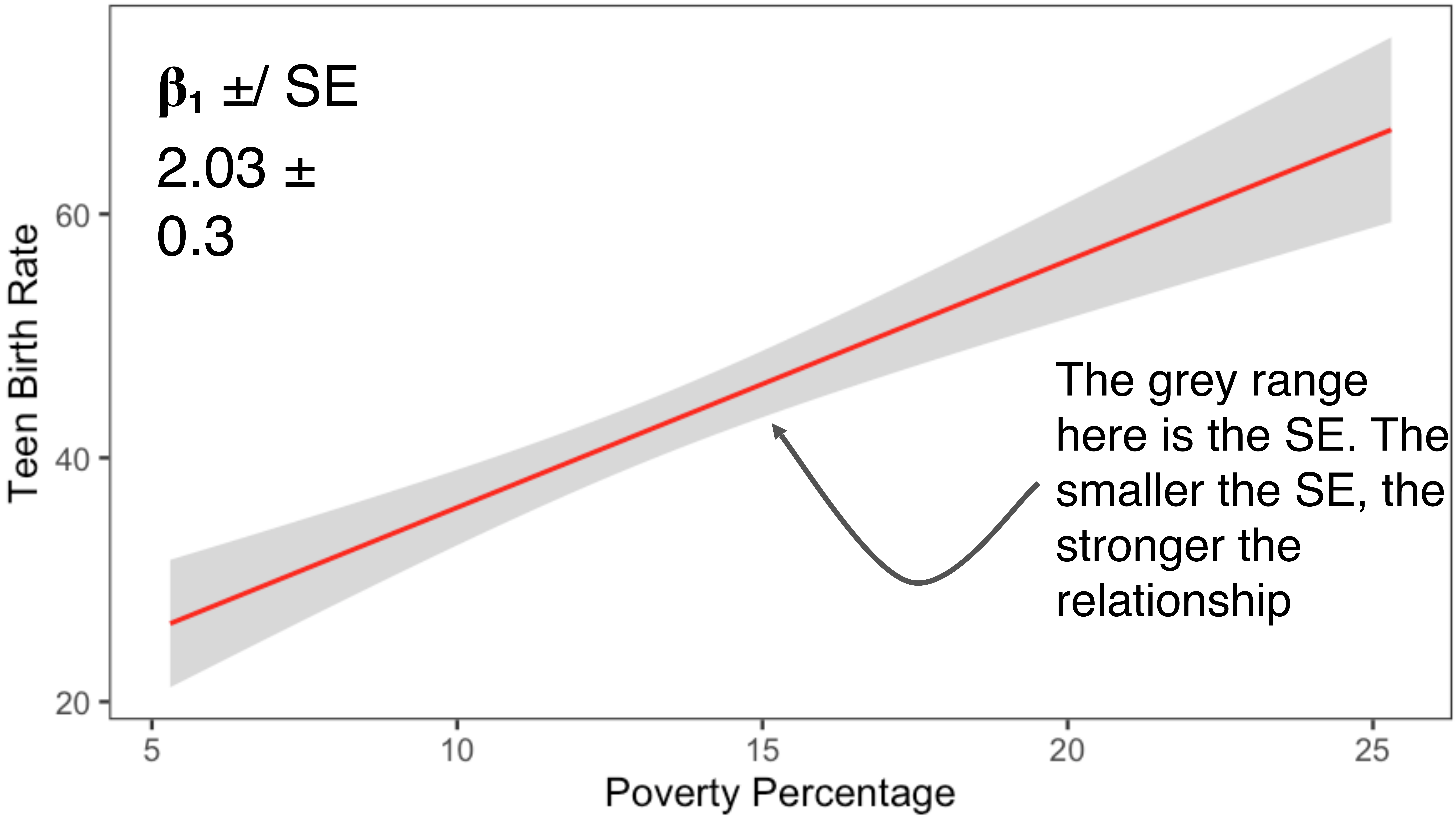

The regression line is the model being used to explain the relationship between Poverty Percentage and Birth Rate
The magnitude of this relationship is measured by the model's effect size (slope of the line, β_1): **2.03**
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage


To interpret this effect size… *for every 1% increase in Poverty Percentage*, the birth rate is expected to increase by **2.03**
The magnitude of this relationship is measured by the model's effect size (slope of the line, β_1): **2.03**
Teen Birth Rate
30
40
50
60
5
10
15
20
25
Poverty Percentage

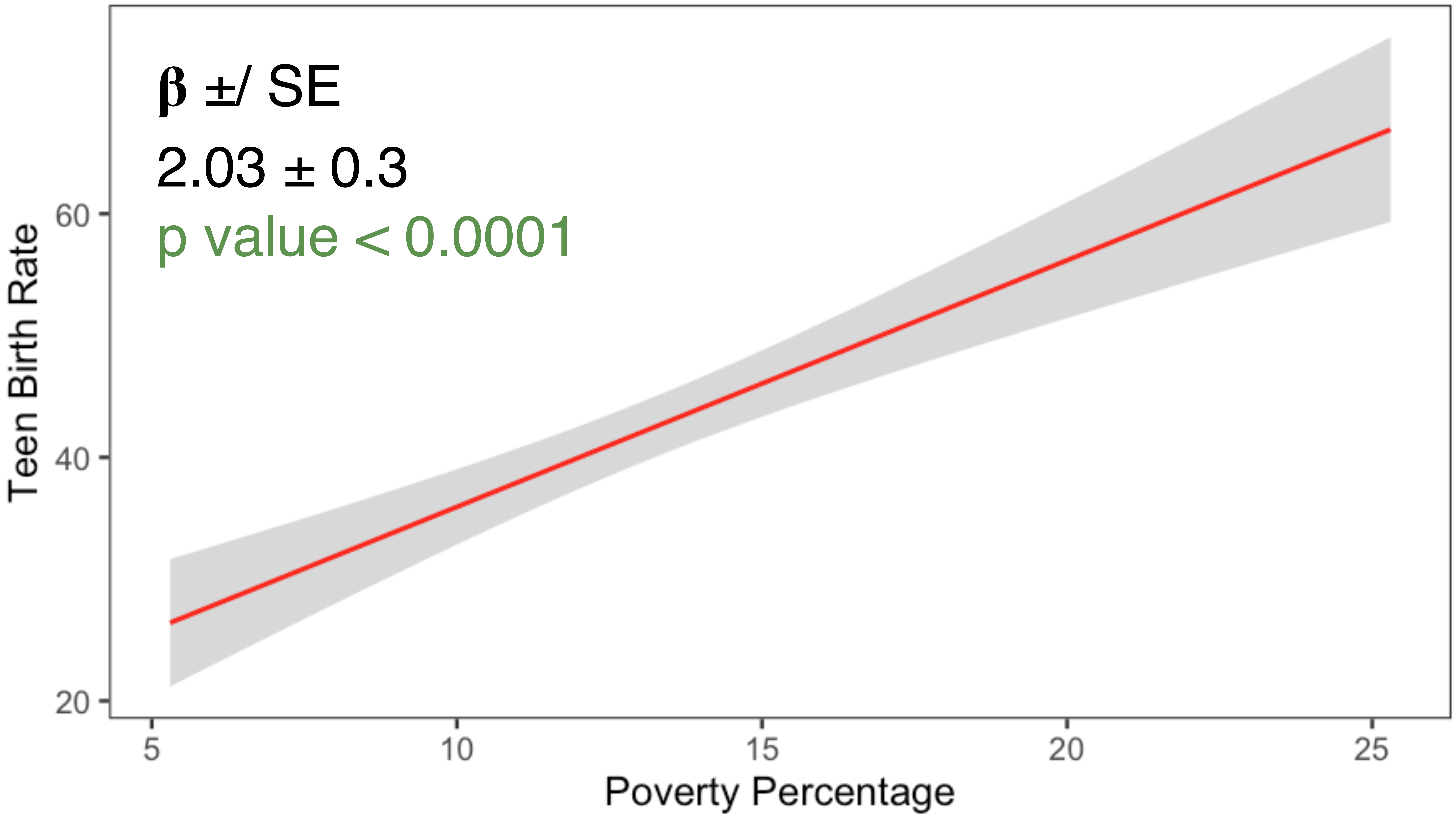


...but *how confident* are we in that estimate of the effect size?

