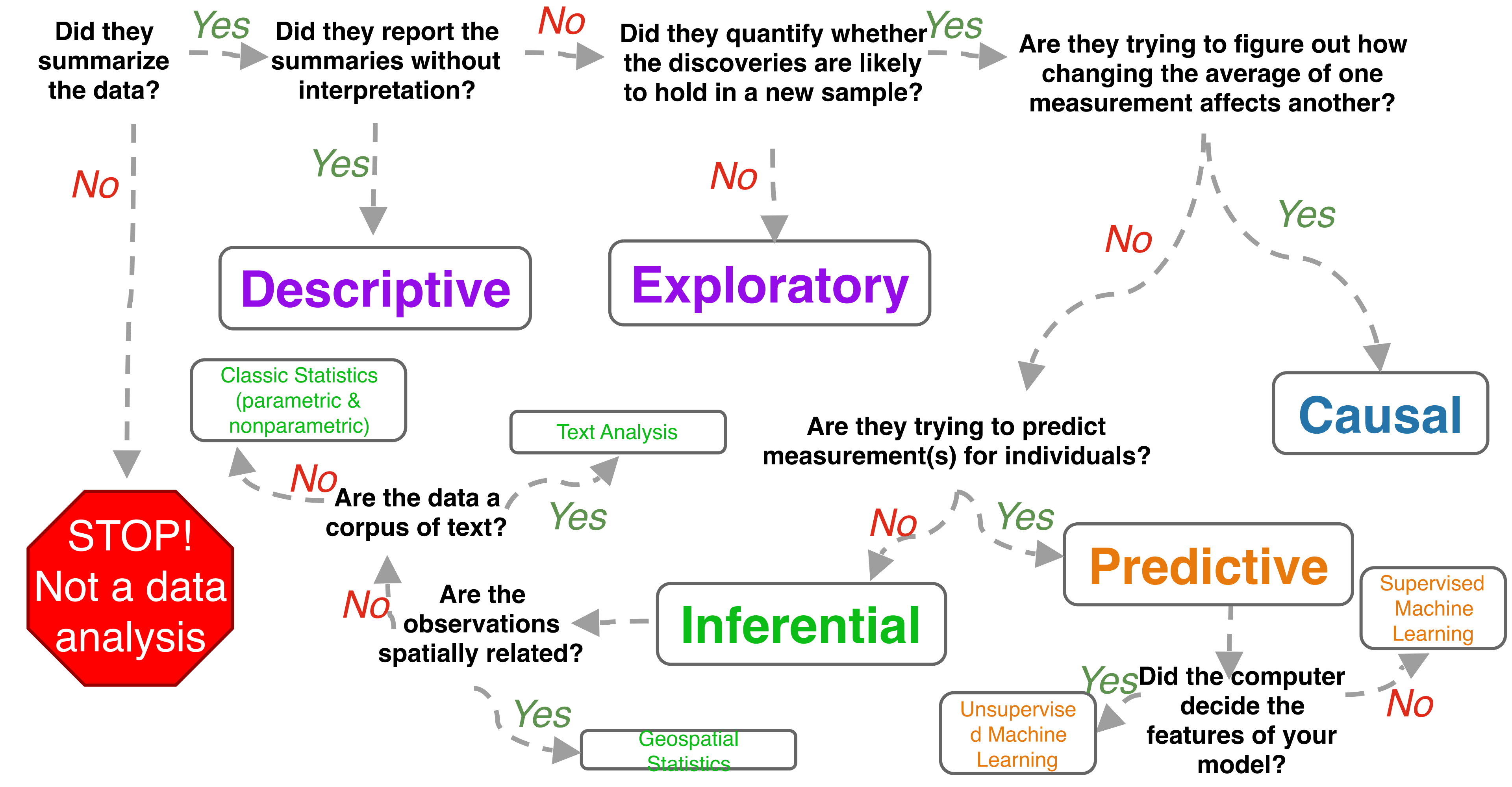
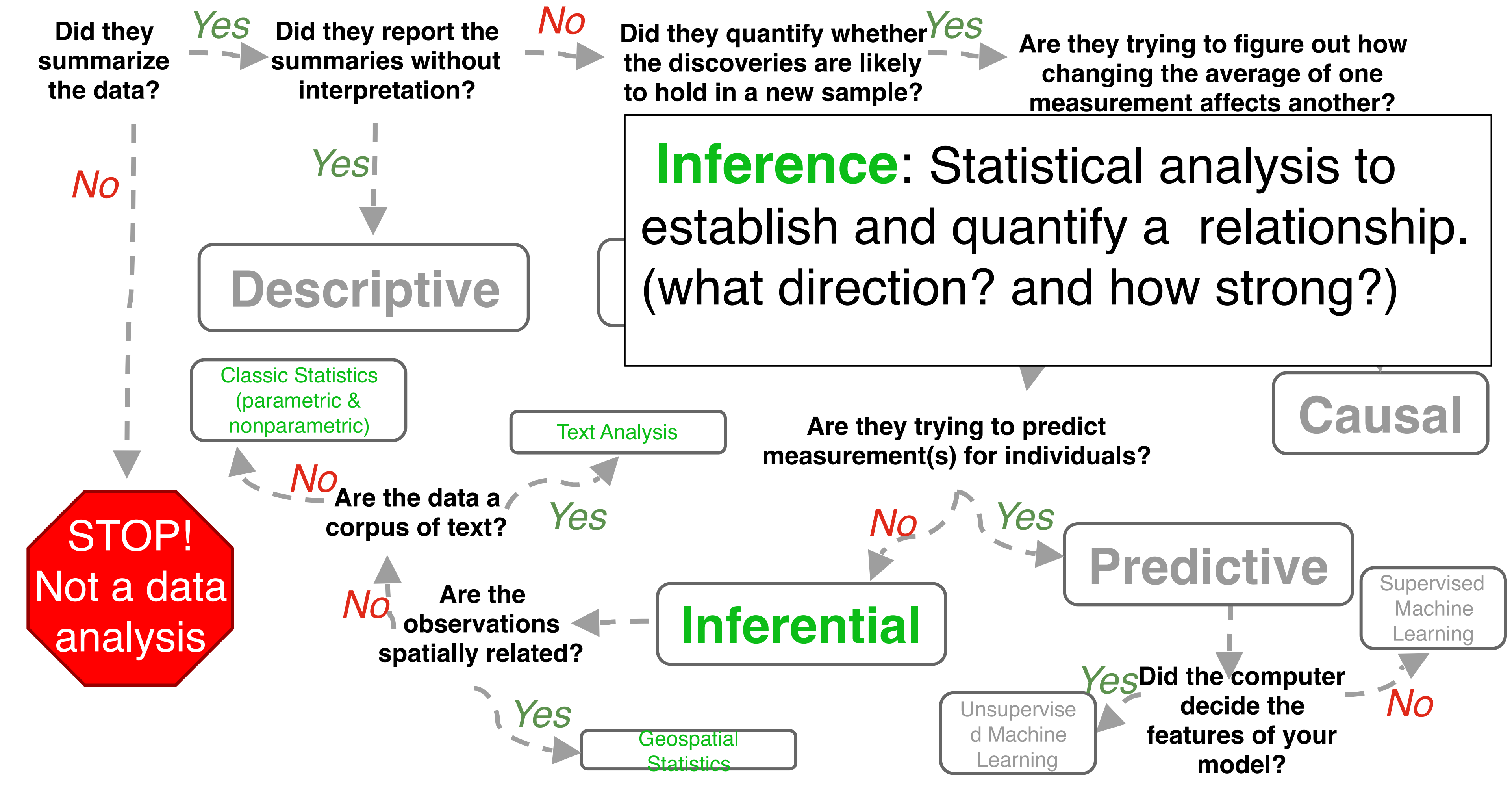
Inferential Analysis

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

Department of Cognitive Science
rdprobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures : https://github.com/COGS108/Lectures-Wi23


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


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?
Inference: Statistical analysis to establish and quantify a relationship. (what direction? and how strong?)
No
Yes
Descriptive
Classic Statistics (parametric & nonparametric)
Text Analysis
Are they trying to predict measurement(s) for individuals?
Causal
No
Are the data a corpus of text?
Yes
No
Yes
Predictive
STOP! Not a data analysis
No
Are the observations spatially related?
Inferential
Supervised Machine Learning
Yes
Did the computer decide the features of your model?
No
Unsupervised Machine Learning
Yes
Geospatial Statistics

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

Sesame Street viewership ?? Test scores

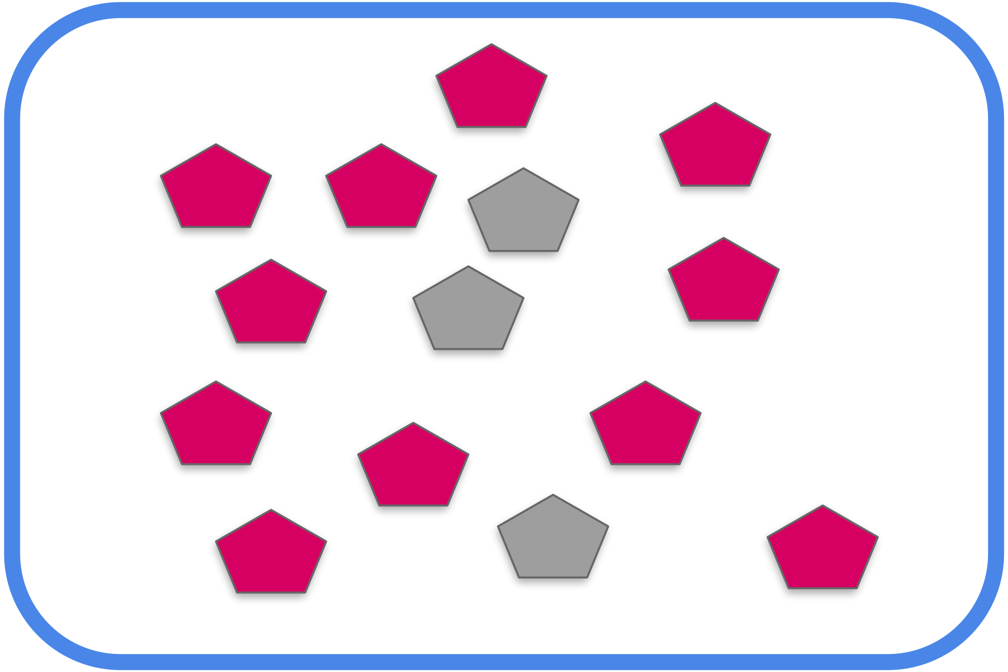
Establishing & Stating Your Null and Alternative Hypotheses Helps Guide Your Analysis

Null Hypothesis:

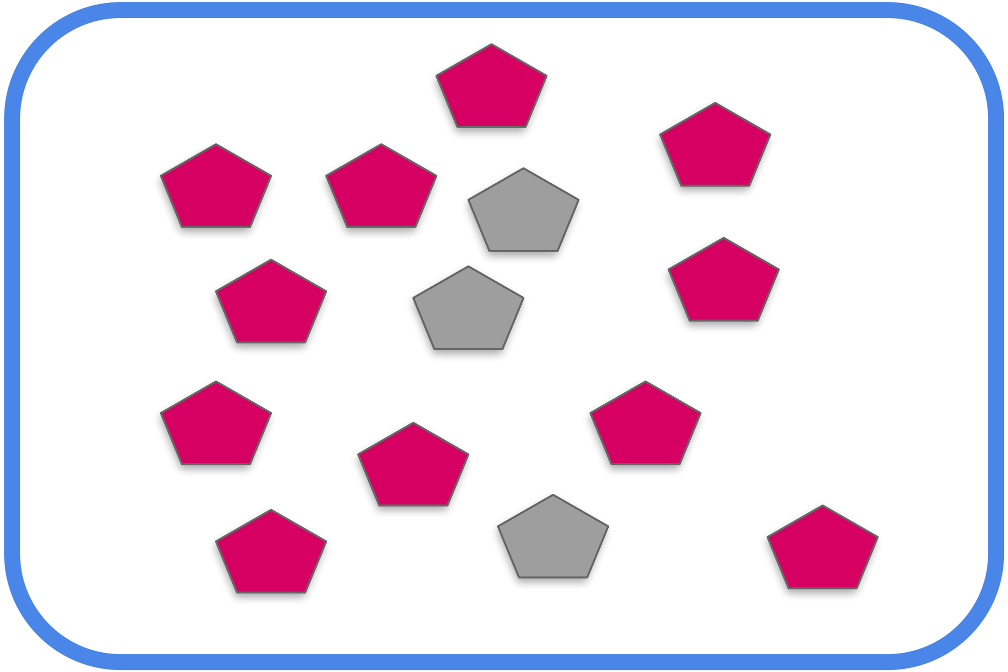
H_0: Sesame Street has *no effect* on kids brain development

Alternative Hypothesis:

H_a: Watching Sesame Street *has an effect* on kids' brain development

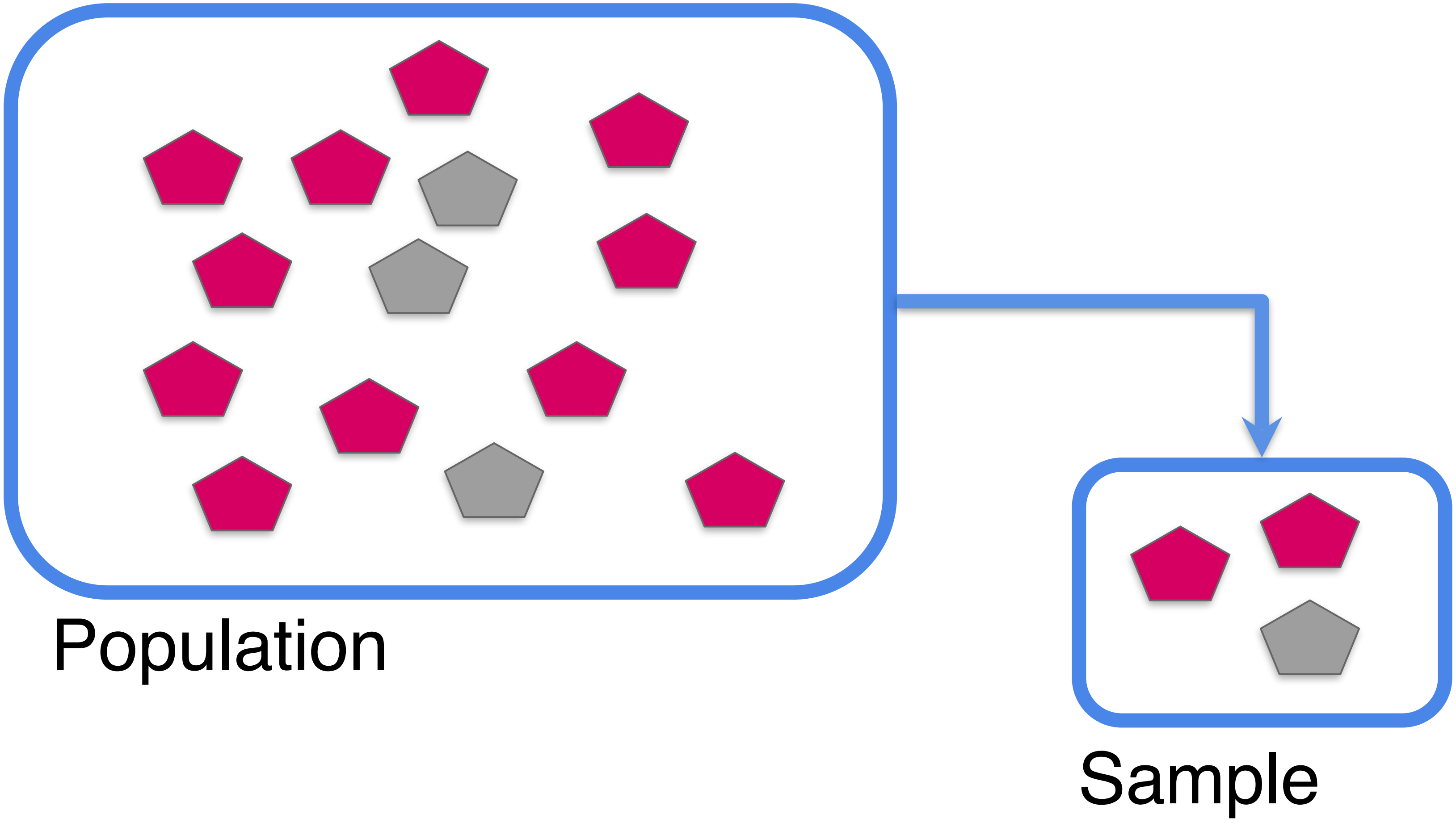


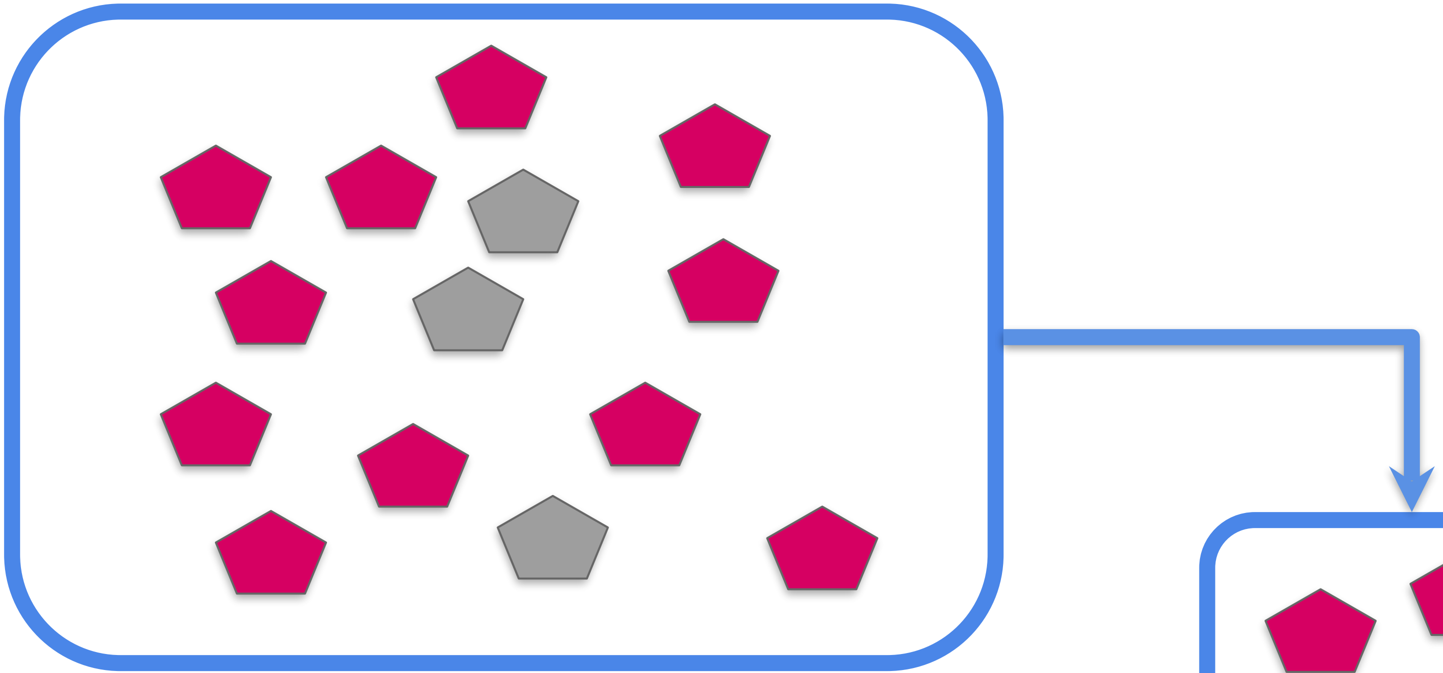
Population



Population

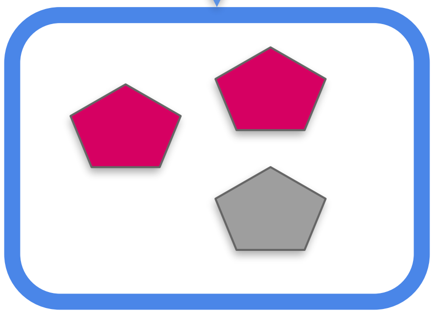
In our Sesame street example, the population would be all children