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

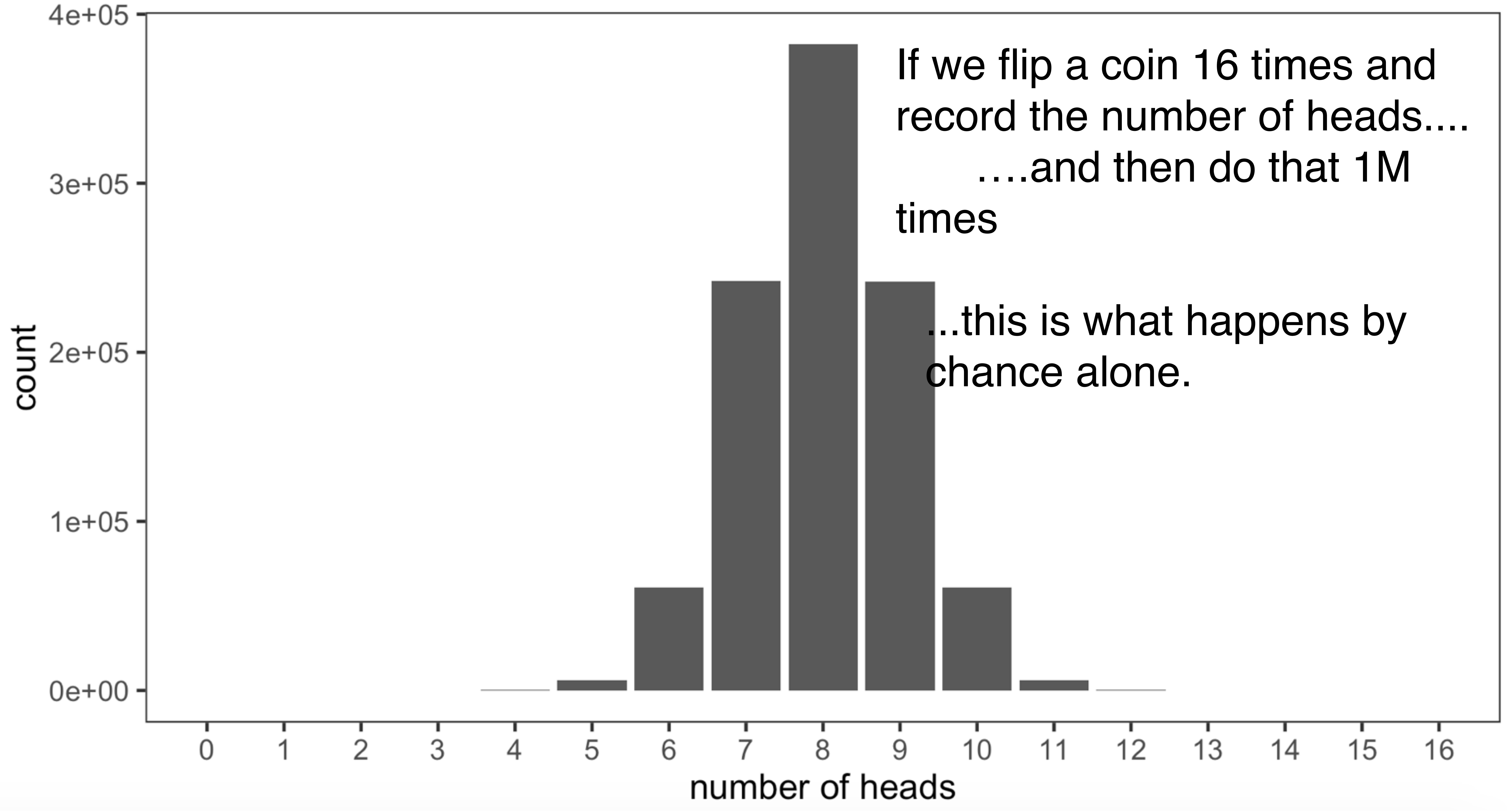


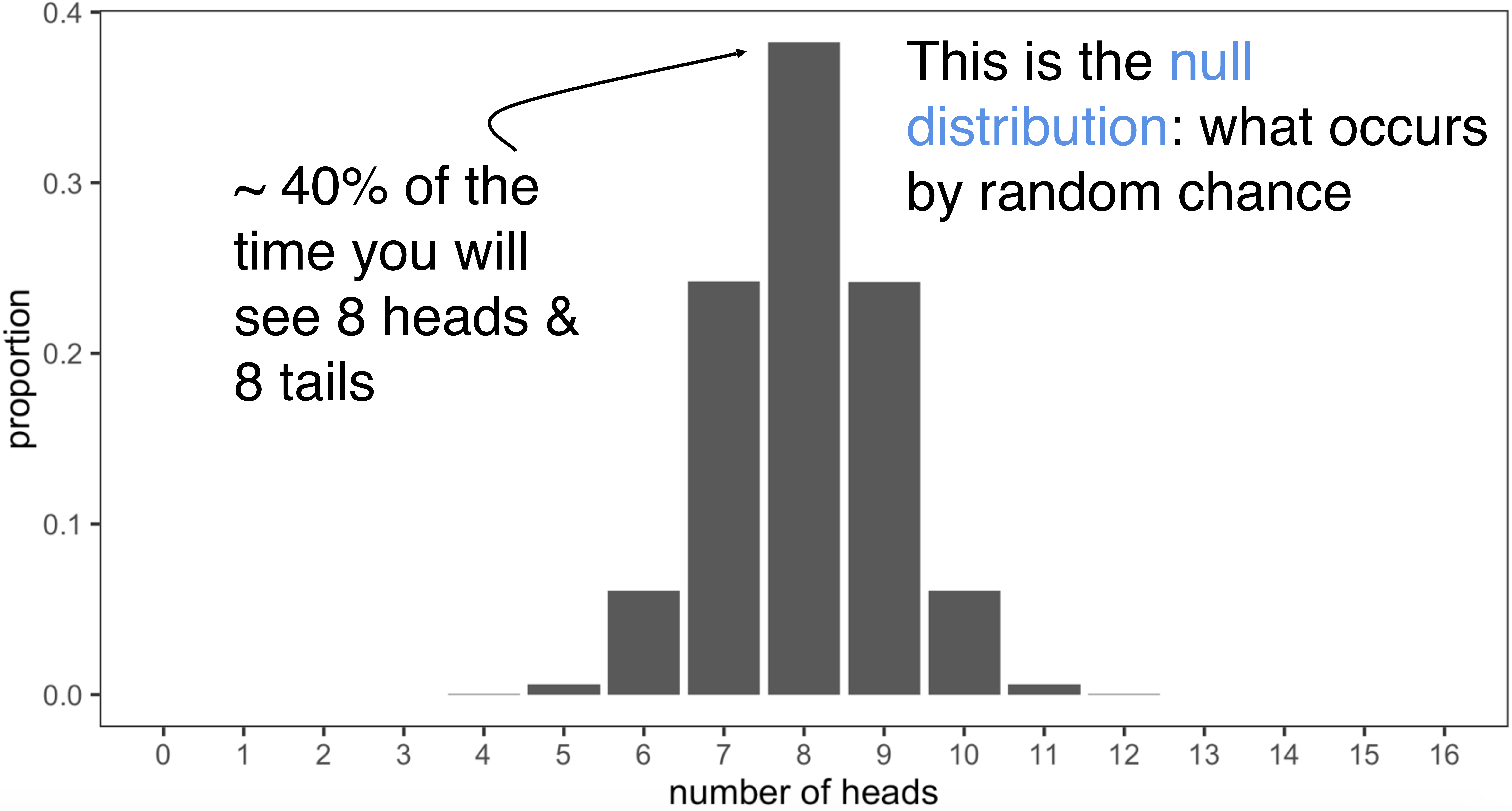
β1 ±/ SE
2.03 ±
0.3
The grey range here is the SE. The smaller the SE, the stronger the relationship
60
40
20
Teen Birth Rate
5
10
15
20
25
Poverty Percentage



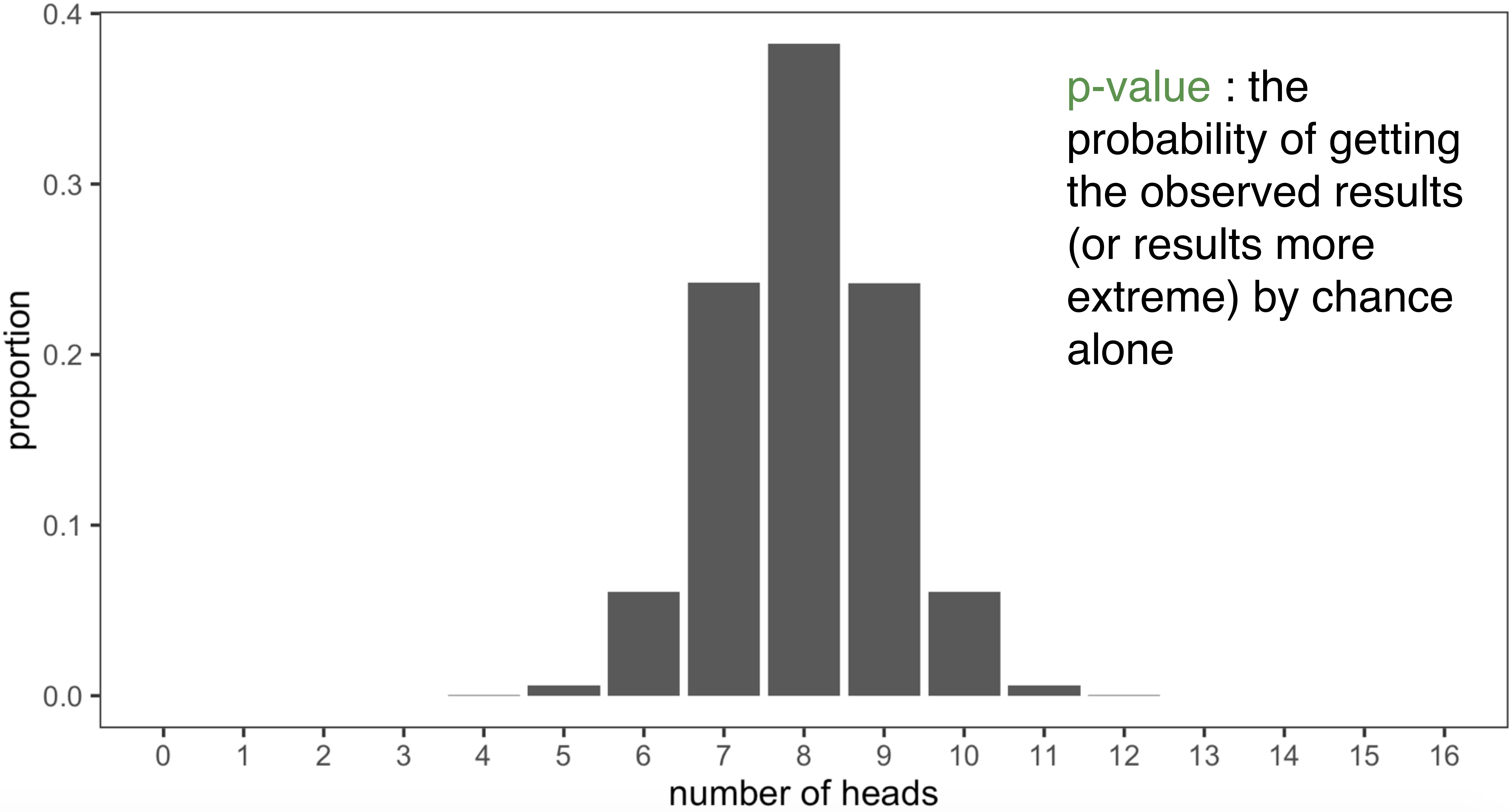
β ±/ SE
2.03 ± 0.3
p value < 0.0001
Teen Birth Rate
20
40
60
5
10
15
20
25
Poverty Percentage