Population
Sample

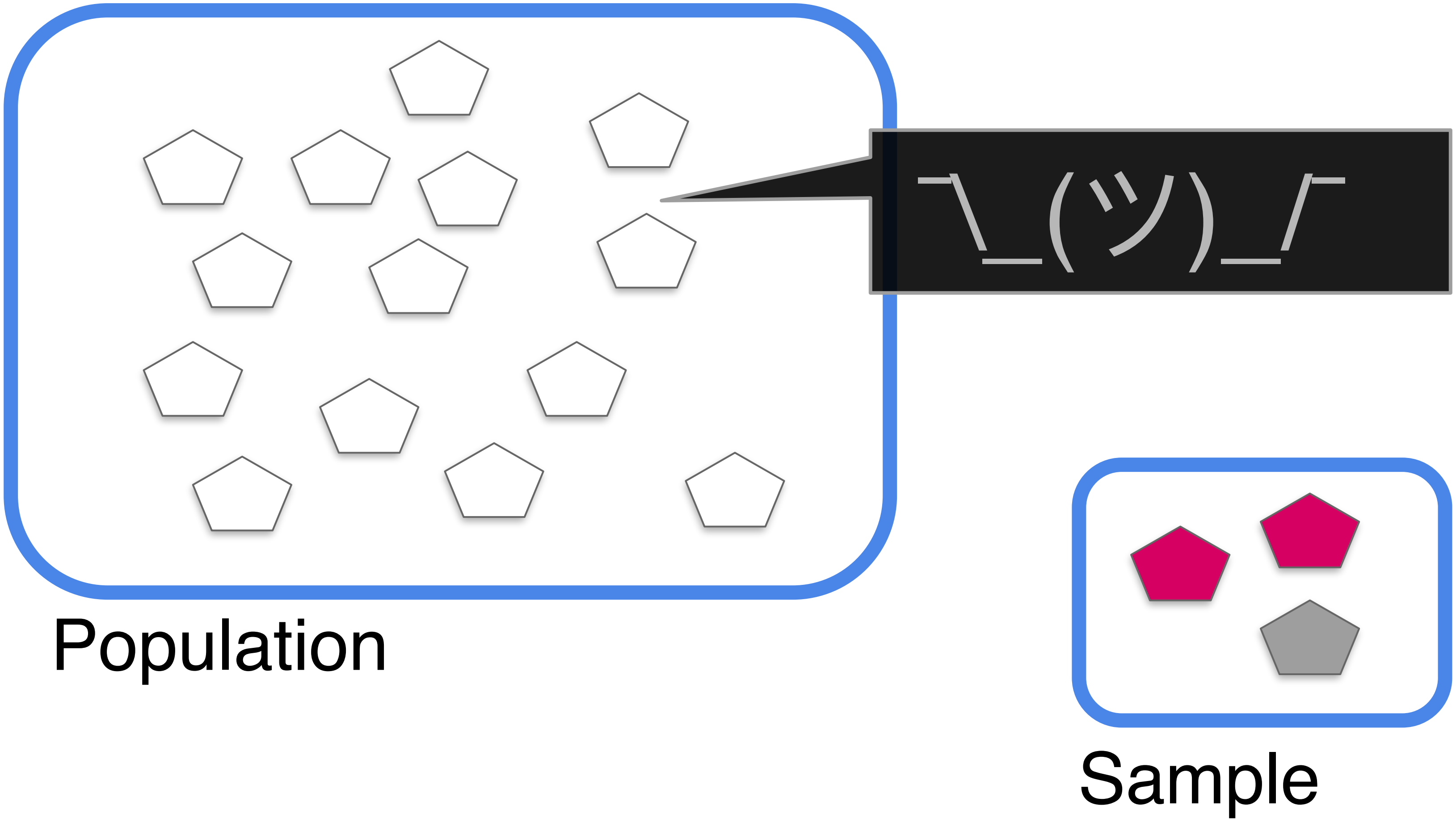


Population

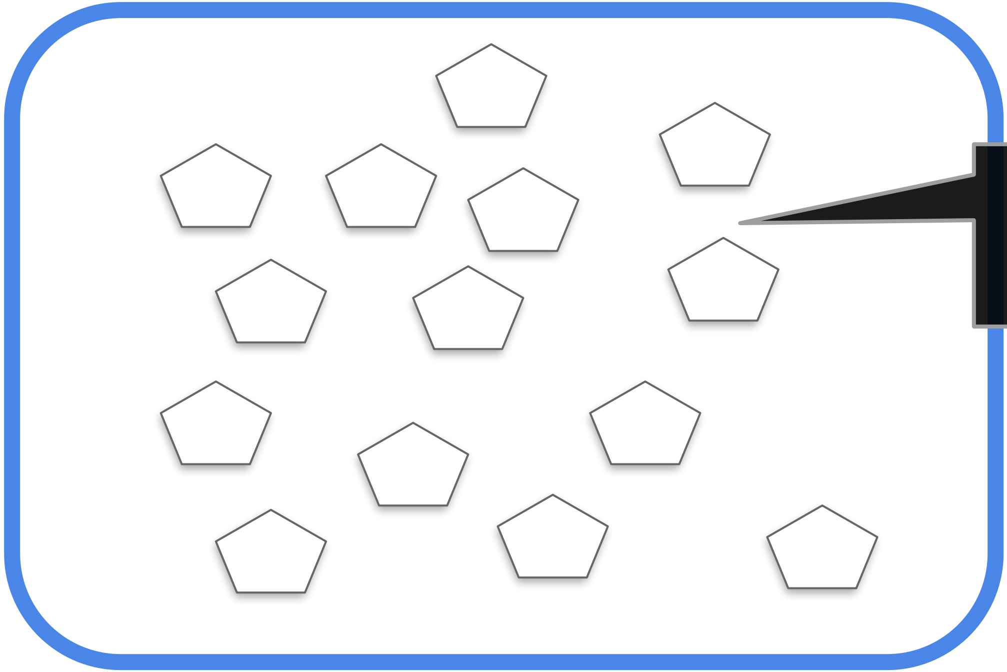
In our Sesame street example, the <u>sample</u> would be the children included in the study



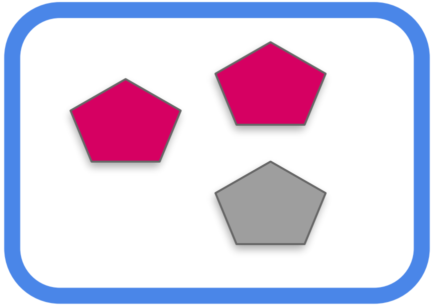
Sample


¯_(ツ)_/¯
Population
Sample

We don’t know how much Sesame street was watched by or the tests scores of <u>all kids</u>

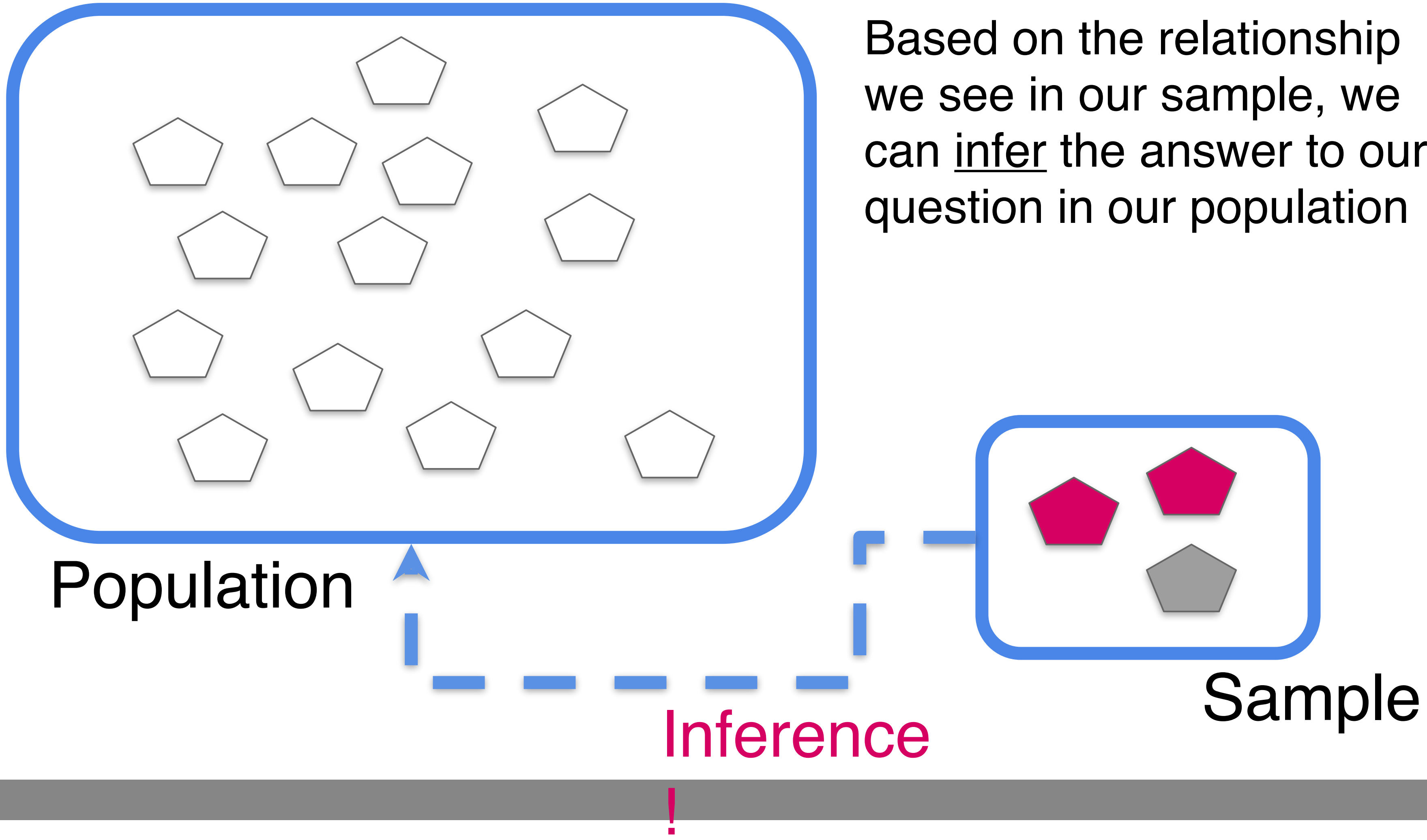


Population

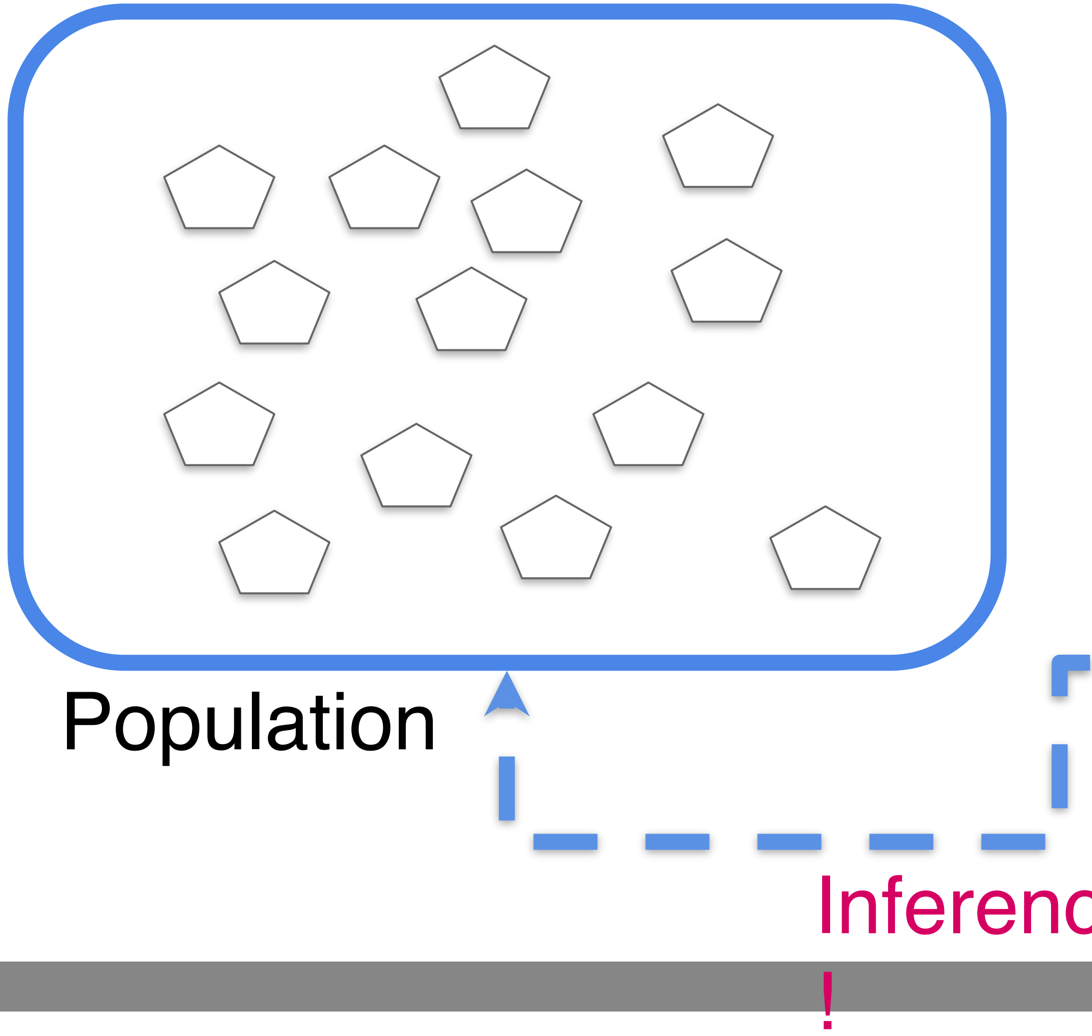


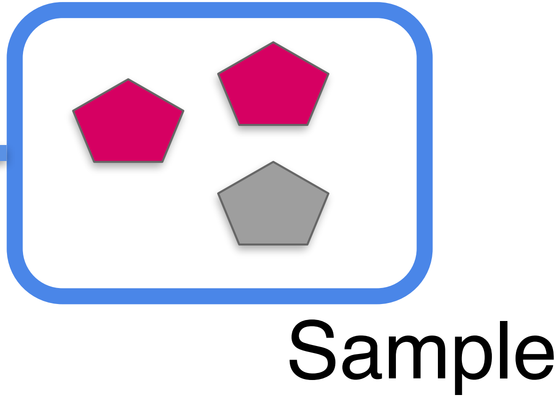
Sample

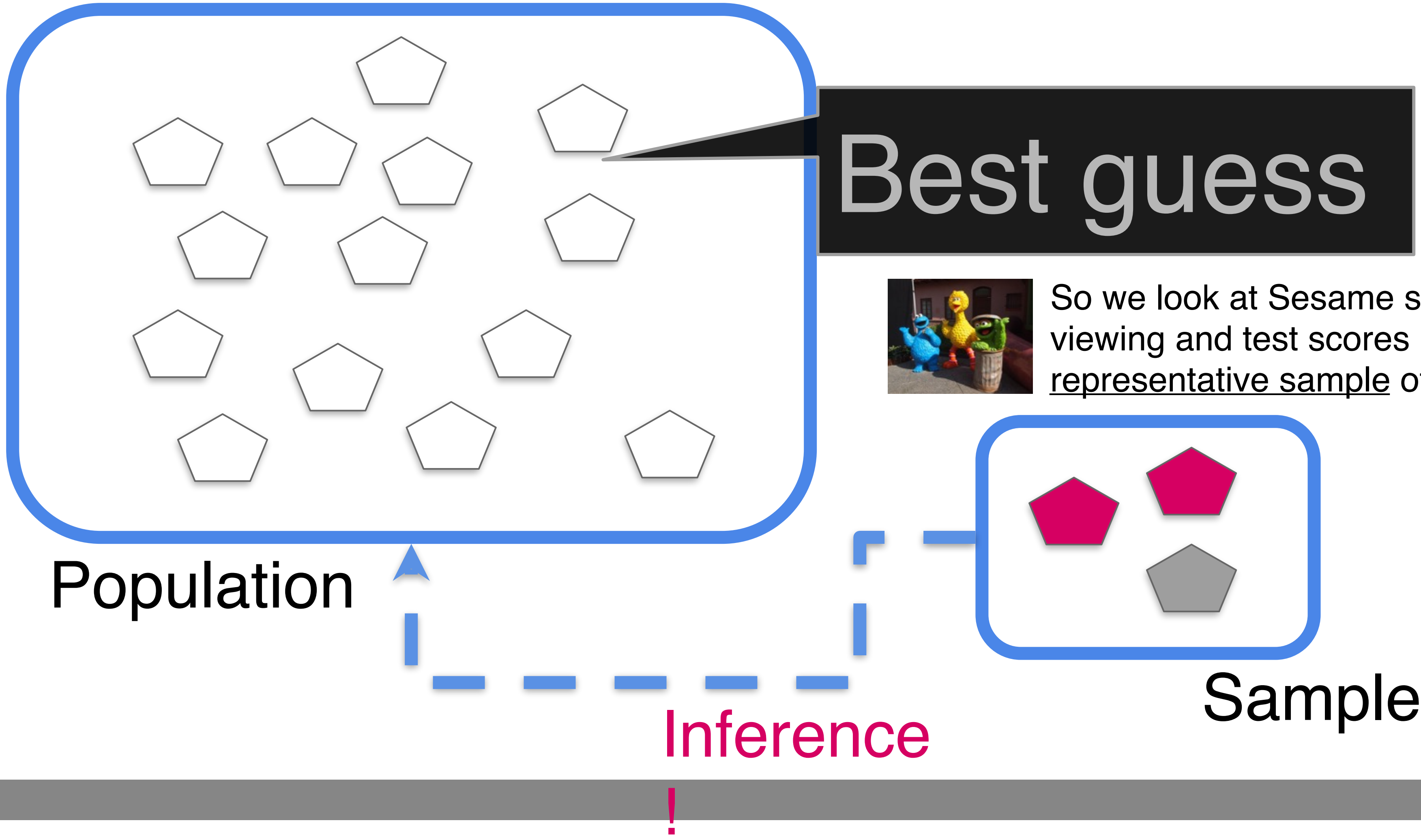
Based on the relationship we see in our sample, we can infer the answer to our question in our population



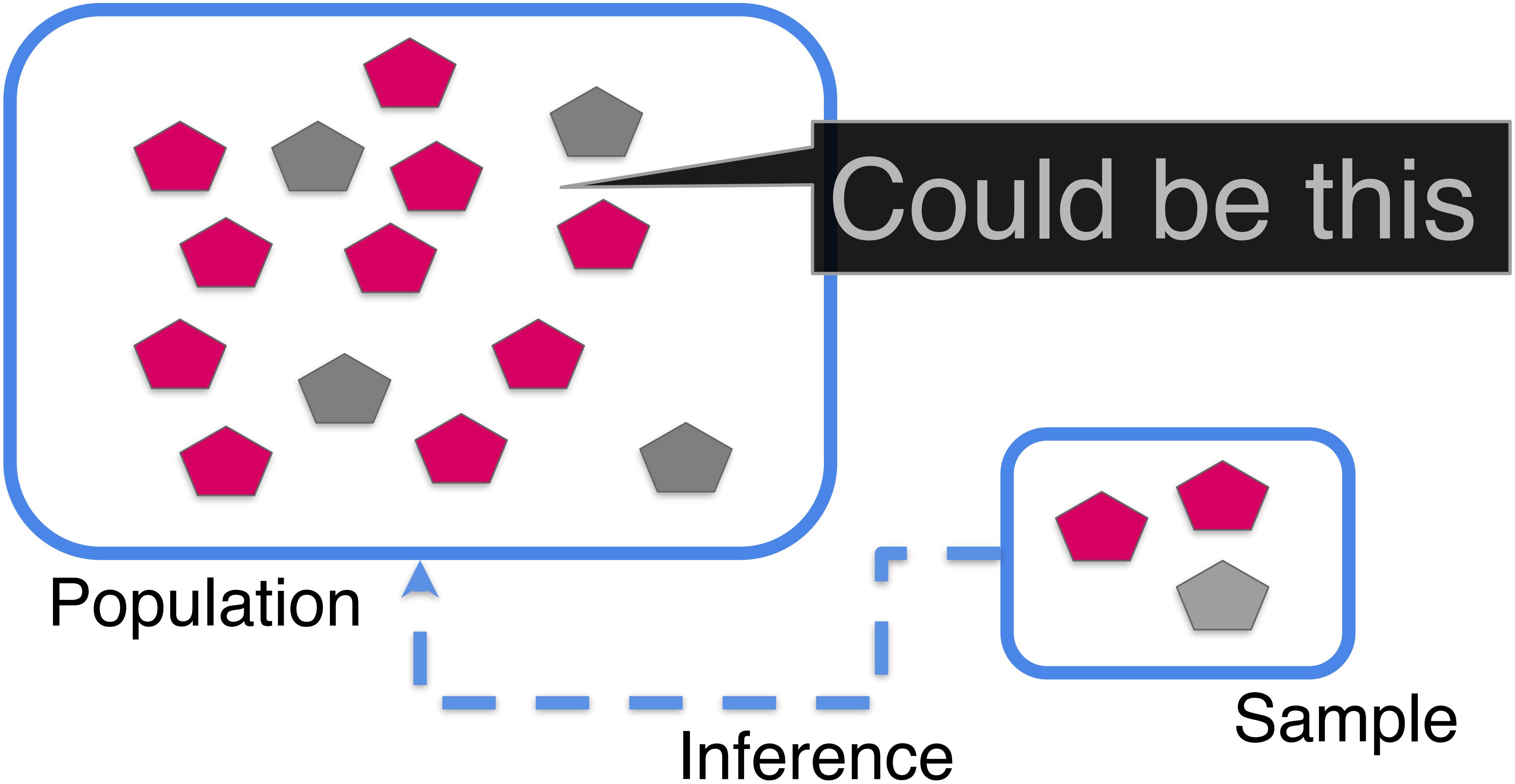
So we look at Sesame street viewing and test scores in a <u>representative sample</u> of kids