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

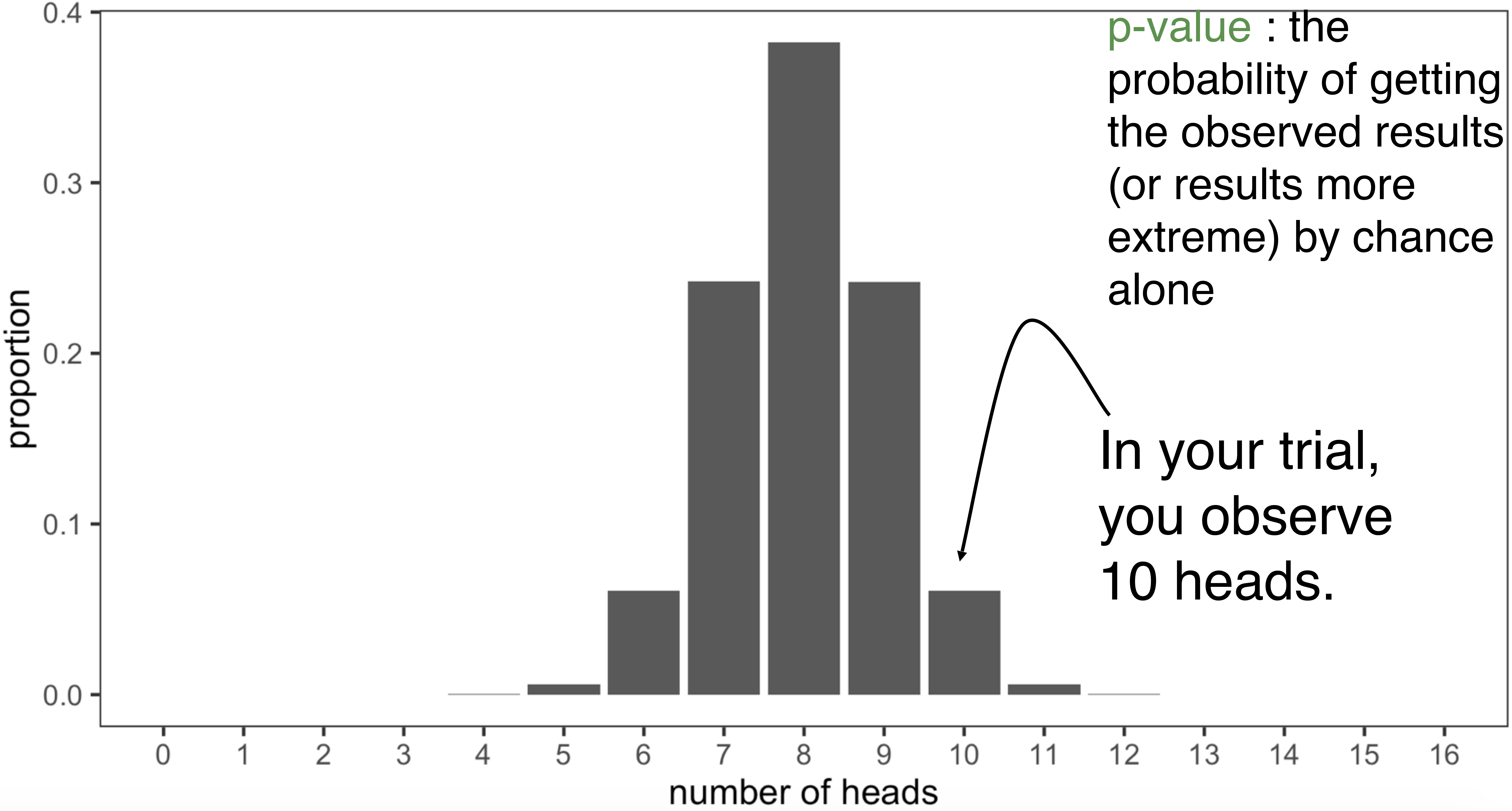

If we flip a coin 16 times and record the number of heads....
….and then do that 1M times
...this is what happens by chance alone.
count
4e+05
3e+05
2e+05
1e+05
0e+00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads



This is the null distribution: what occurs by random chance
~ 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

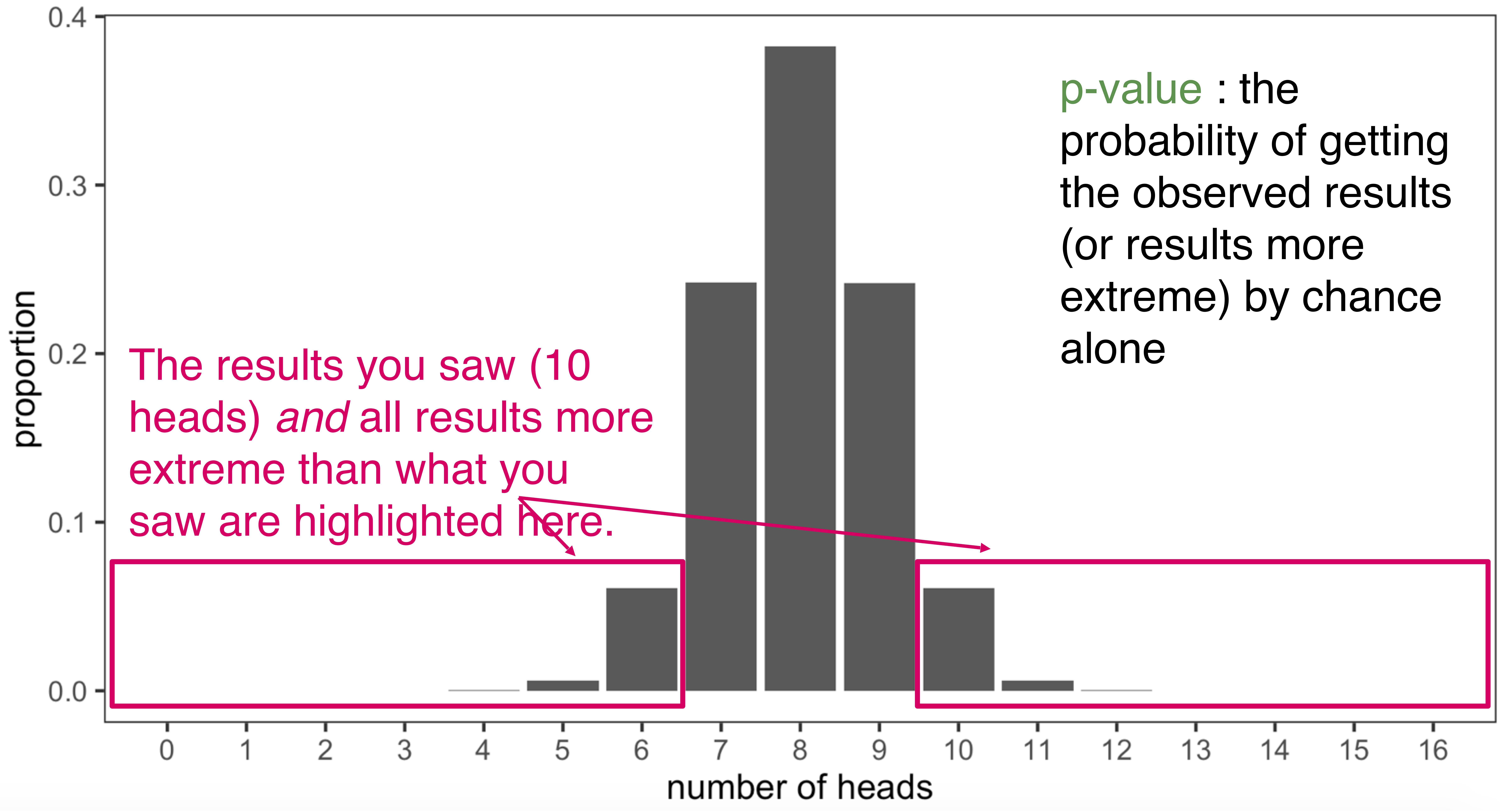


p-value : the probability of getting the observed results (or results more extreme) by chance alone

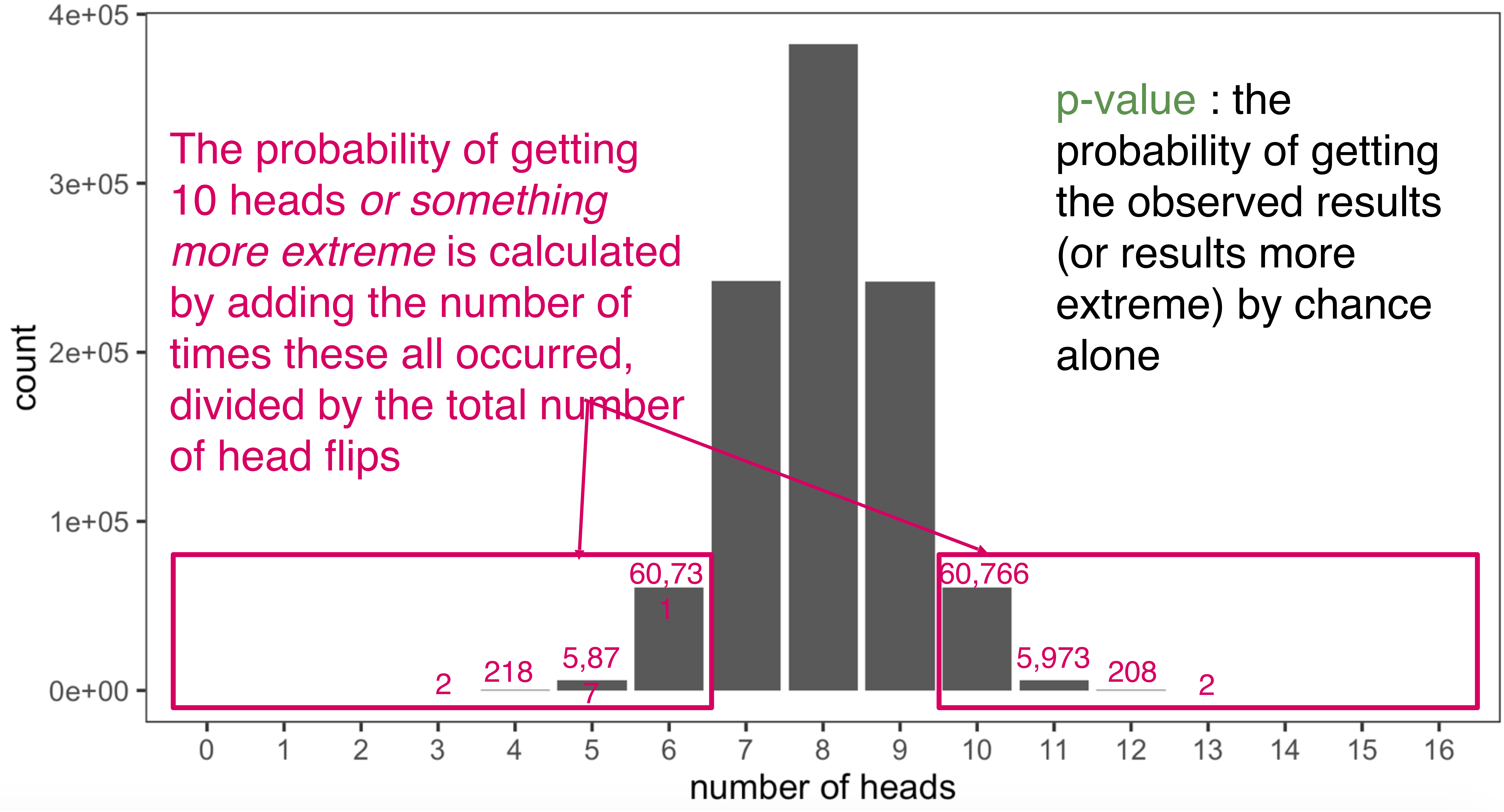


p-value : the probability of getting the observed results (or results more extreme) by chance alone

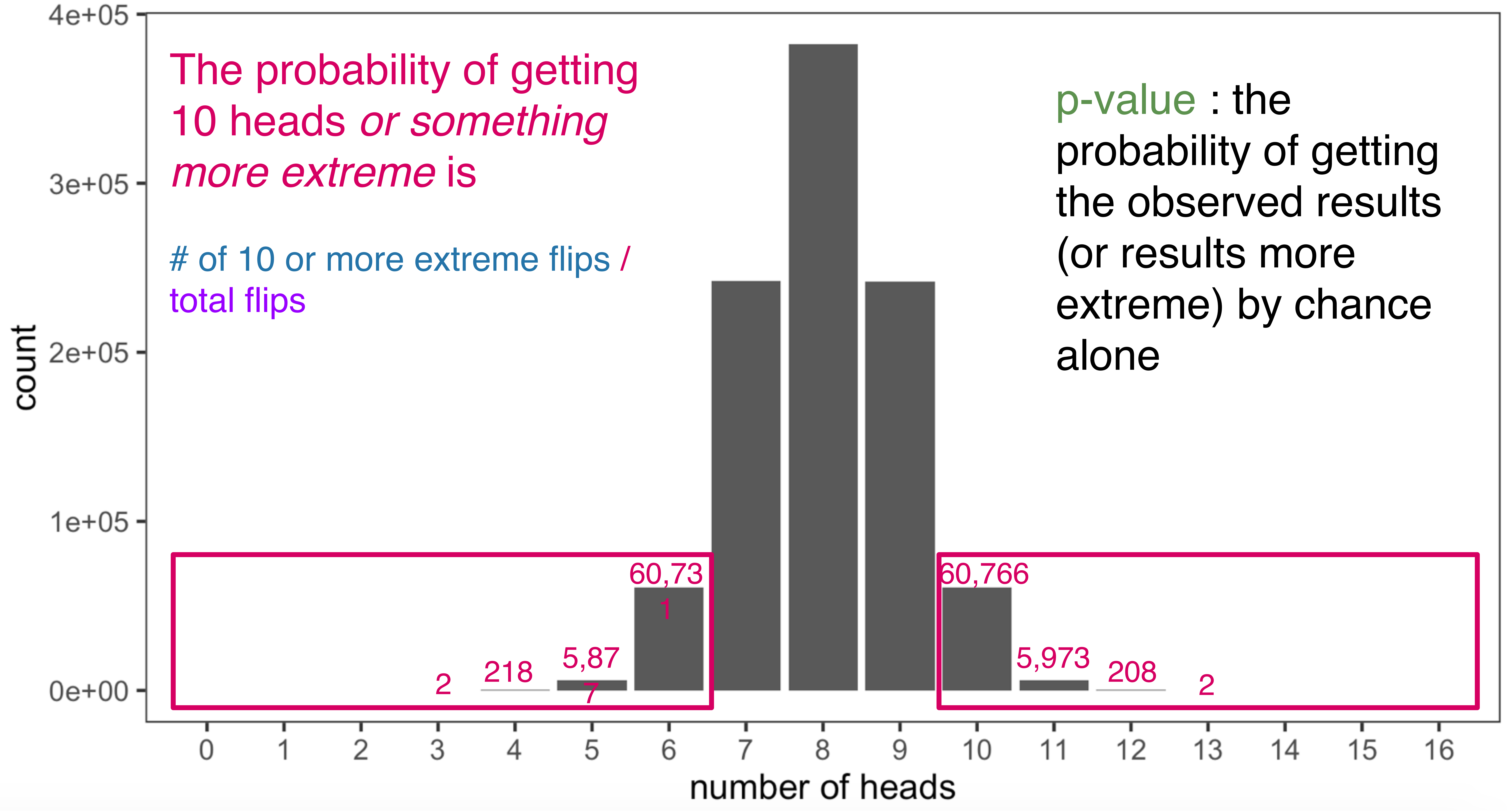
In your trial, you observe 10 heads.