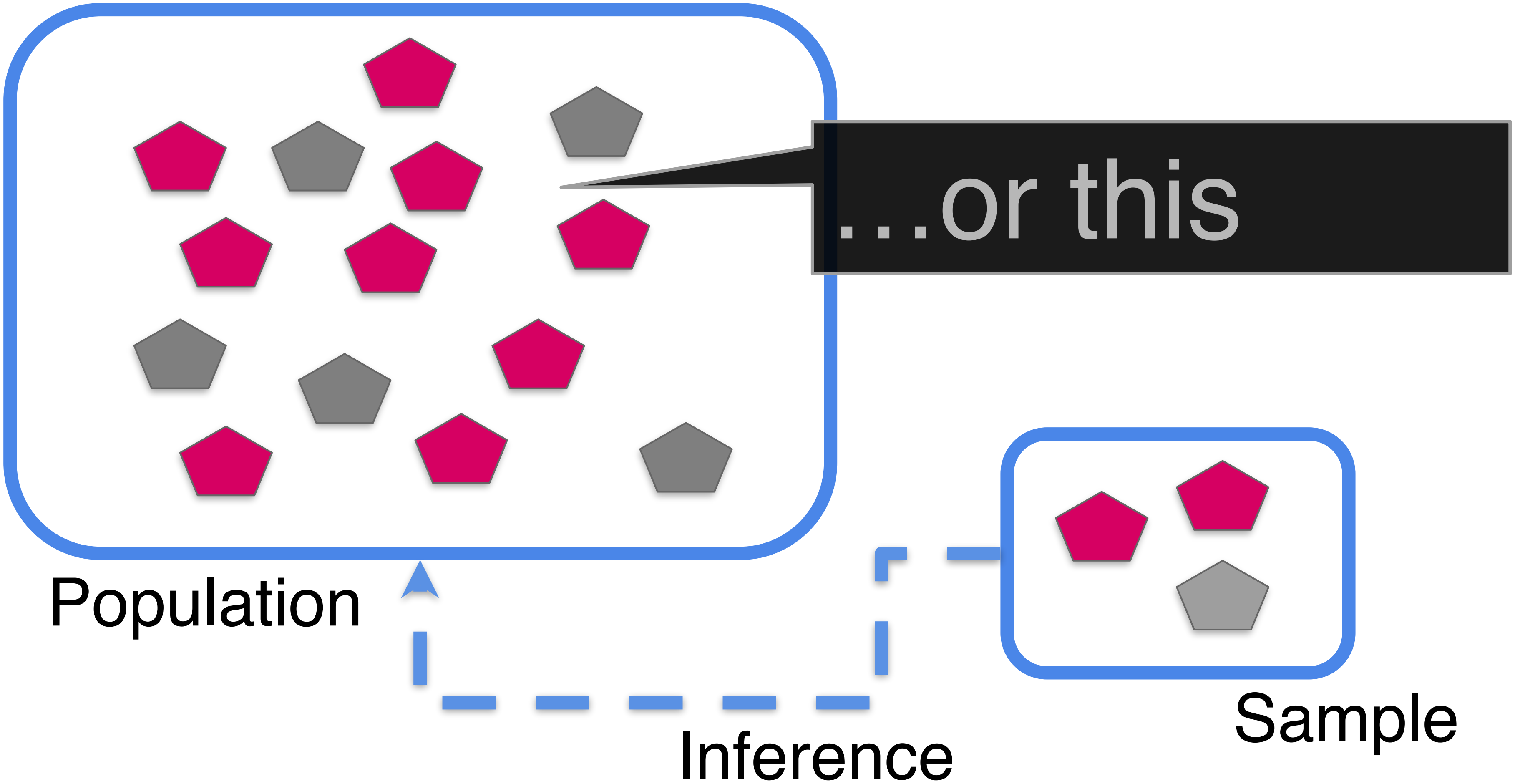


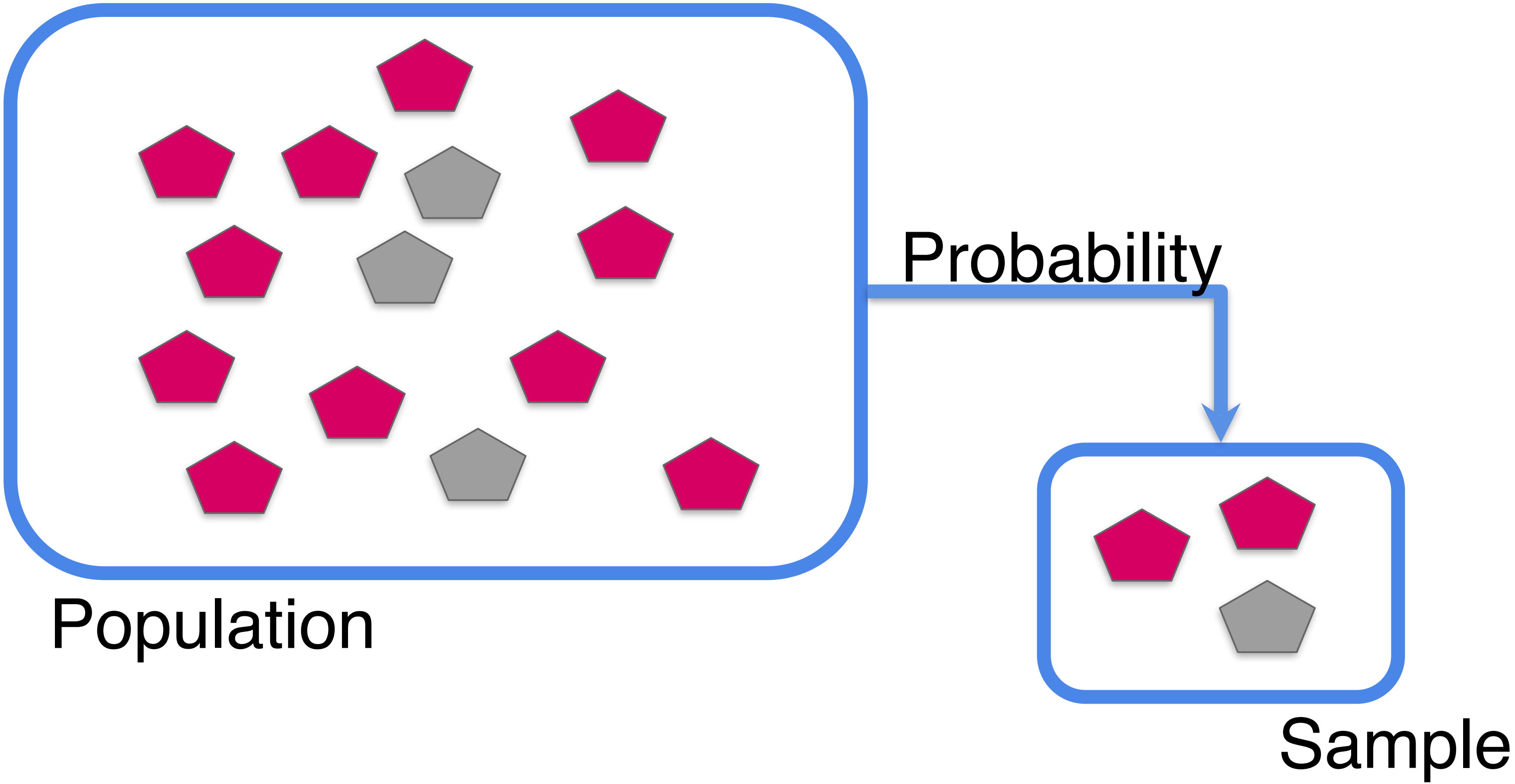


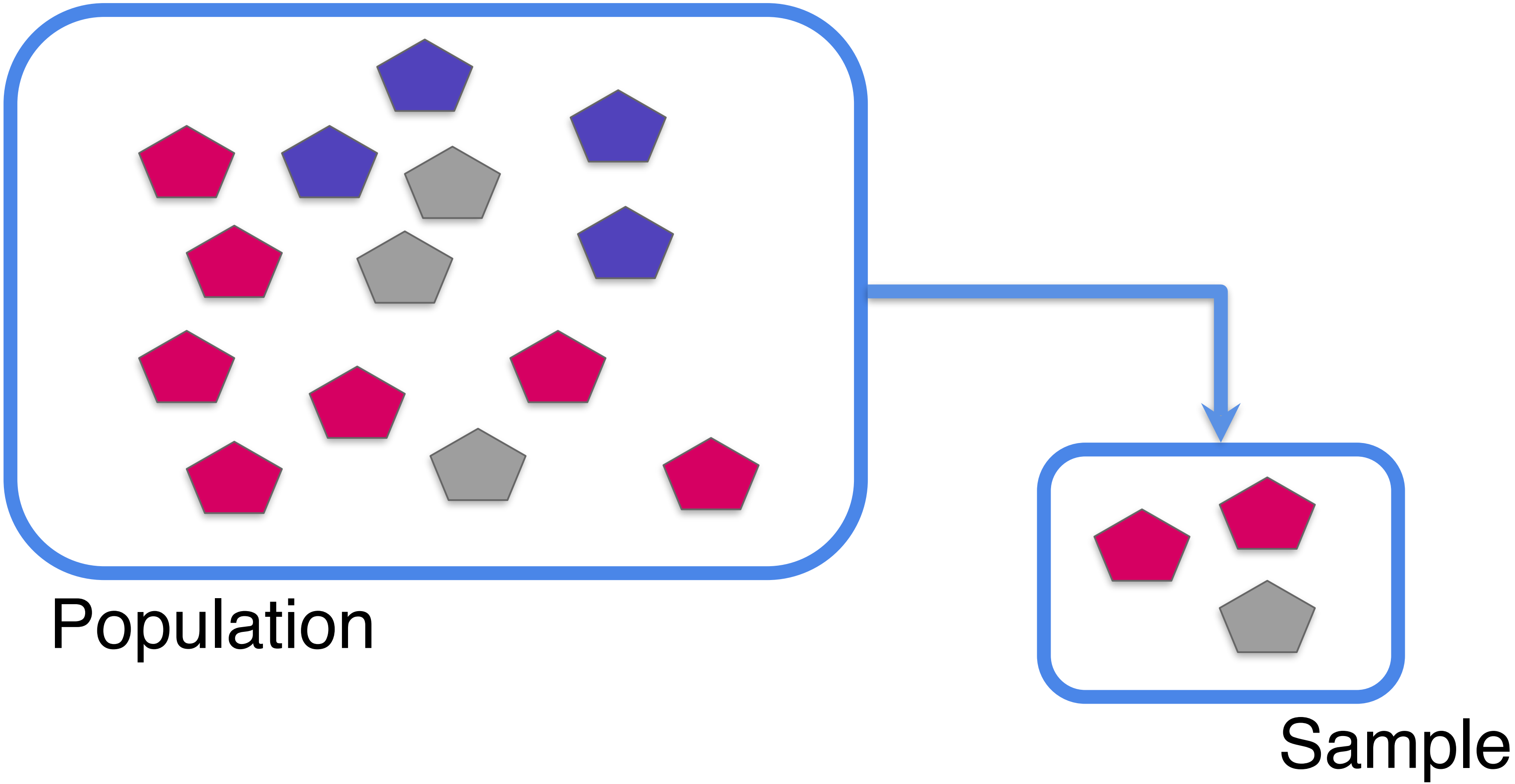


So we look at Sesame street viewing and test scores in a representative sample of kids

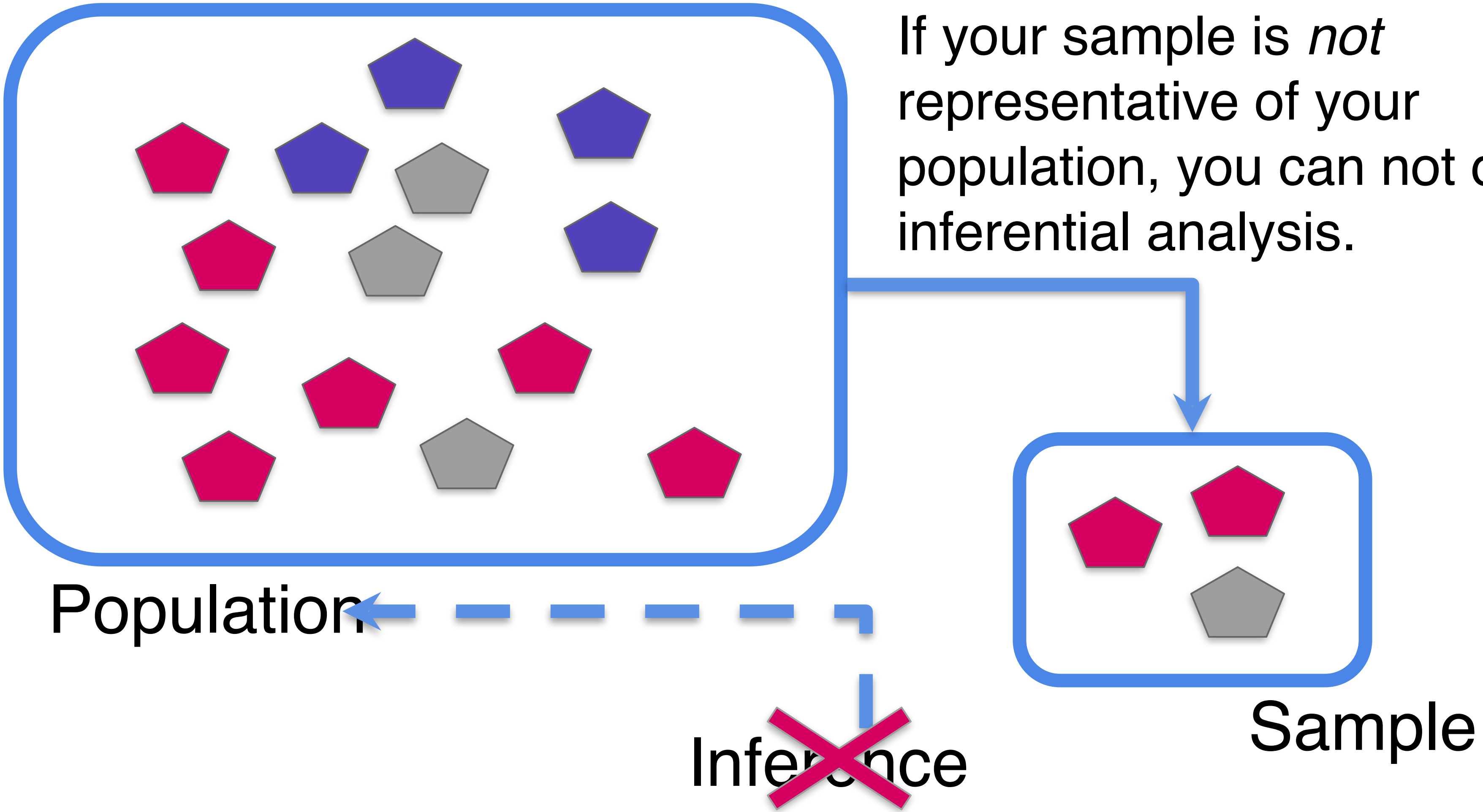

Could be this
Population
Inference
Sample


...or this
Population
Inference
Sample


Probability
Population
Sample


Population
Sample

If your sample is *not* representative of your population, you can not do inferential analysis.



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 VARIABLES

i.e. t-test, ANOVA

REGRESSION

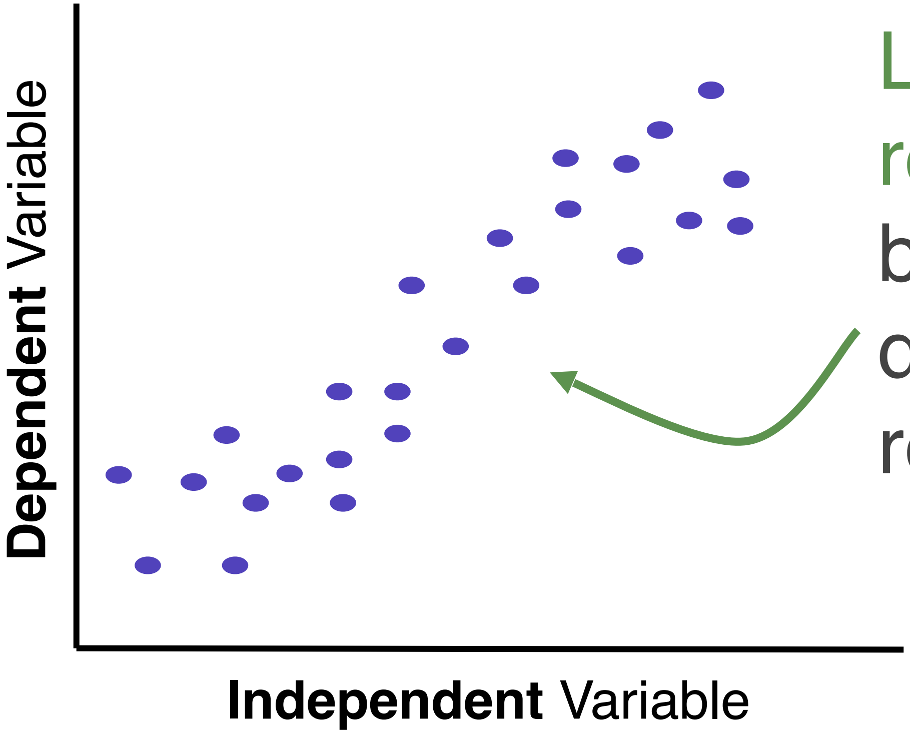
DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

I.e. simple regression, multiple regression

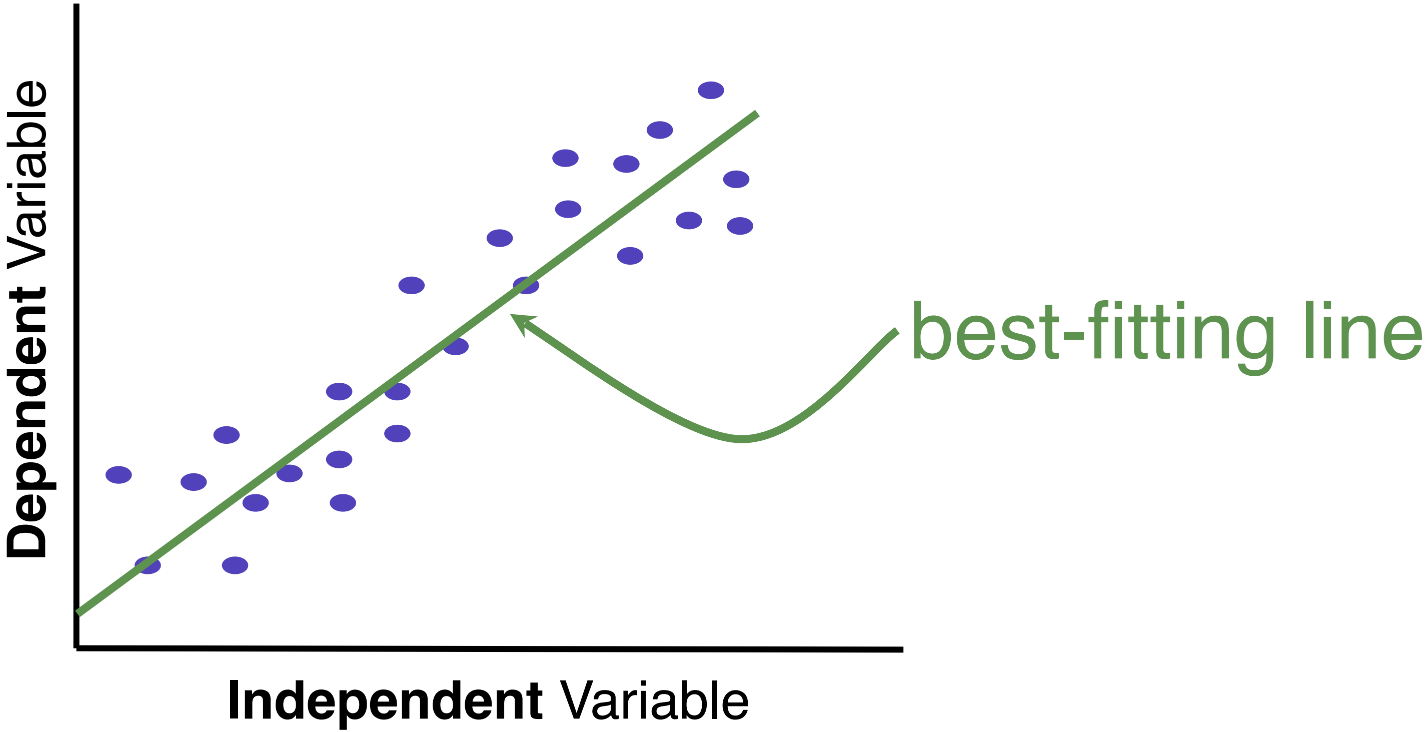
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

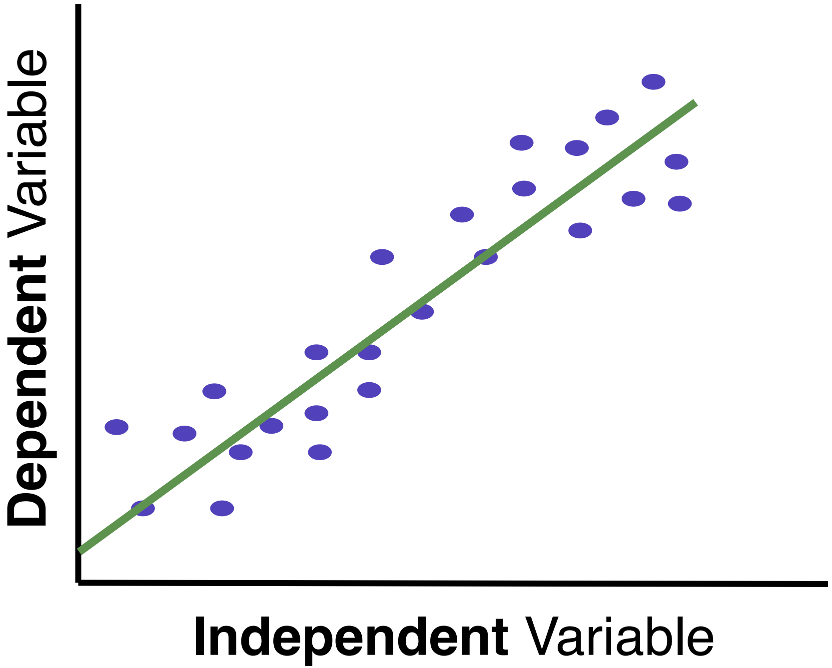
i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test



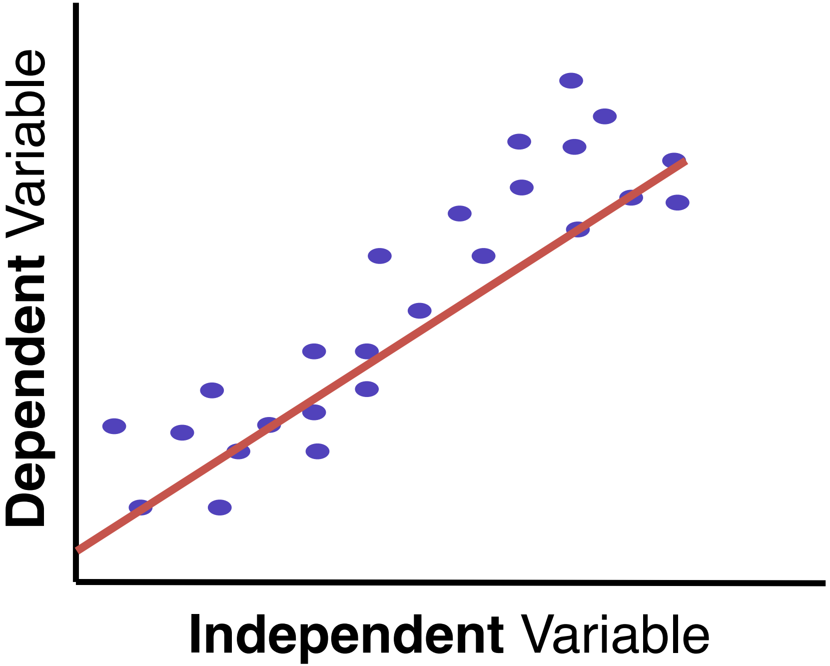
Linear regression can be used to describe this relationship