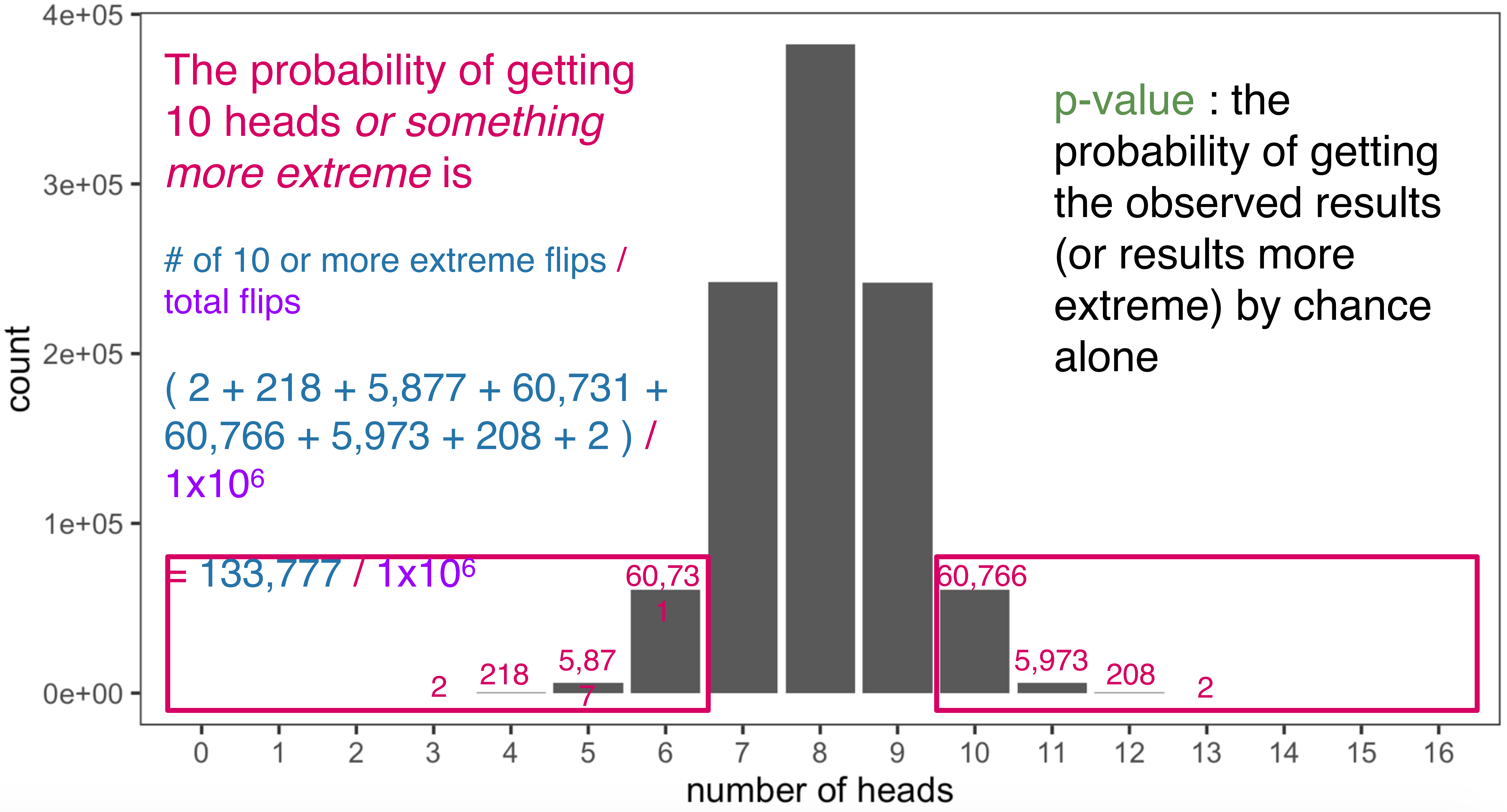


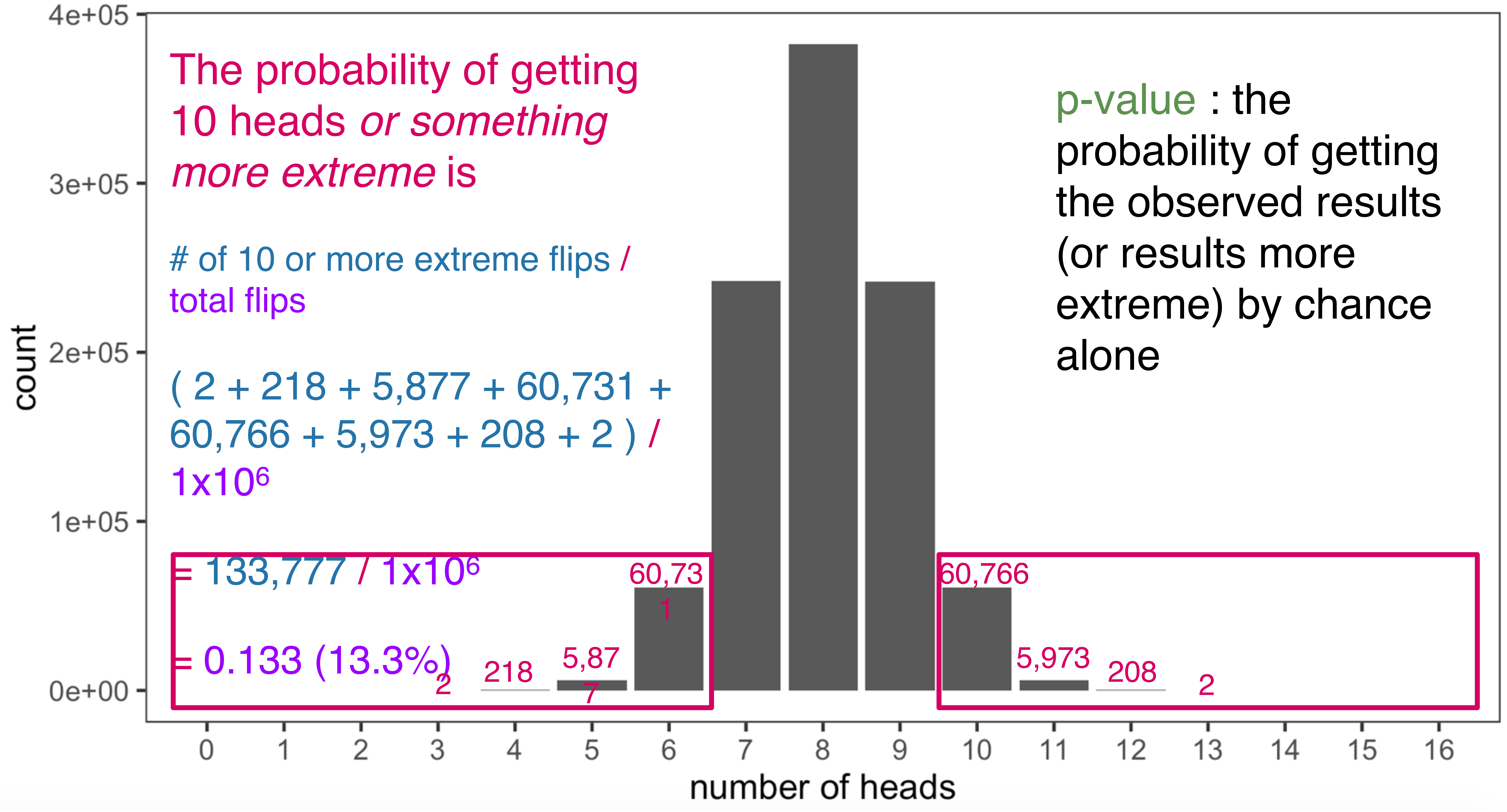
p-value : the probability of getting the observed results (or results more extreme) by chance alone
The results you saw (10 heads) *and* all results more extreme than what you saw are highlighted here.
proportion
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads



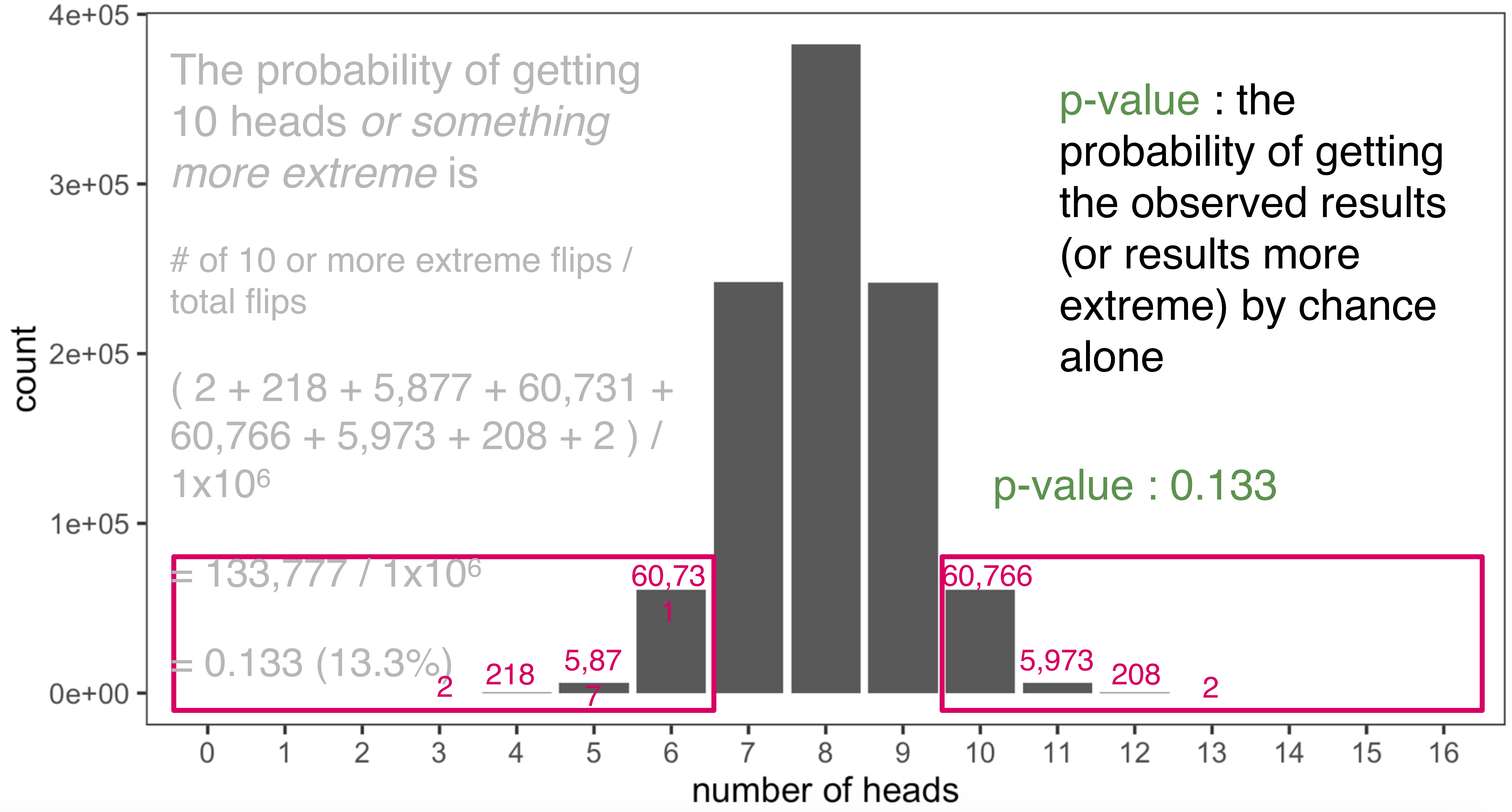
The probability of getting 10 heads *or something more extreme* is calculated by adding the number of times these all occurred, divided by the total number of head flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
4e+05
3e+05
2e+05
1e+05
0e+00
count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

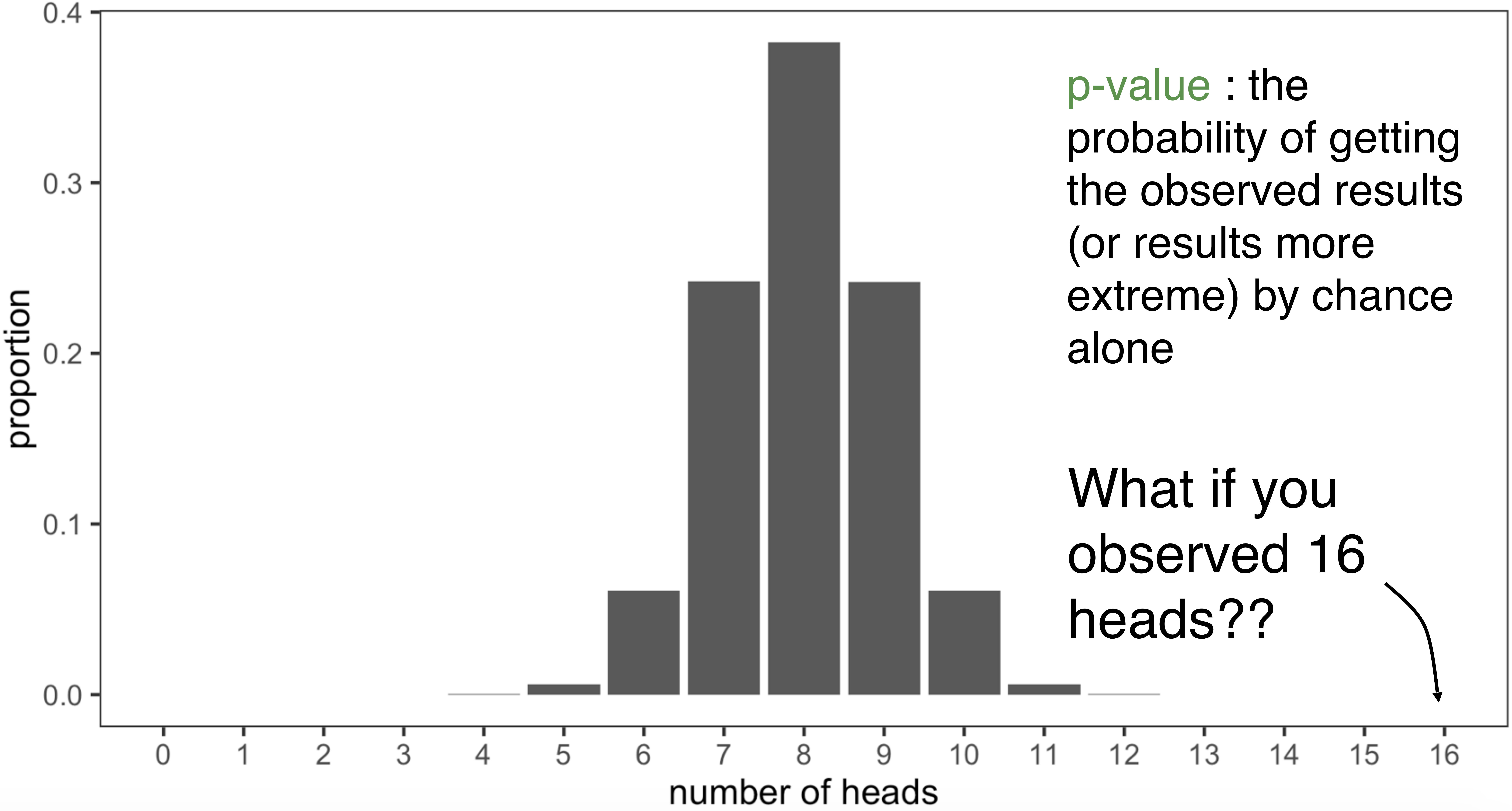

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
60,731
60,766
5,877
5,973
218
208
2
2
count
4e+05
3e+05
2e+05
1e+05
0e+00
0 1 2 3 4 5 6 7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



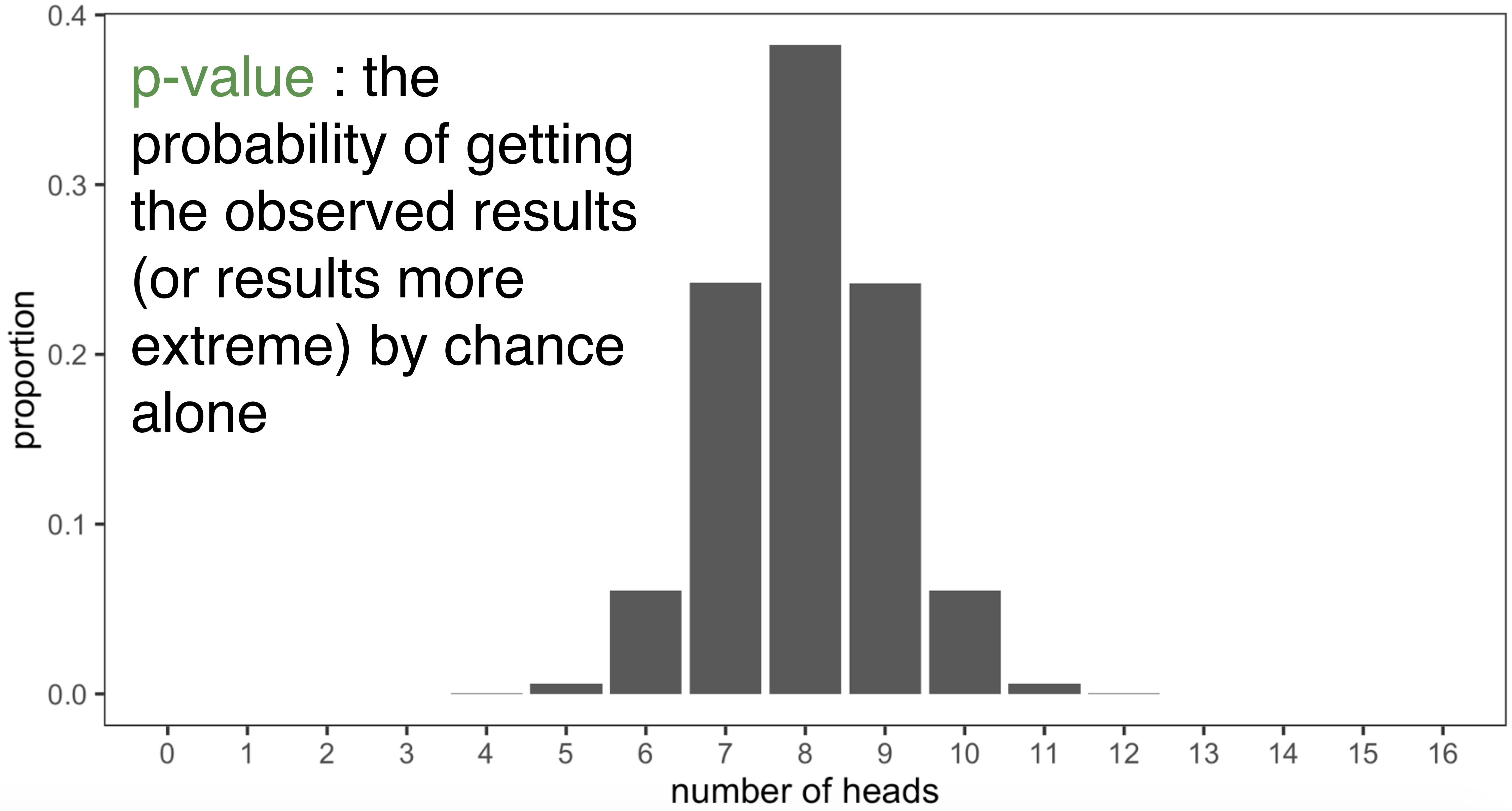
The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / $1x10^6$
= 133,777 / $1x10^6$
= 0.133 (13.3%)
p-value : the probability of getting the observed results (or results more extreme) by chance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


4e+05
3e+05
2e+05
1e+05
0e+00
count
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%)
2
218
5,877
60,731
60,766
5,973
208
2
p-value : the probability of getting the observed results (or results more extreme) by chance alone
p-value : 0.133
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

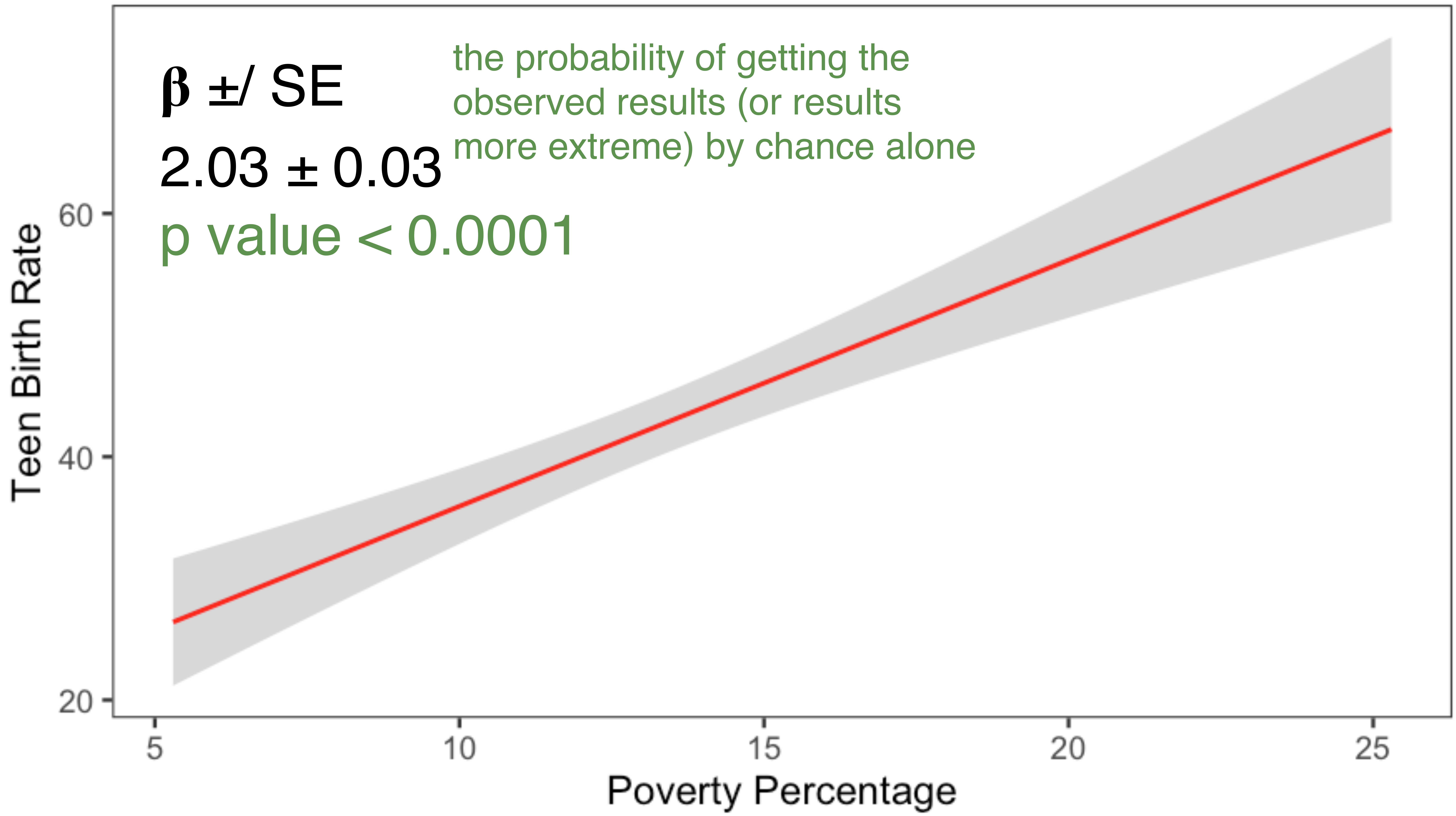
What if you observed 16 heads??



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

A
< 0.13

B
> 0.13


β ±/ SE
2.03 ± 0.03
the probability of getting the observed results (or results more extreme) by chance alone
p value < 0.0001
60
40
20
Teen Birth Rate
5
10
15
20
25
Poverty Percentage

Takes into account
the effect size (β_1)
and the SE

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

Confounding

Confounding

popsicles → crime rate

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

A
popsicle preference

B
new gun laws

C
temperature

D
changes in popsicle prices

E
new law enforcement officers

We'll discuss additional approaches of how to account for confounding in your analysis in another lecture.

Ignoring confounders will lead you to draw incorrect conclusions from your analyses

Spine Surgery Results

Sample: 400 patients with index vertebral fractures

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

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

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 quickly

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