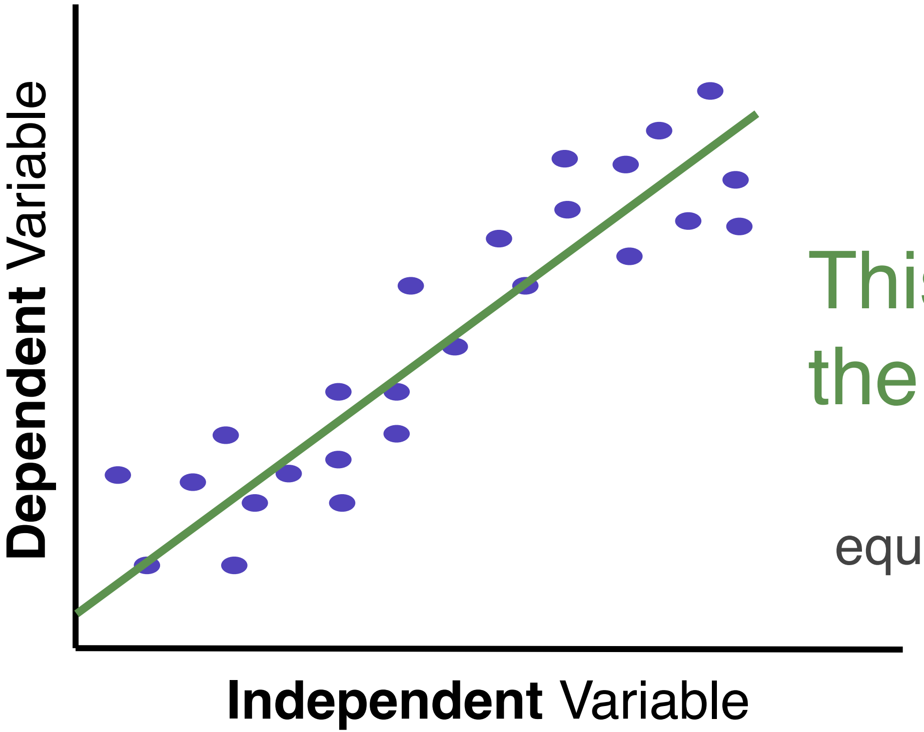

Dependent Variable
Independent Variable
best-fitting line

Best-fitting line



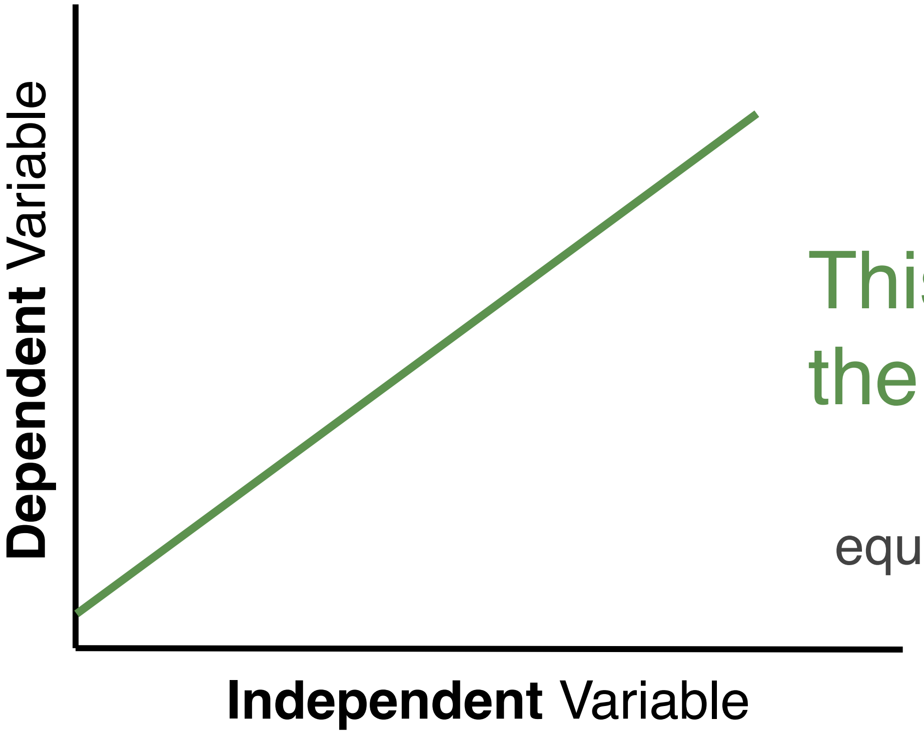
NOT a best-fitting line





This line is a model of the data

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

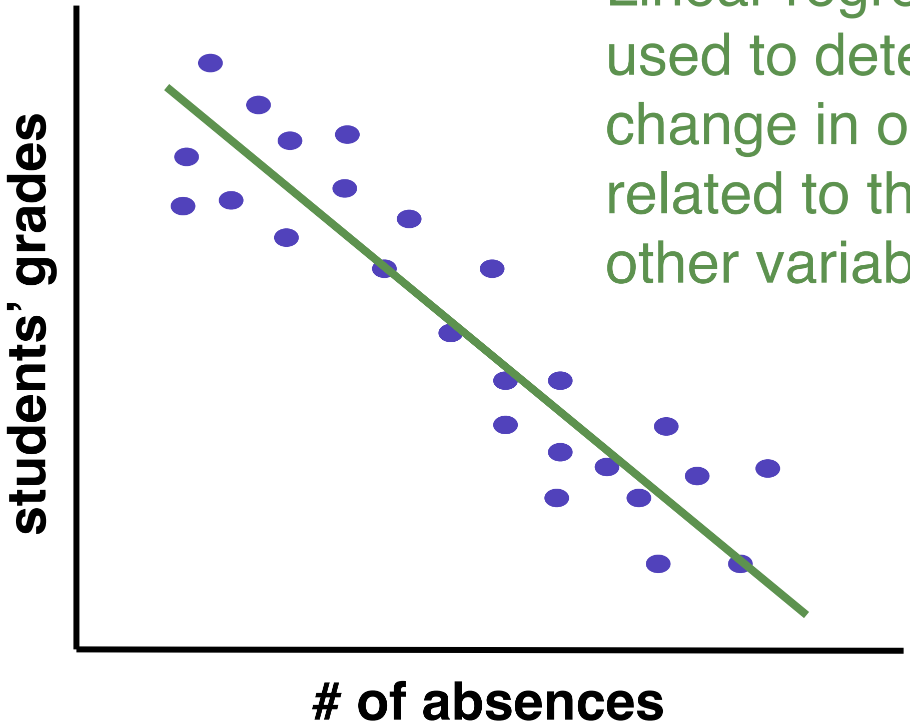


This line is a model of the data

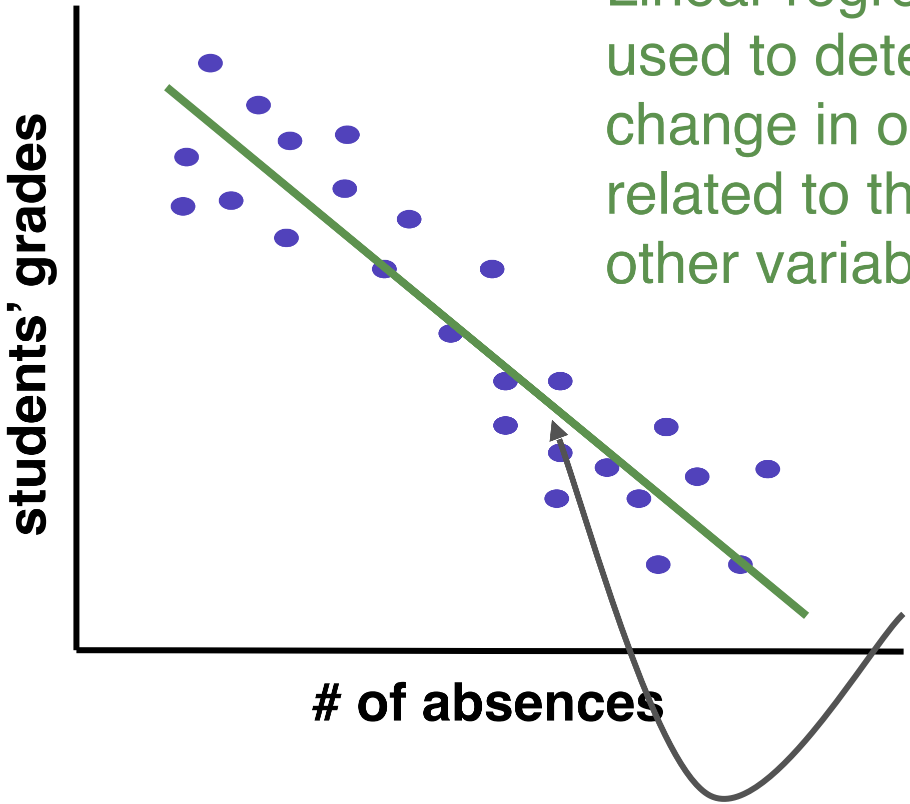
Models are mathematical equations generated to *represent* the real life situation

“*All models are wrong, but some are useful*”

-George Box (British Statistician, *JASA* 1976)

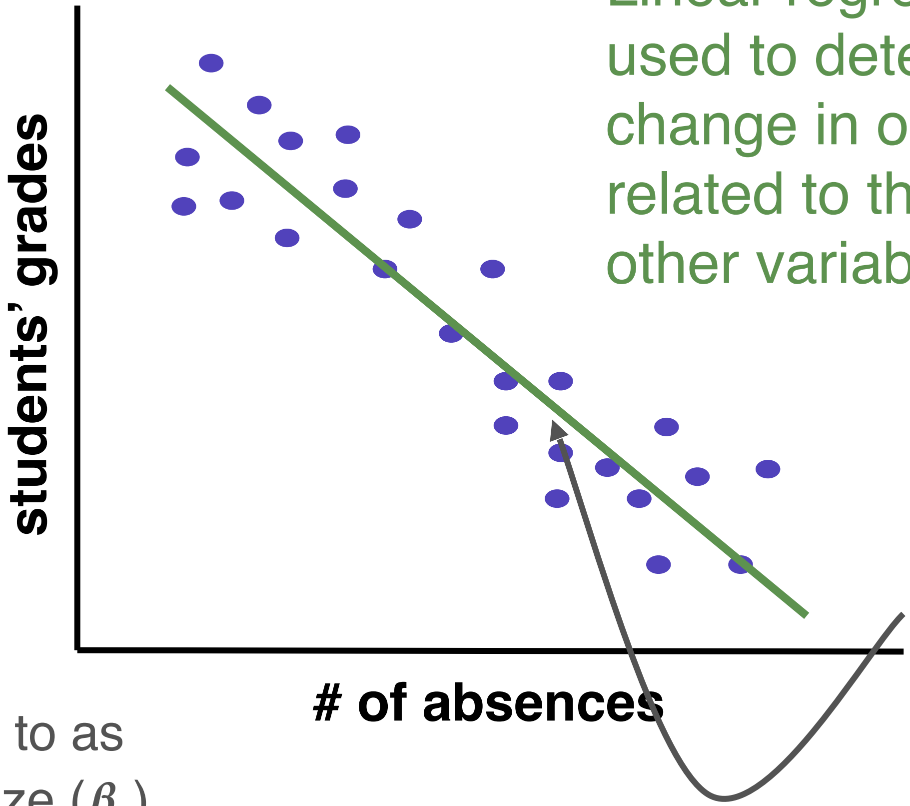


Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



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

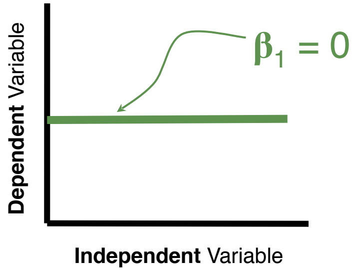


Linear regression can be used to determine whether a change in one variable is related to the change in the other variable

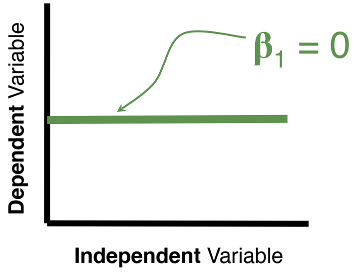
The magnitude of the relationship is measured by the slope of the line

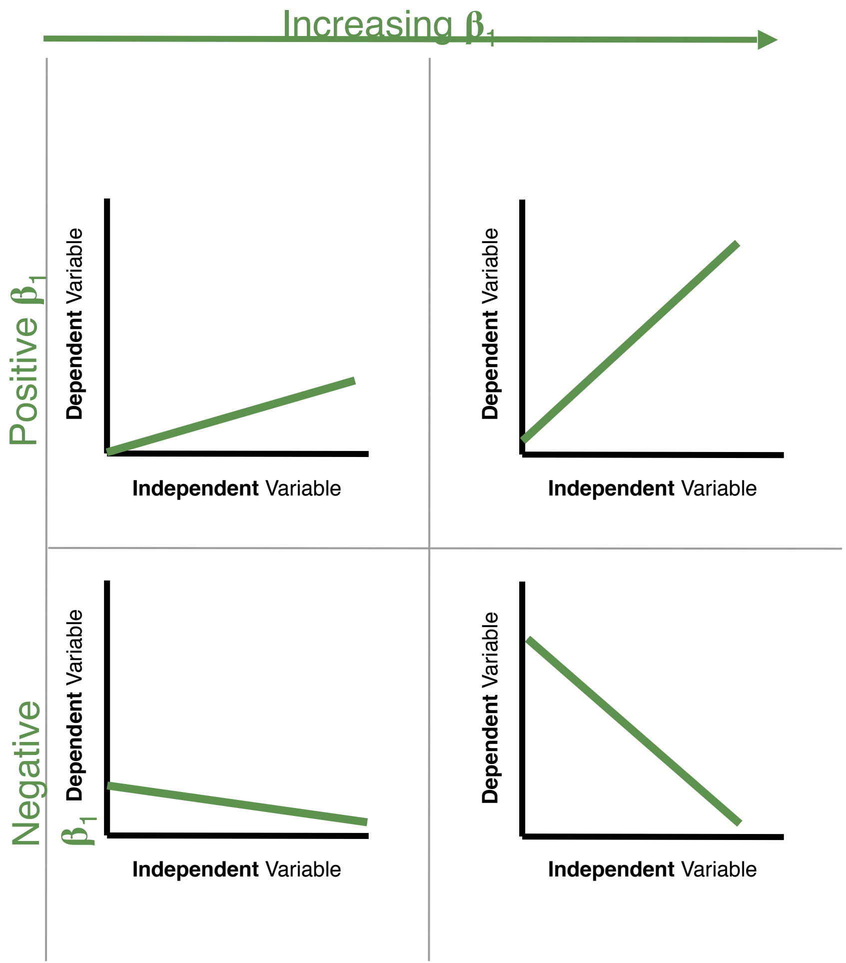
This is also referred to as the model's effect size (β_1)

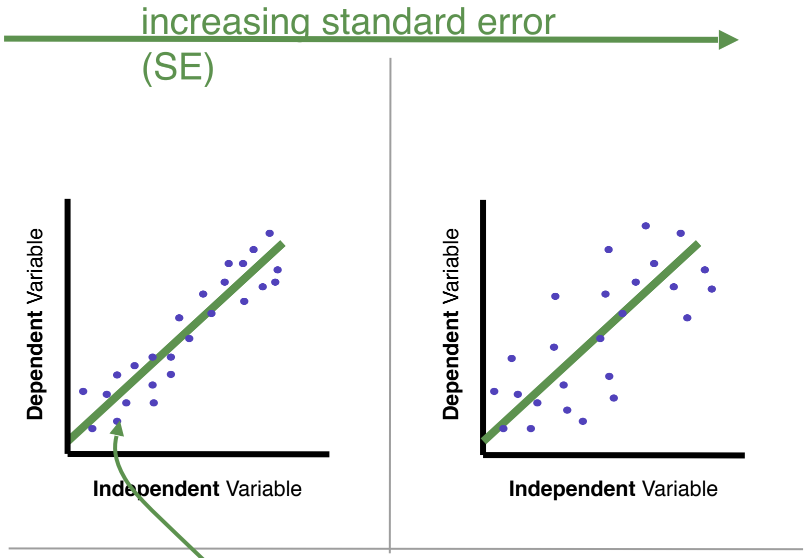
Effect size (β_1) can be estimated using the slope of the line



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



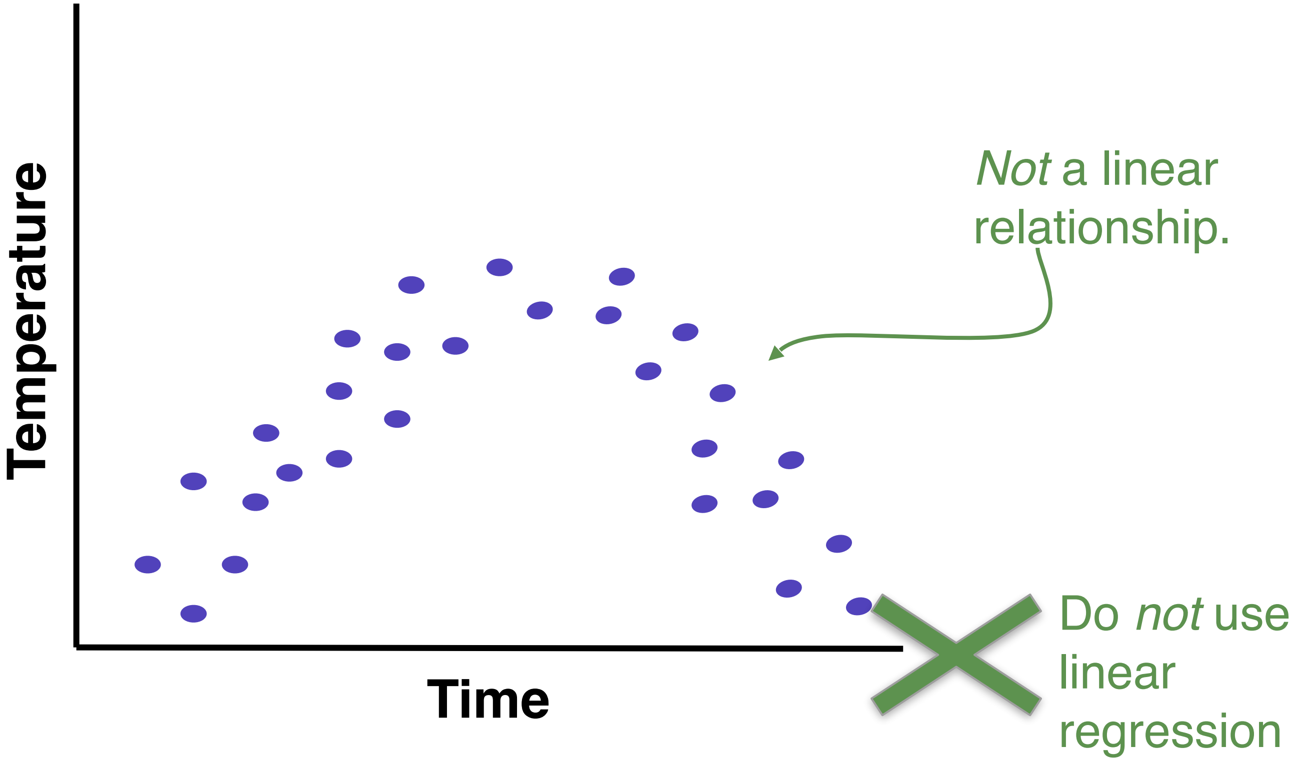




The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

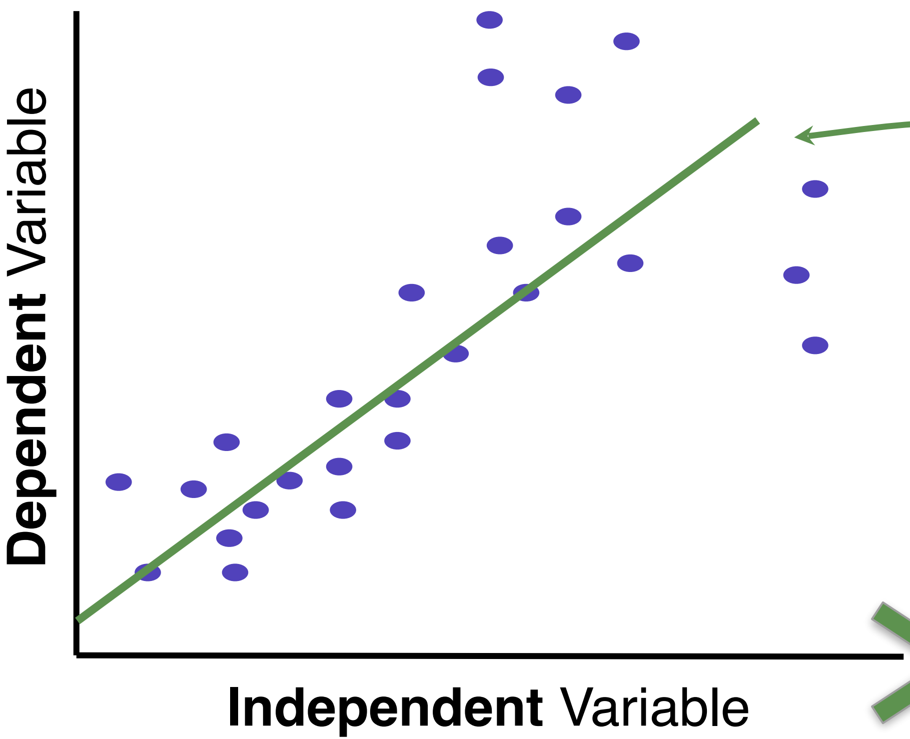
Assumptions of linear regression

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


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

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

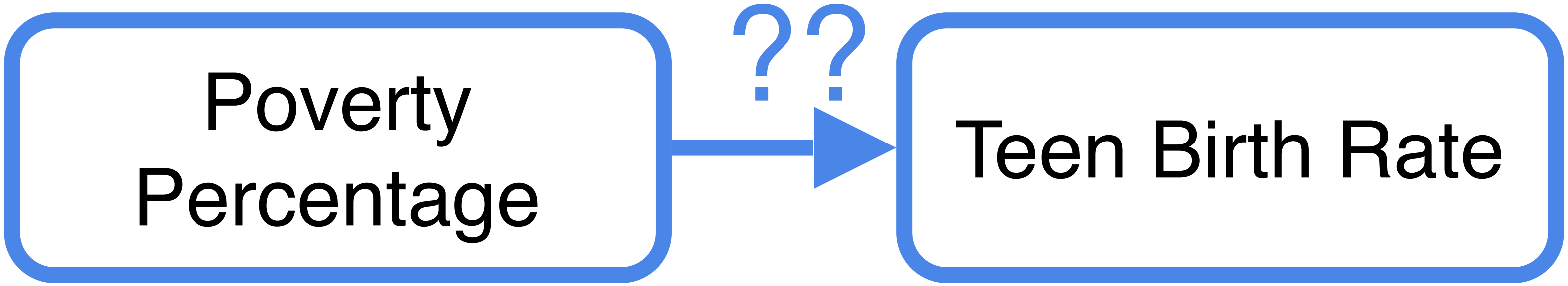
Autocorrelation occurs when the observations are *not* independent of one another (i.e. stock prices)



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?



Null Hypothesis:

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

Alternative Hypothesis:

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

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

What’s your hypothesis?

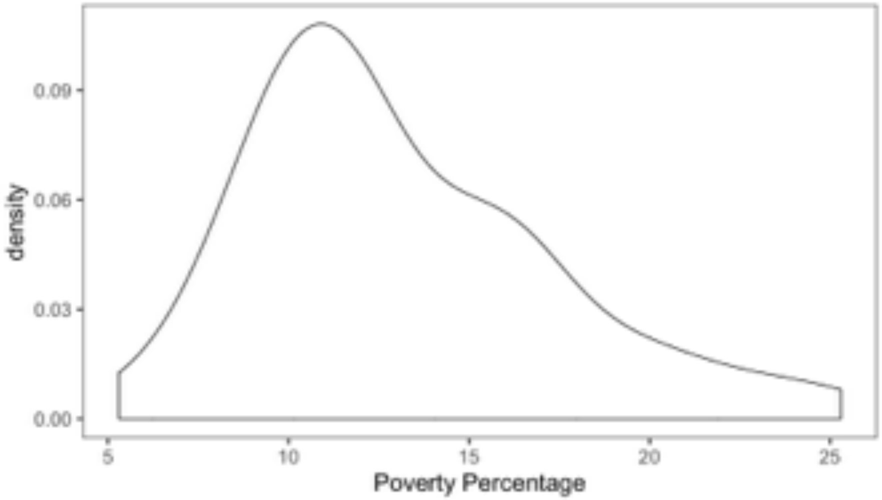
A
No
relationship

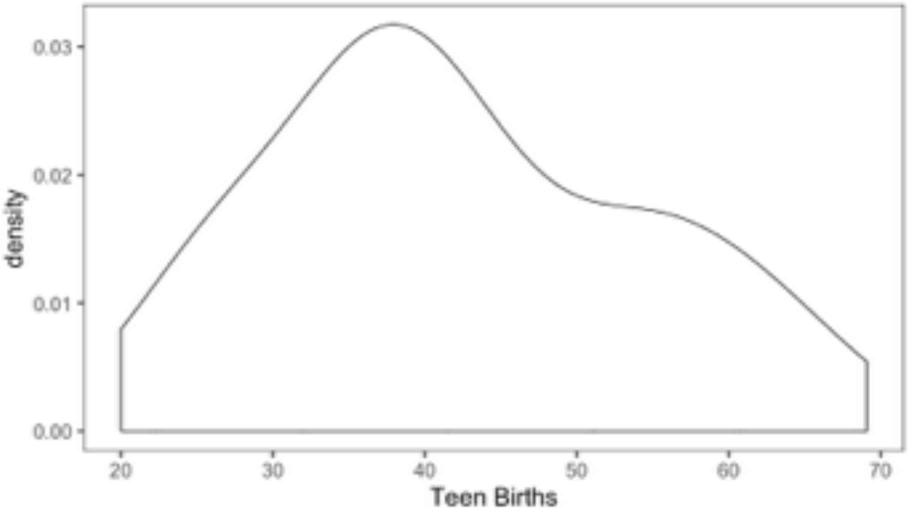
B
Positive
relationshi
p

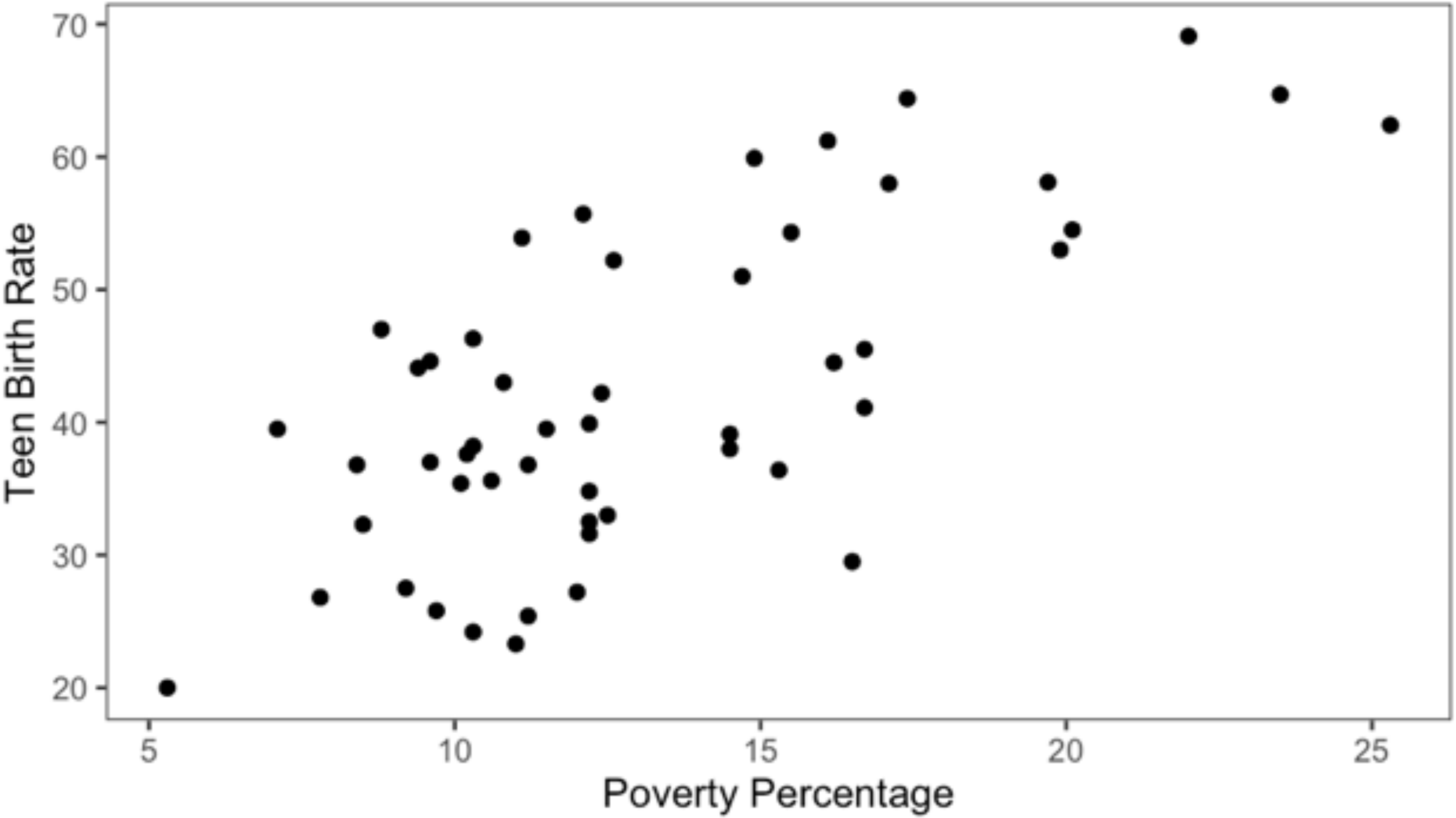
C
Negative
relationshi
p

	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

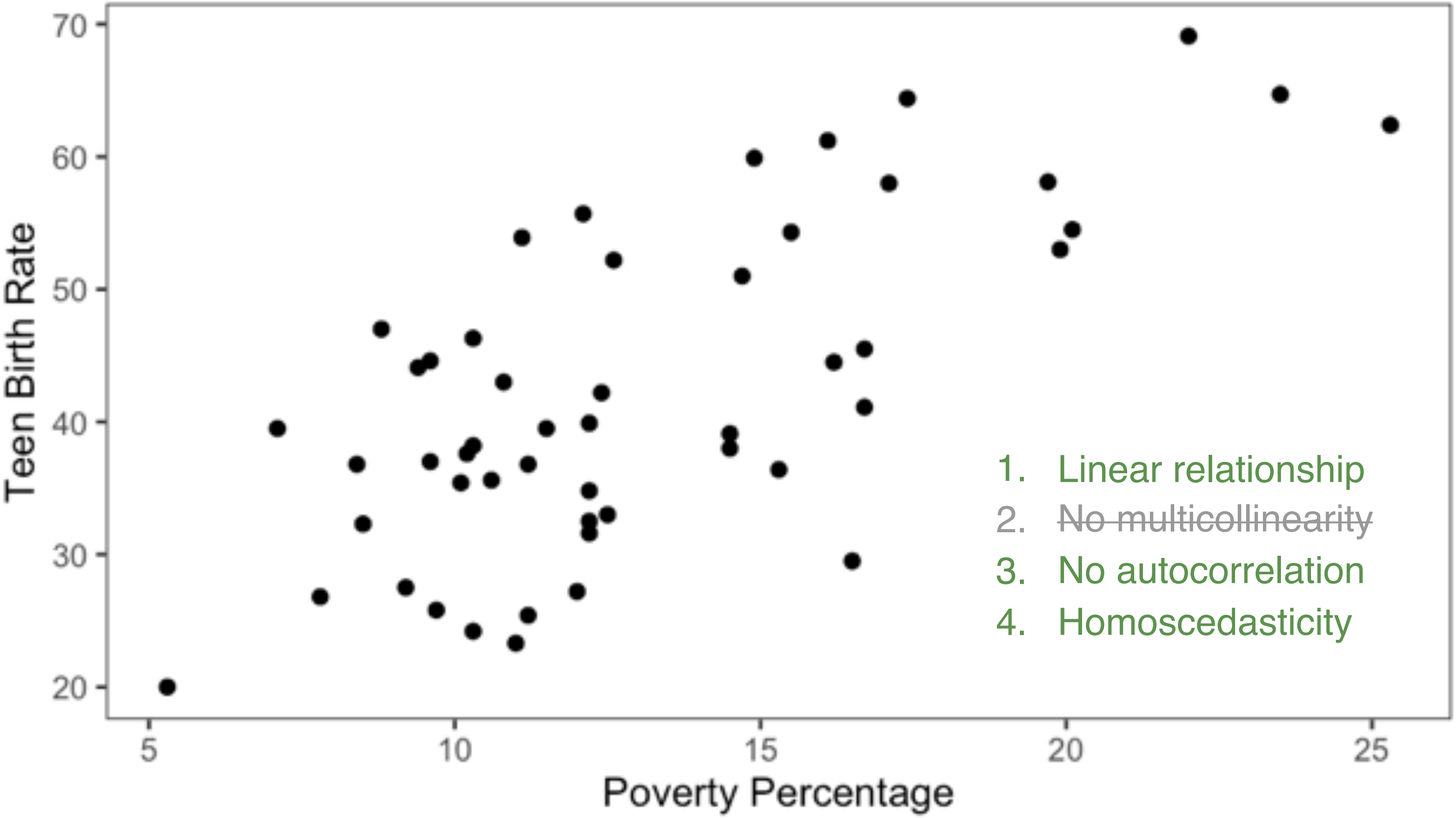
EDA: distributions



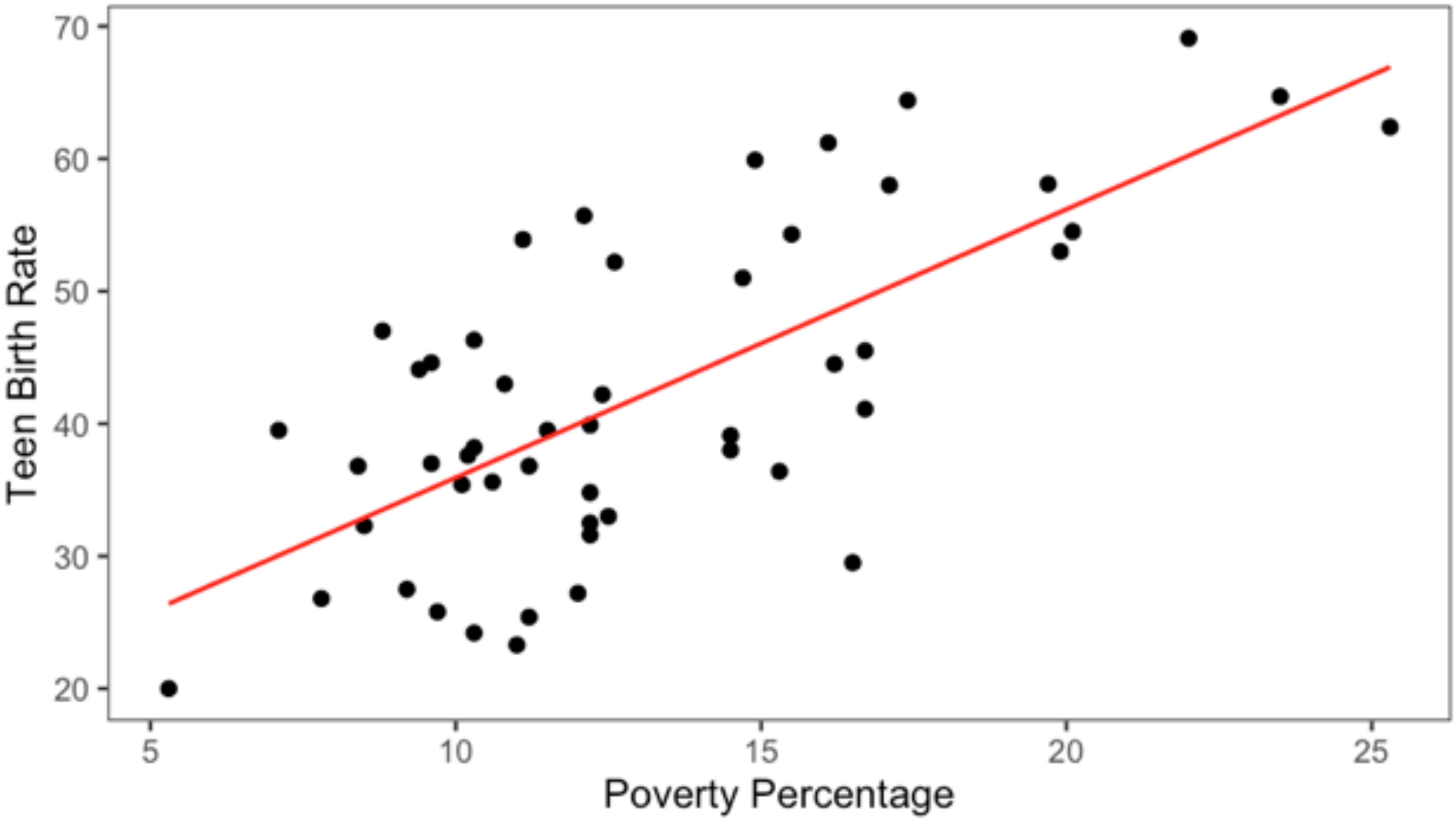




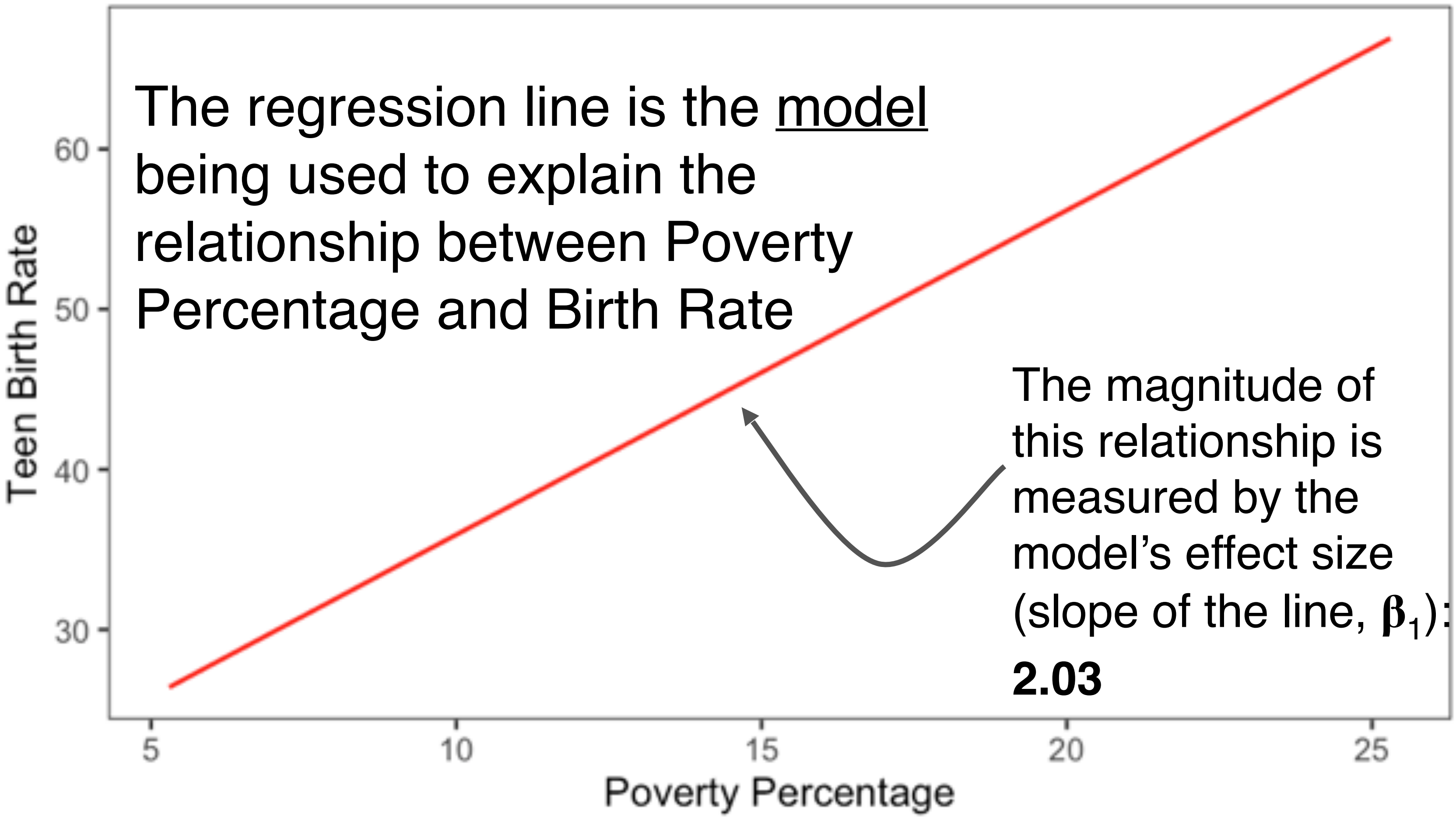
Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.

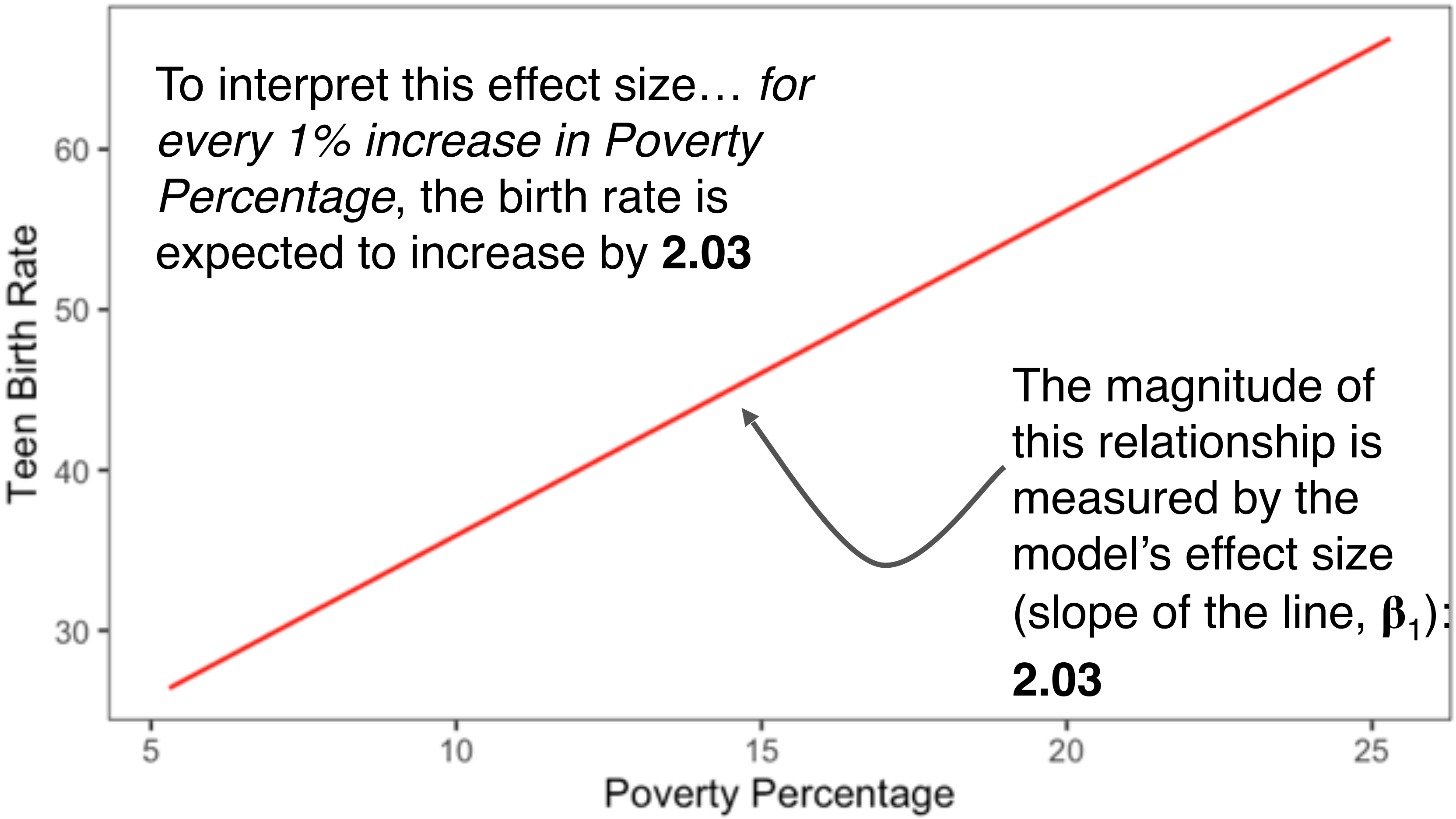


Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard.



70
60
50
40
30
20
Teen Birth Rate
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
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

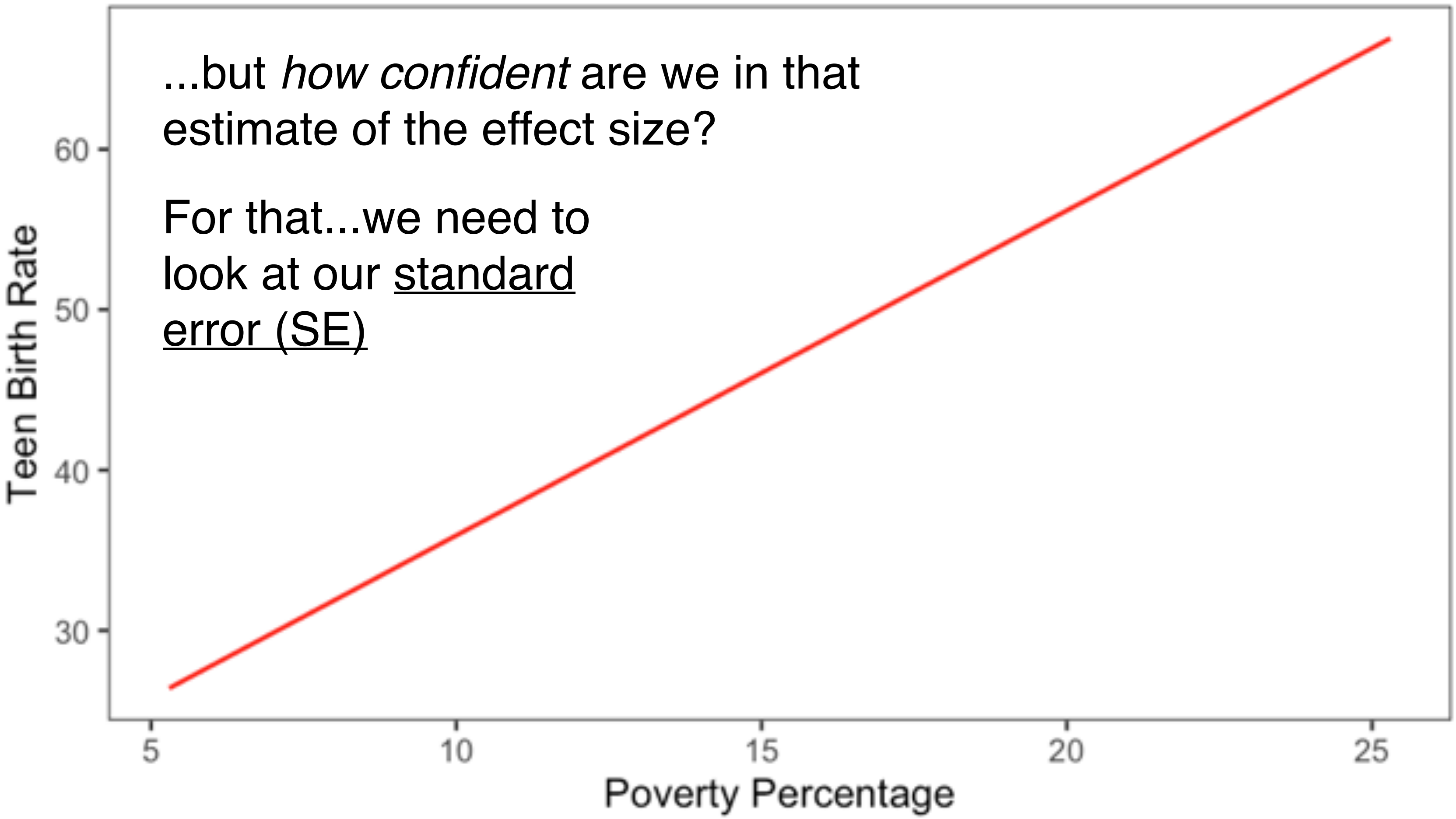


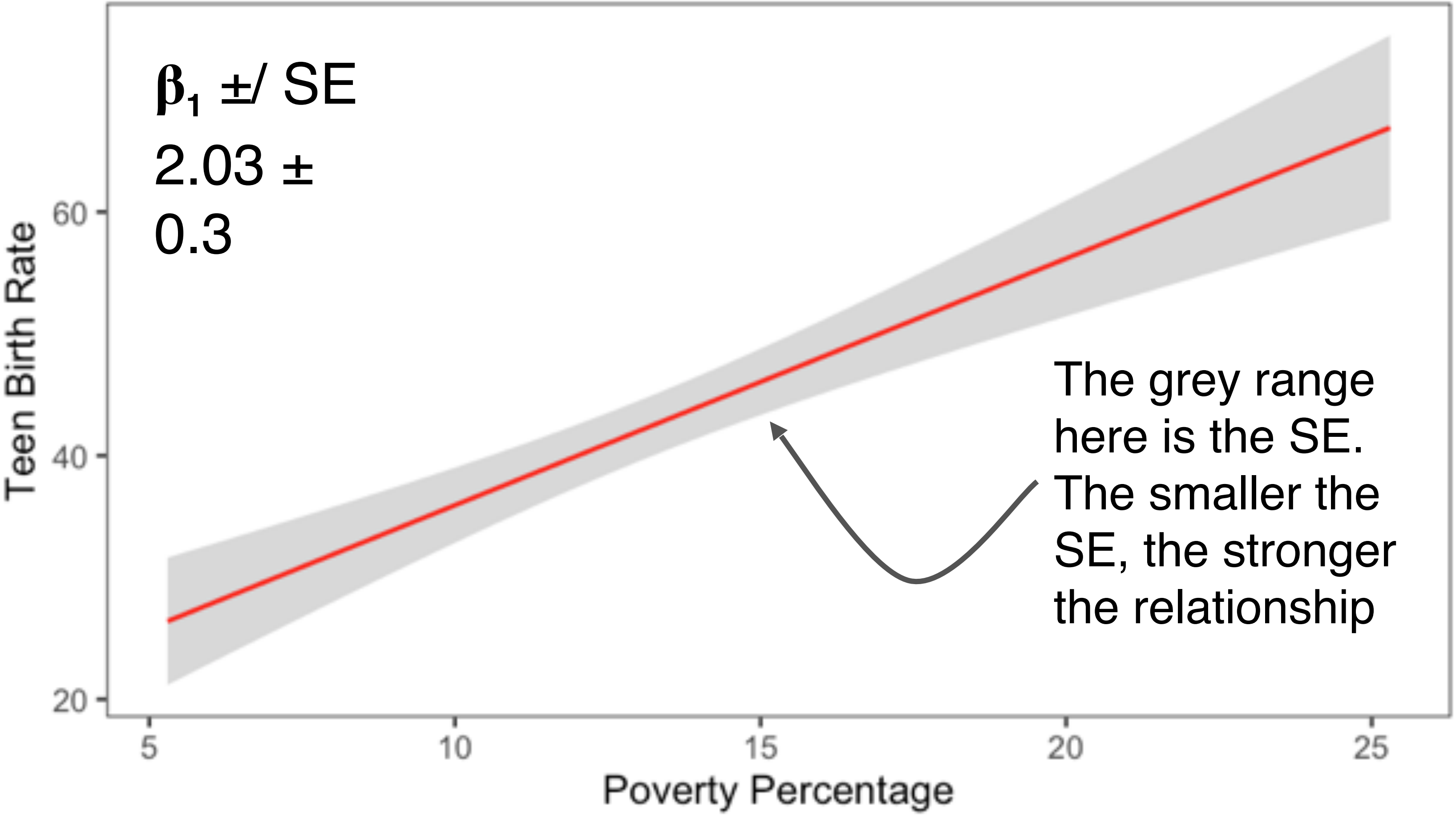
To interpret this effect size… *for every 1% increase in Poverty Percentage*, the birth rate is expected to increase by **2.03**

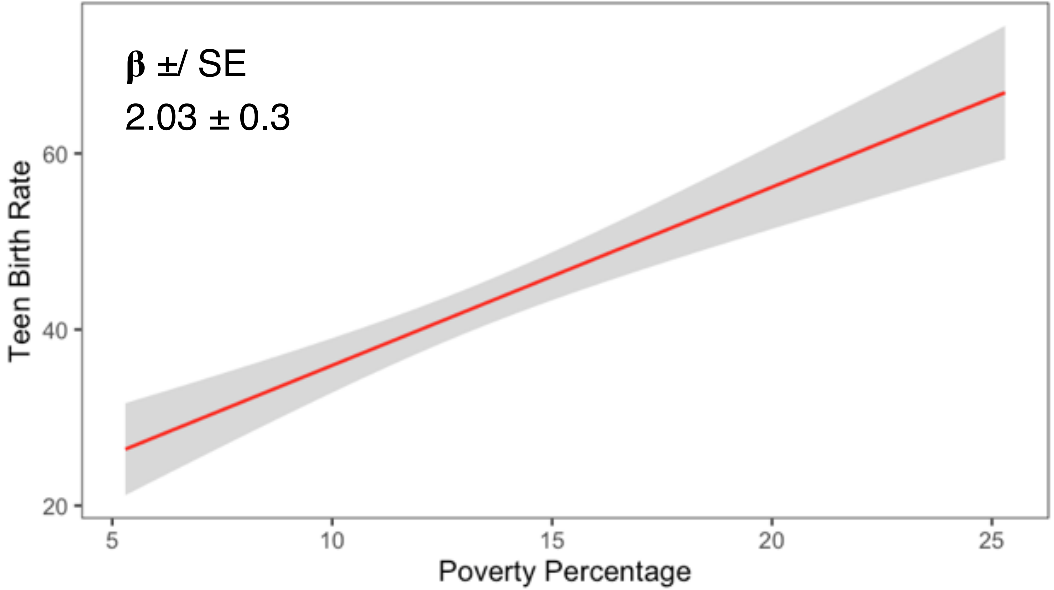
The magnitude of this relationship is measured by the model's effect size (slope of the line, $\boldsymbol{\beta}_1$): **2.03**

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

For that...we need to look at our <u>standard error (SE)</u>



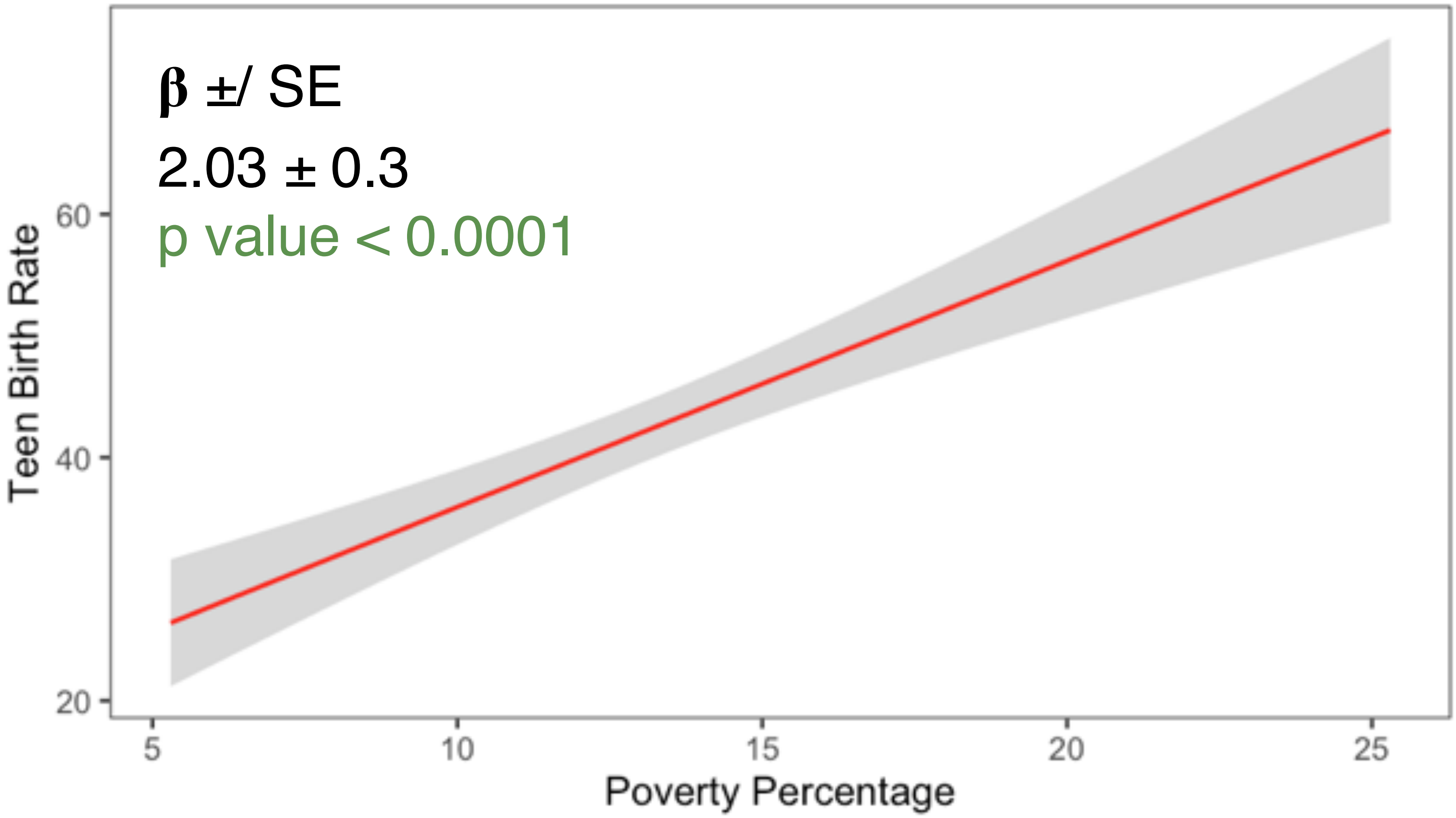

β1 ±/ SE
2.03 ±
0.3
The grey range here is the SE. The smaller the SE, the stronger the relationship
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage



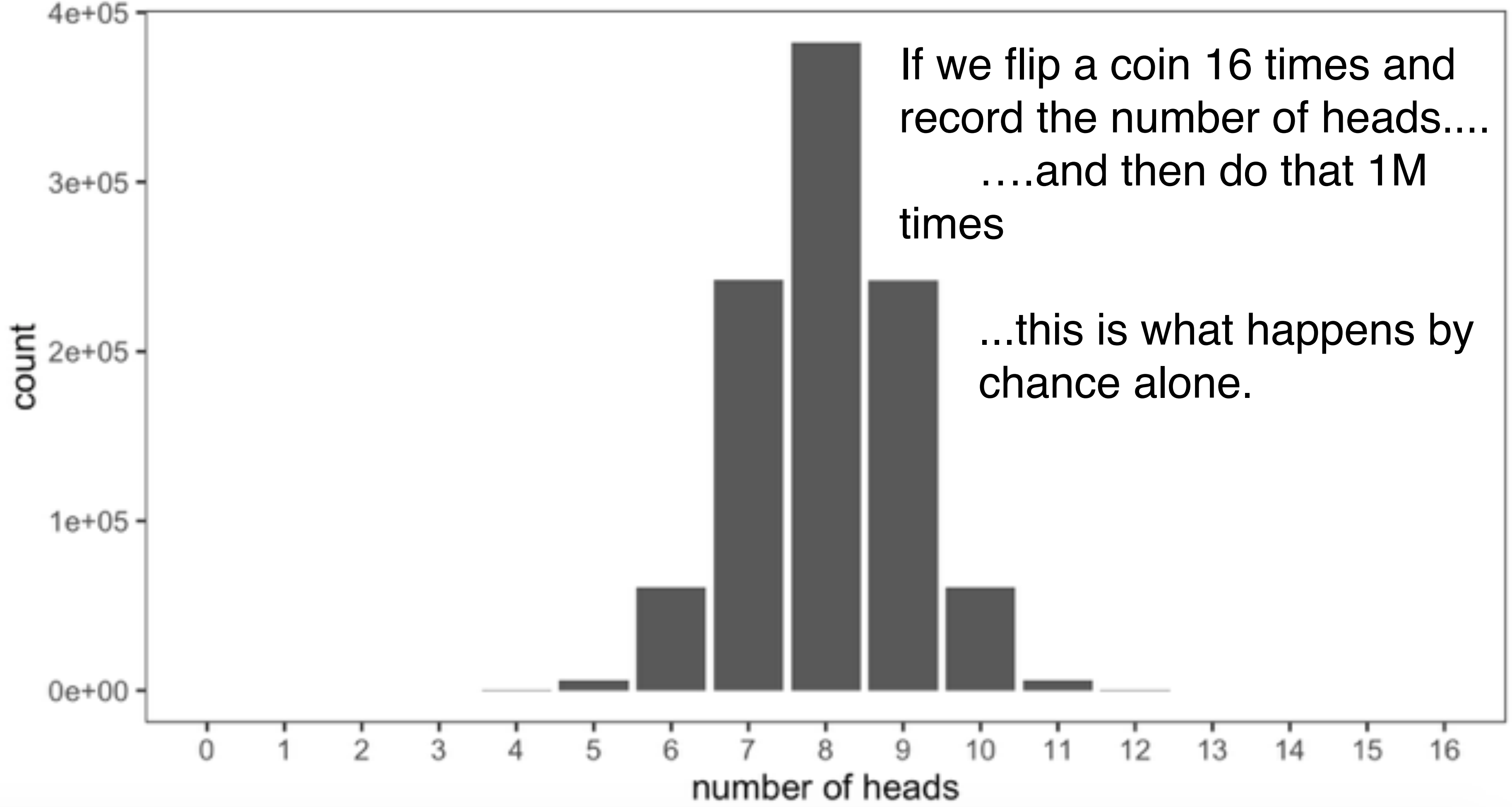
If there were a stronger effect of Poverty on Birth rate, what would β_1 be?

A
< 2.03

B
> 2.03

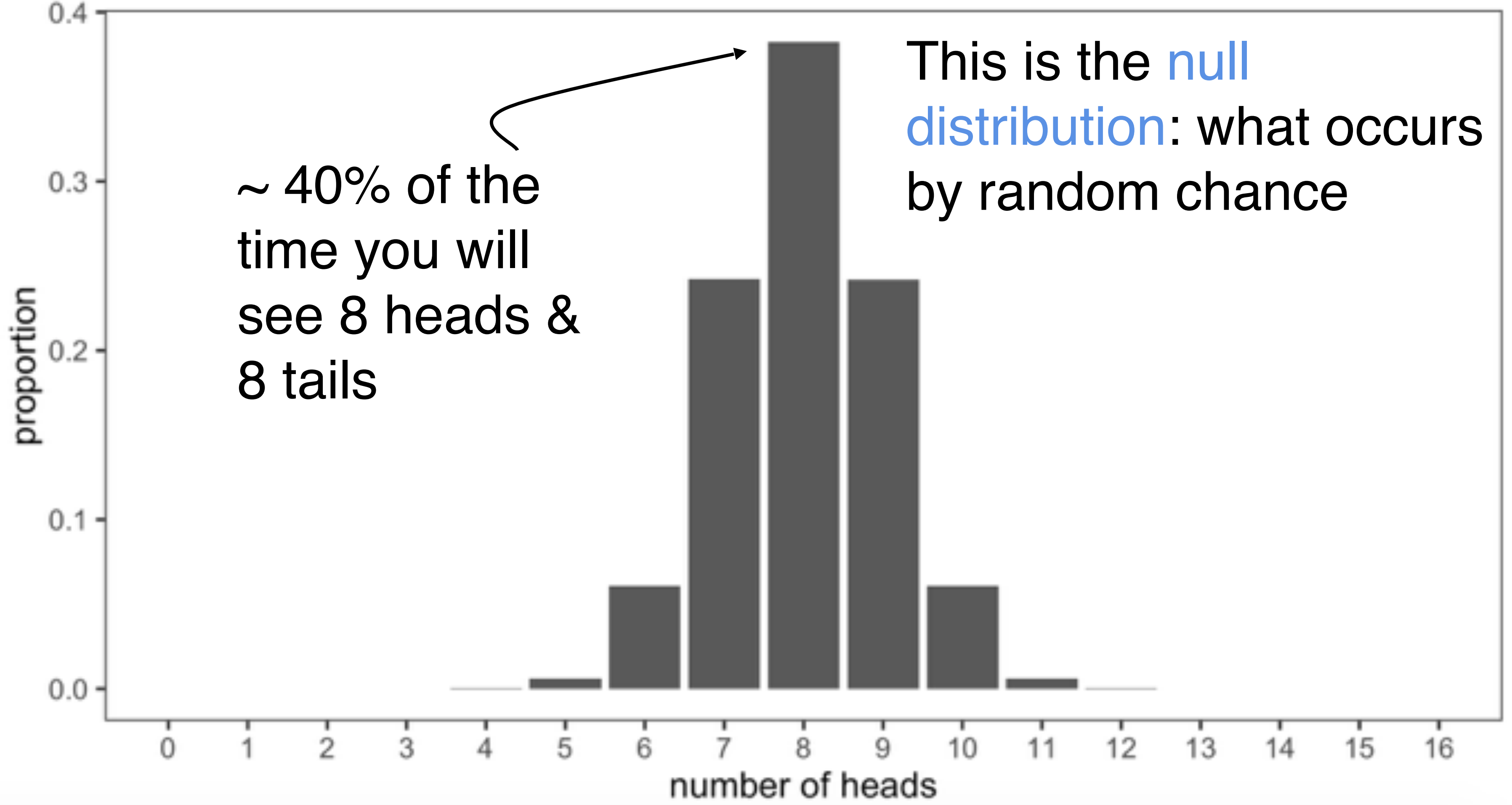

β ±/ SE
2.03 ± 0.3
p value < 0.0001
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage

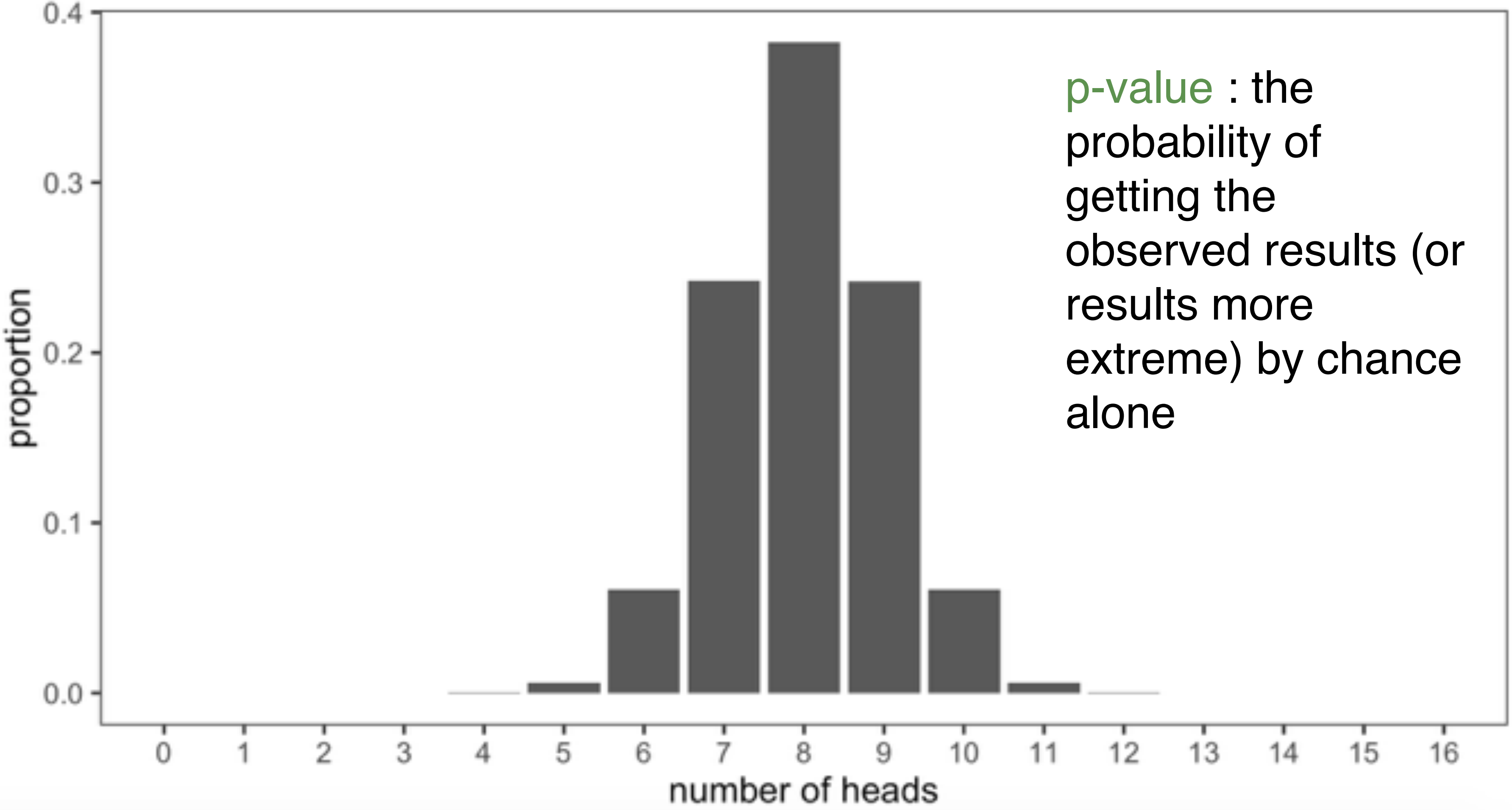
p-value : the probability of getting the observed results (or results more extreme) by chance alone



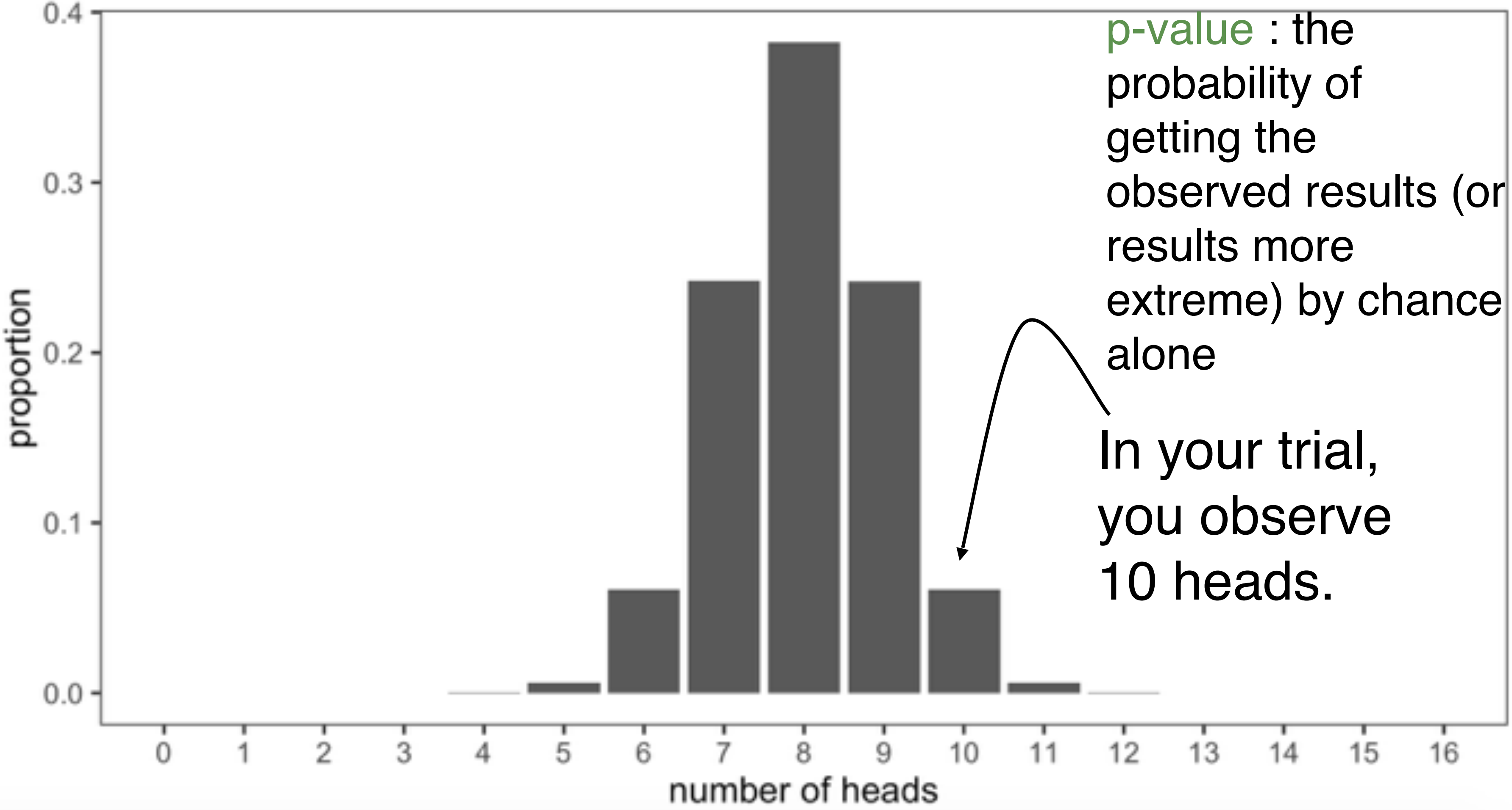
If we flip a coin 16 times and record the number of heads....
….and then do that 1M times

...this is what happens by chance alone.


~ 40% of the time you will see 8 heads & 8 tails
This is the null distribution: what occurs by random chance
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

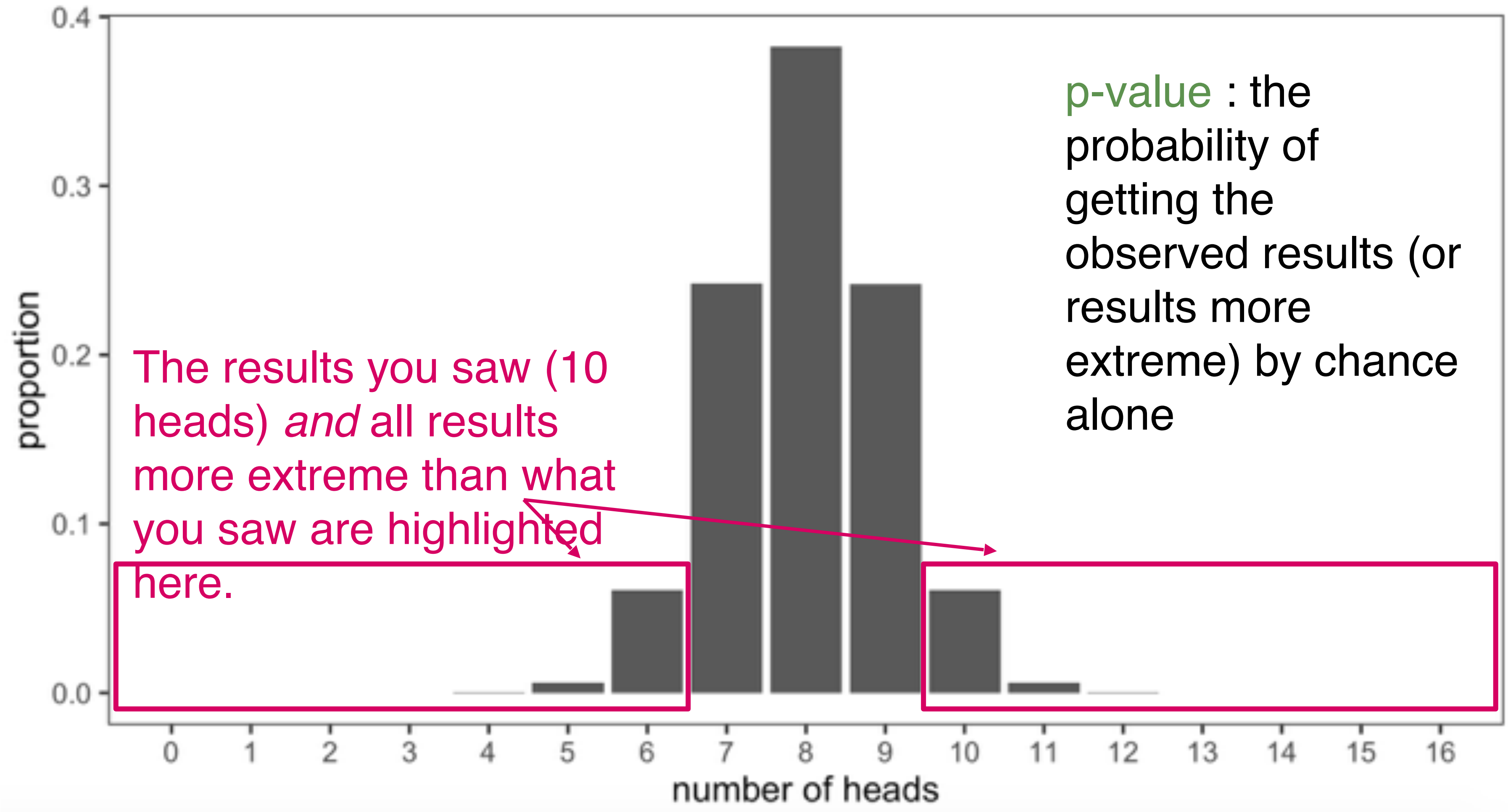


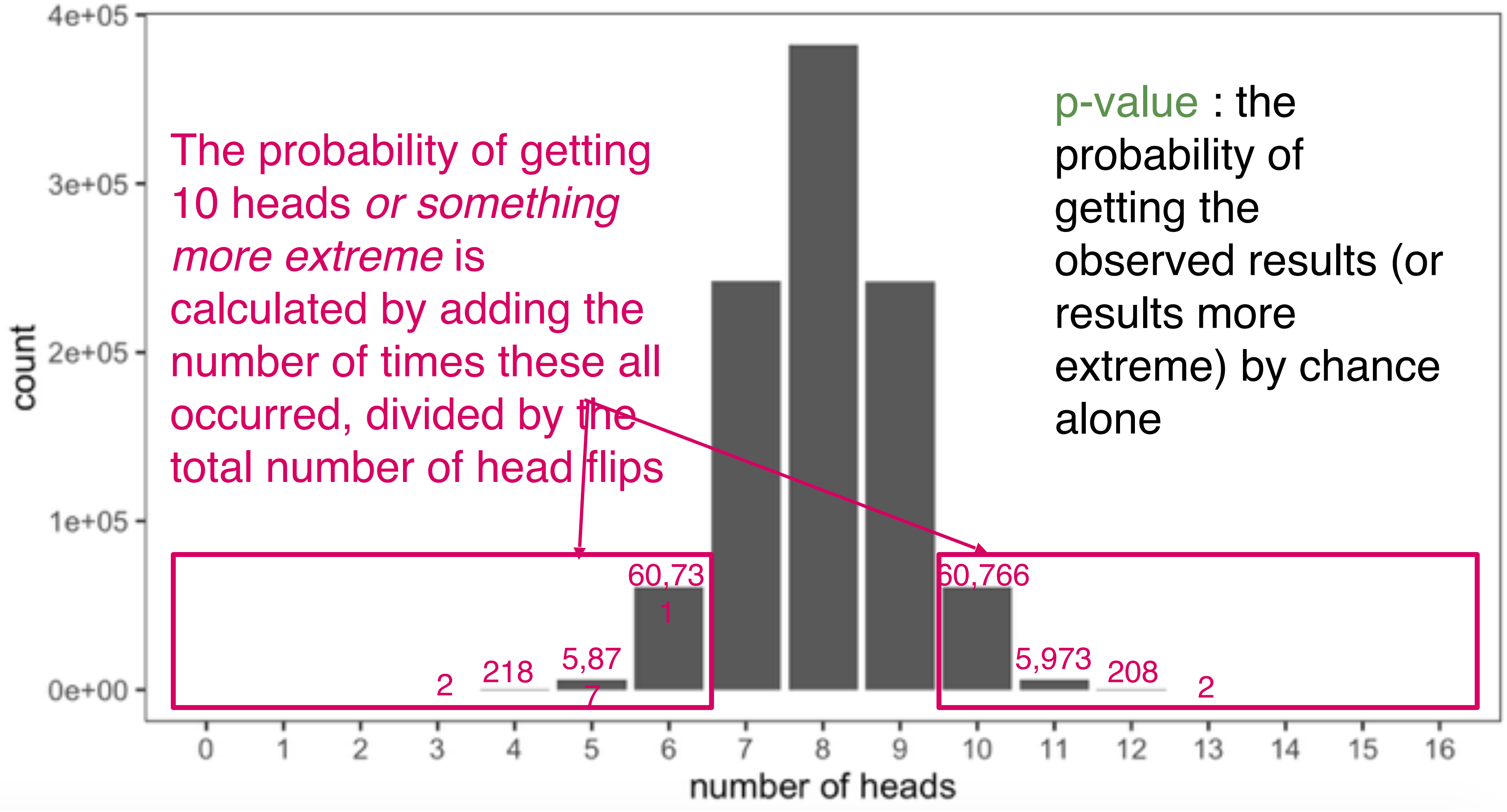
p-value : the probability of getting the observed results (or results more extreme) by chance alone

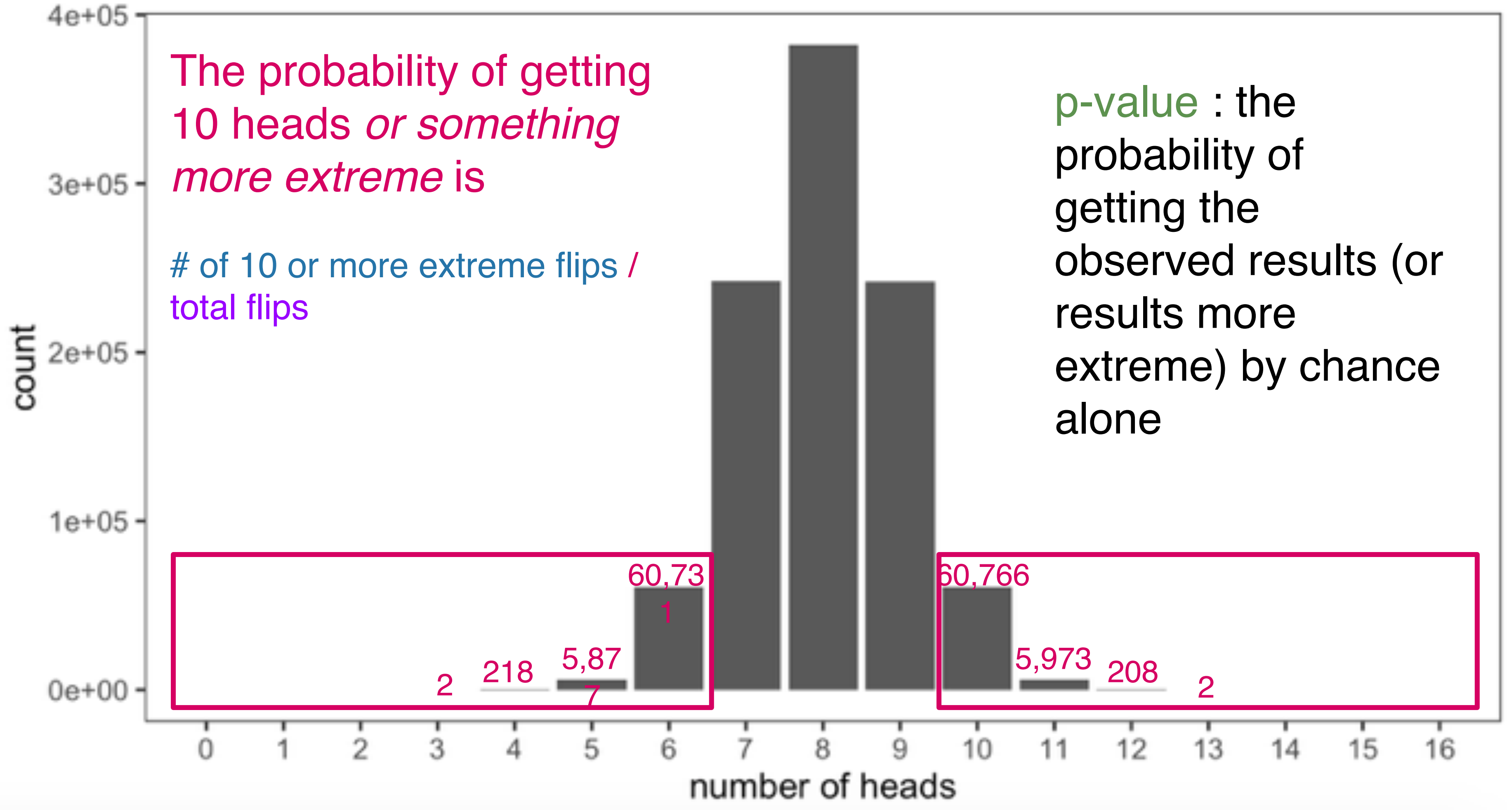
In your trial, you observe 10 heads.

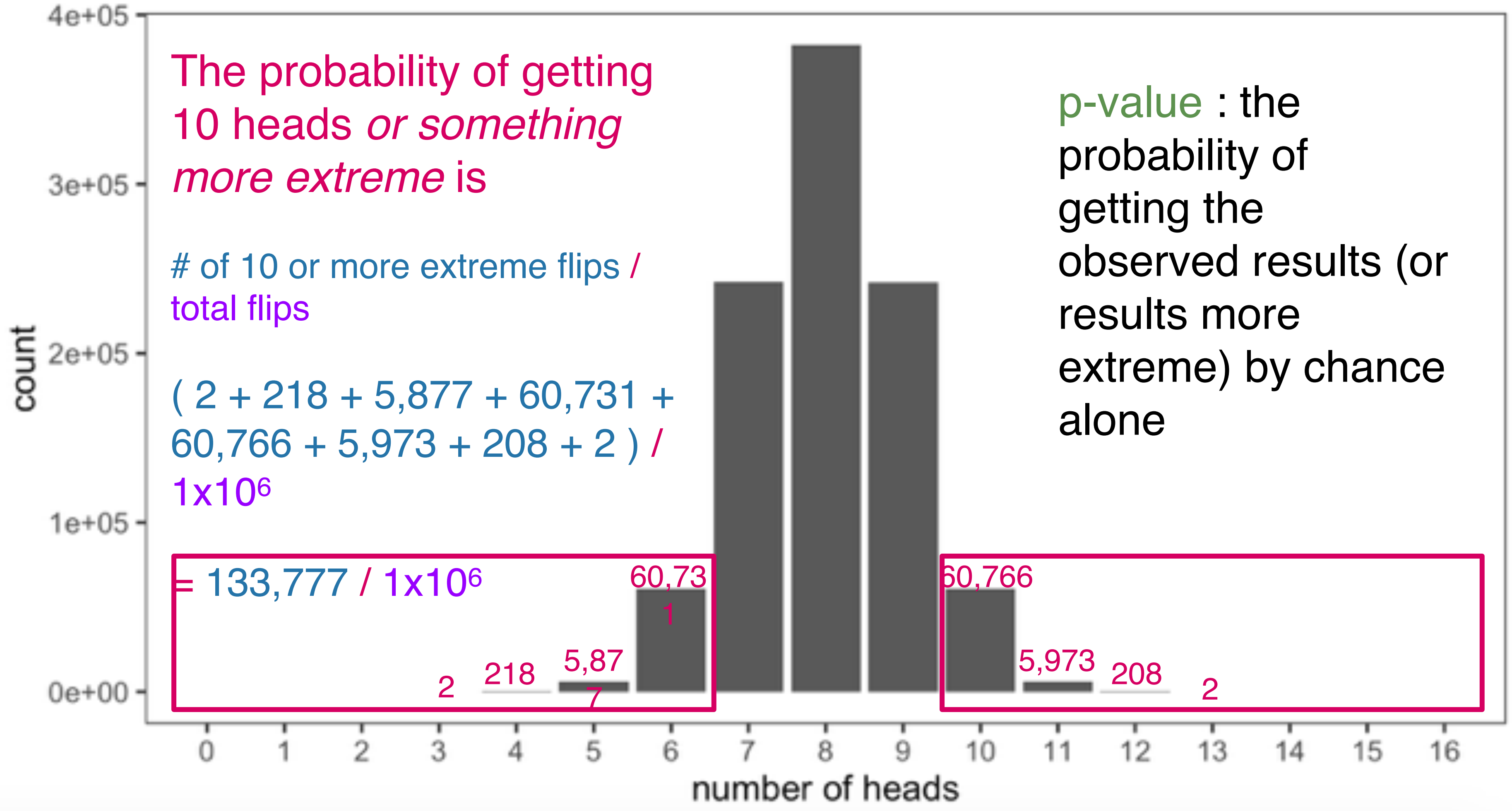
p-value : the probability of getting the observed results (or results more extreme) by chance alone

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




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
60,731
60,766
2
218
5,877
5,973
208
2
4e+05
3e+05
2e+05
1e+05
0e+00
count
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
2
218
5,877
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



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
60,731
5,877
218
2
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

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

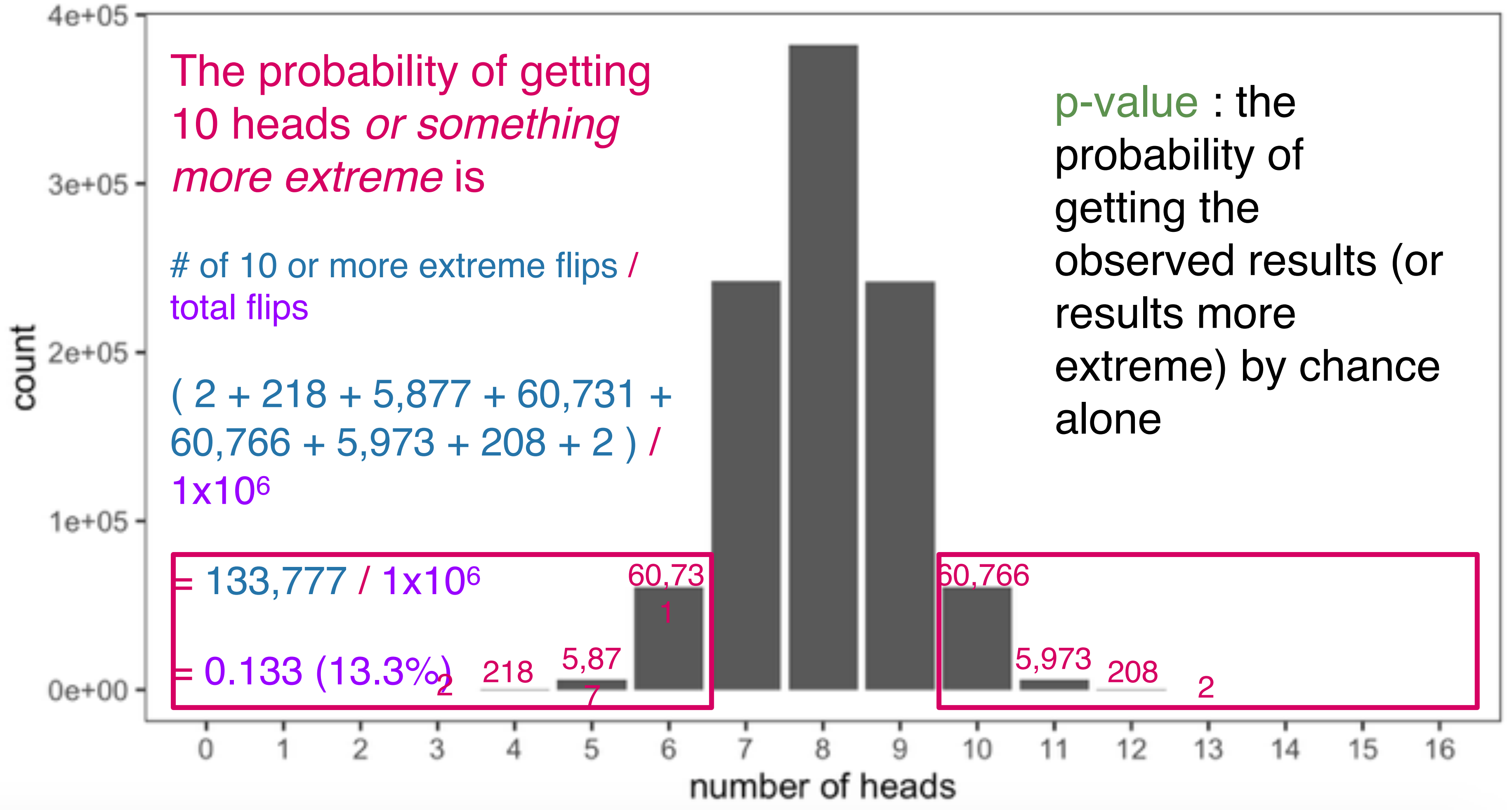
# of 10 or more extreme flips / total flips

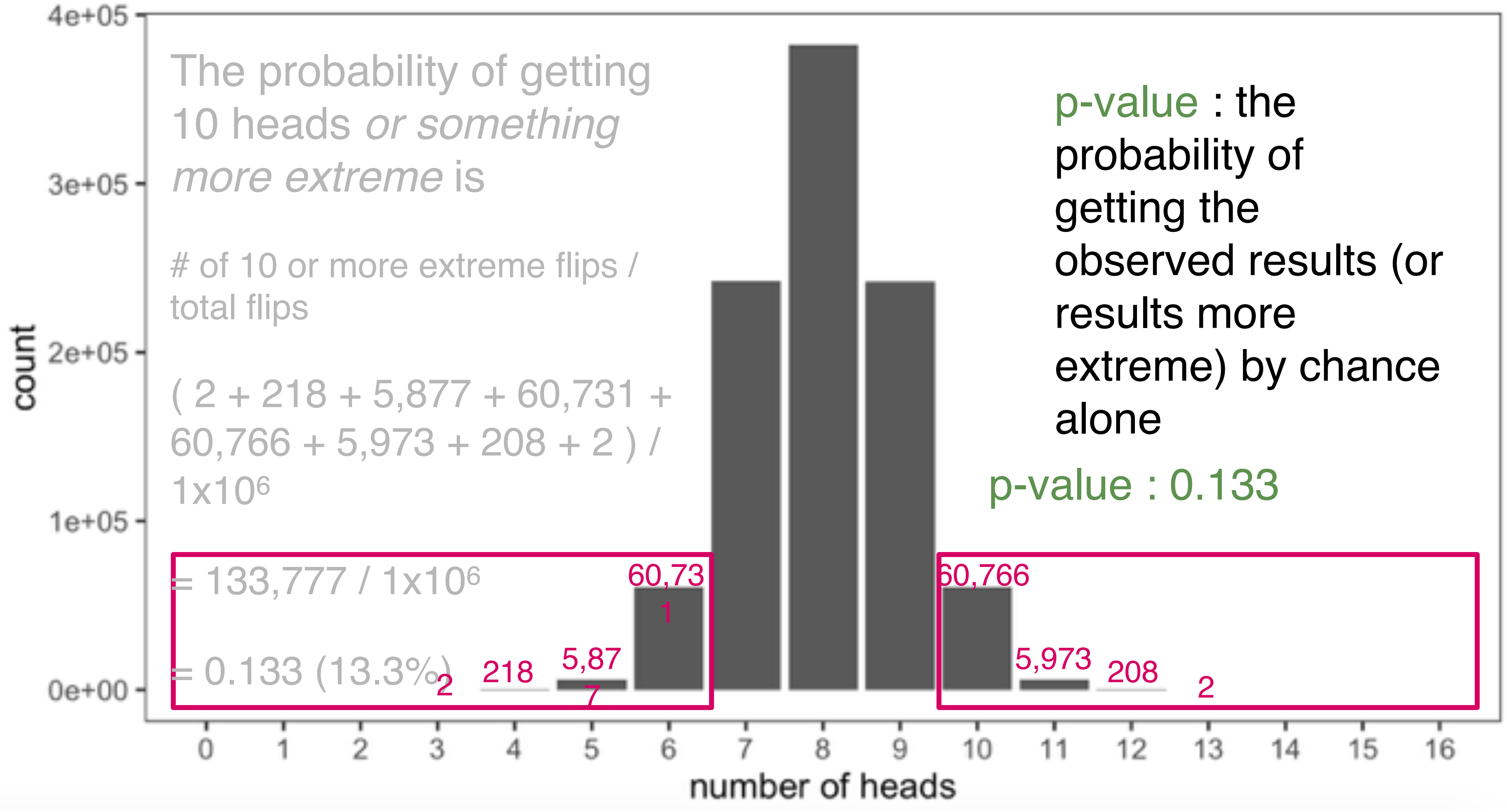
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1×10^6

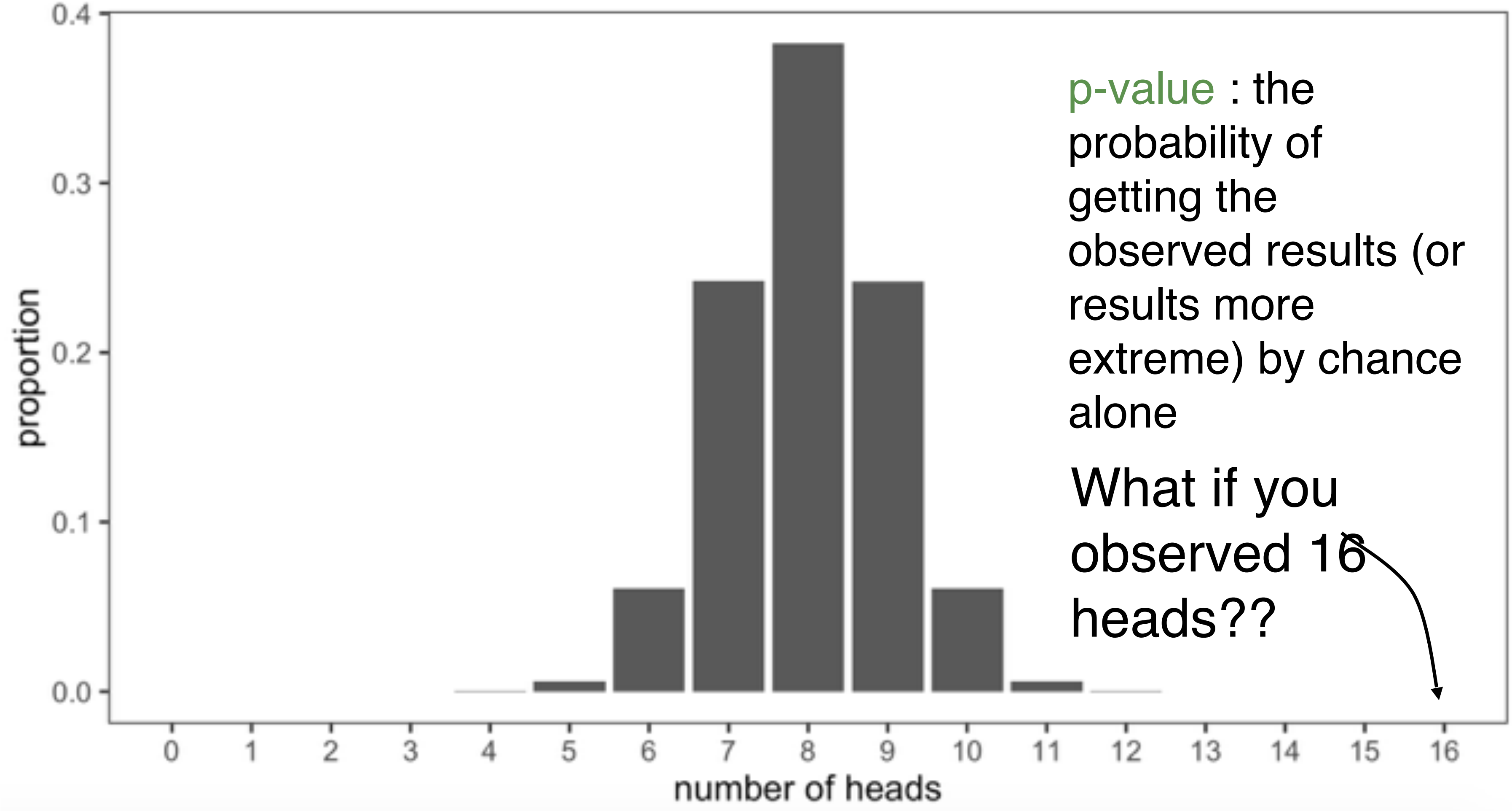
= 133,777 / 1×10^6

= 0.133 (13.3%)

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

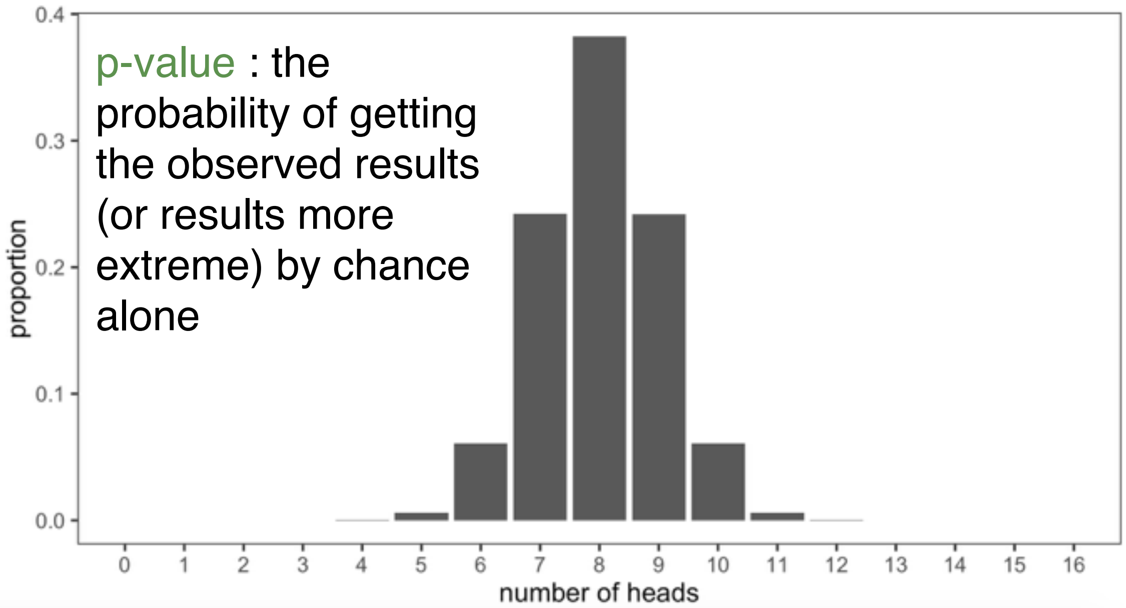



The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
(2 + 218 + 5,877 + 60,731 + 60,766 + 5,973 + 208 + 2) / 1x10^6
= 133,777 / 1x10^6
= 0.133 (13.3%)
p-value : the probability of getting the observed results (or results more extreme) by chance alone
p-value : 0.133
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



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

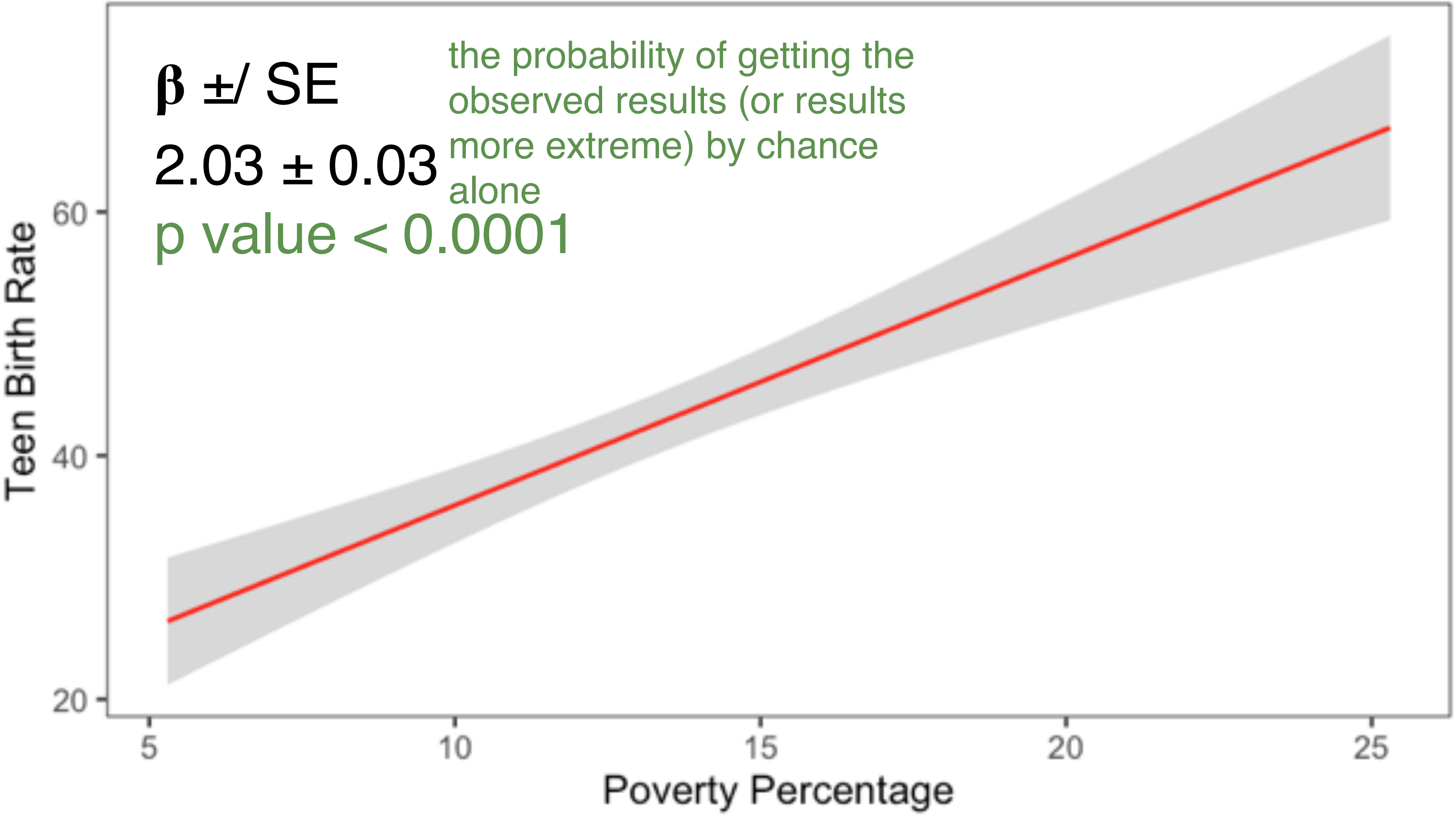
What if you observed 16 heads??



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

A
< 0.13

B
> 0.13

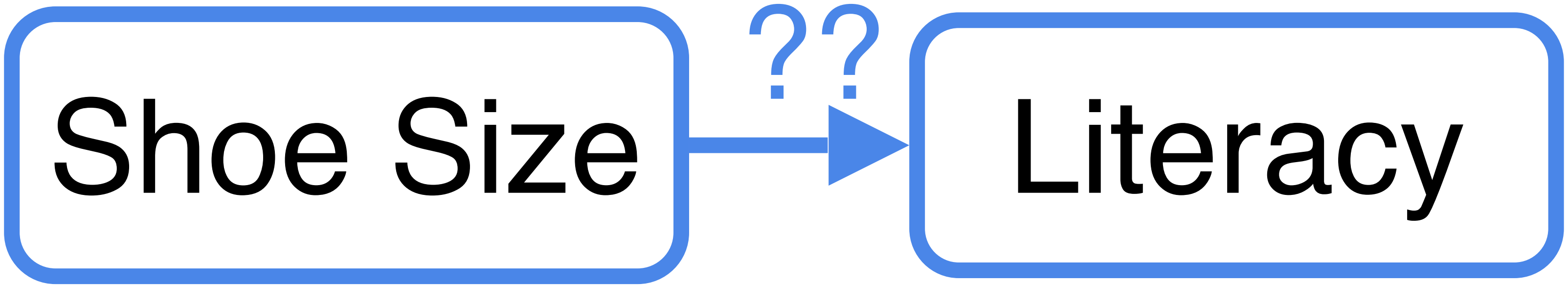

β ±/ SE
2.03 ± 0.03
the probability of getting the observed results (or results more extreme) by chance alone
p value < 0.0001
60
40
20
Teen Birth Rate
5
10
15
20
25
Poverty Percentage

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

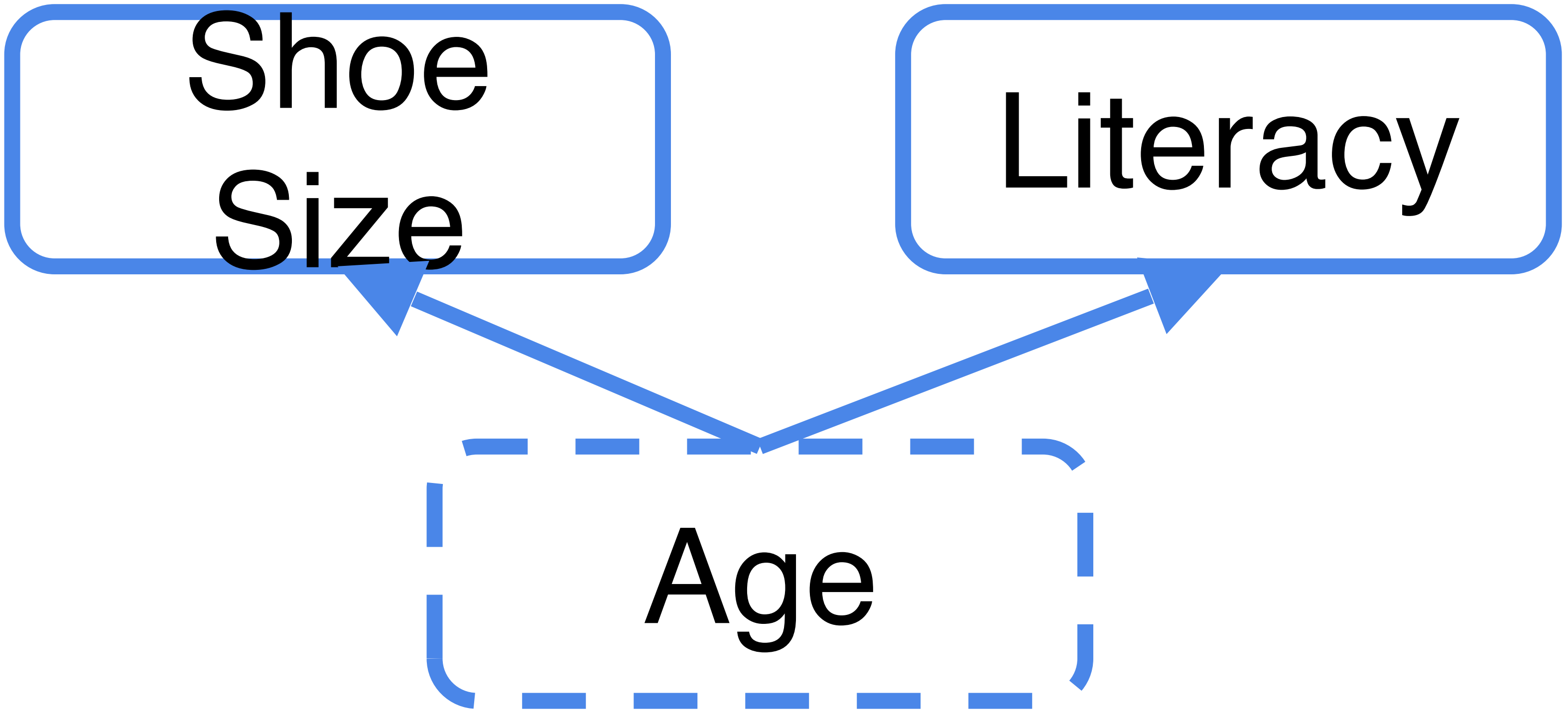
p-value : the probability of getting the observed results (or results more extreme) by chance alone

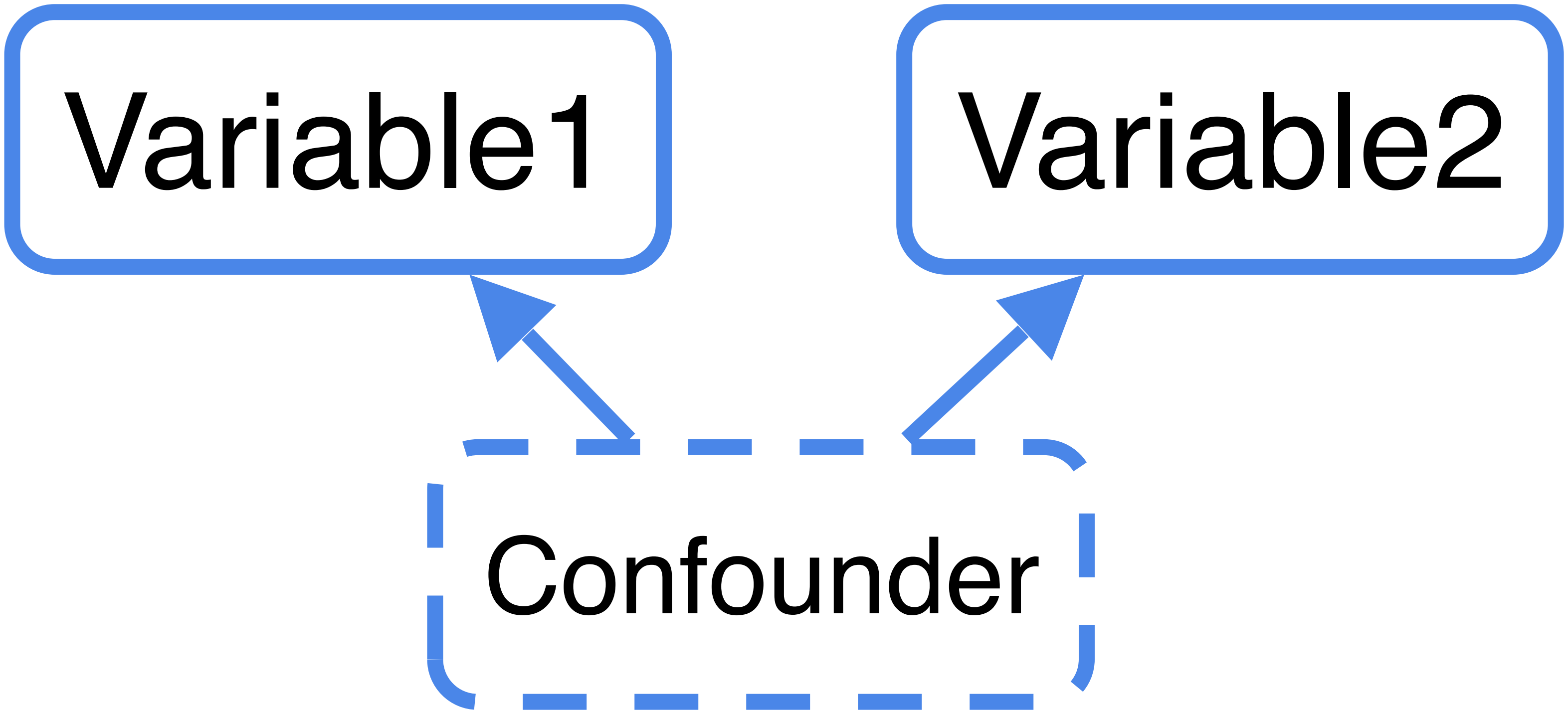
Confounding


Small shoes
Not literate
Big shoes
Literate


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?

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 the next 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 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

Confounding

What are possible confounders for our analysis of the effect of poverty on teen birth rate?