Reminders

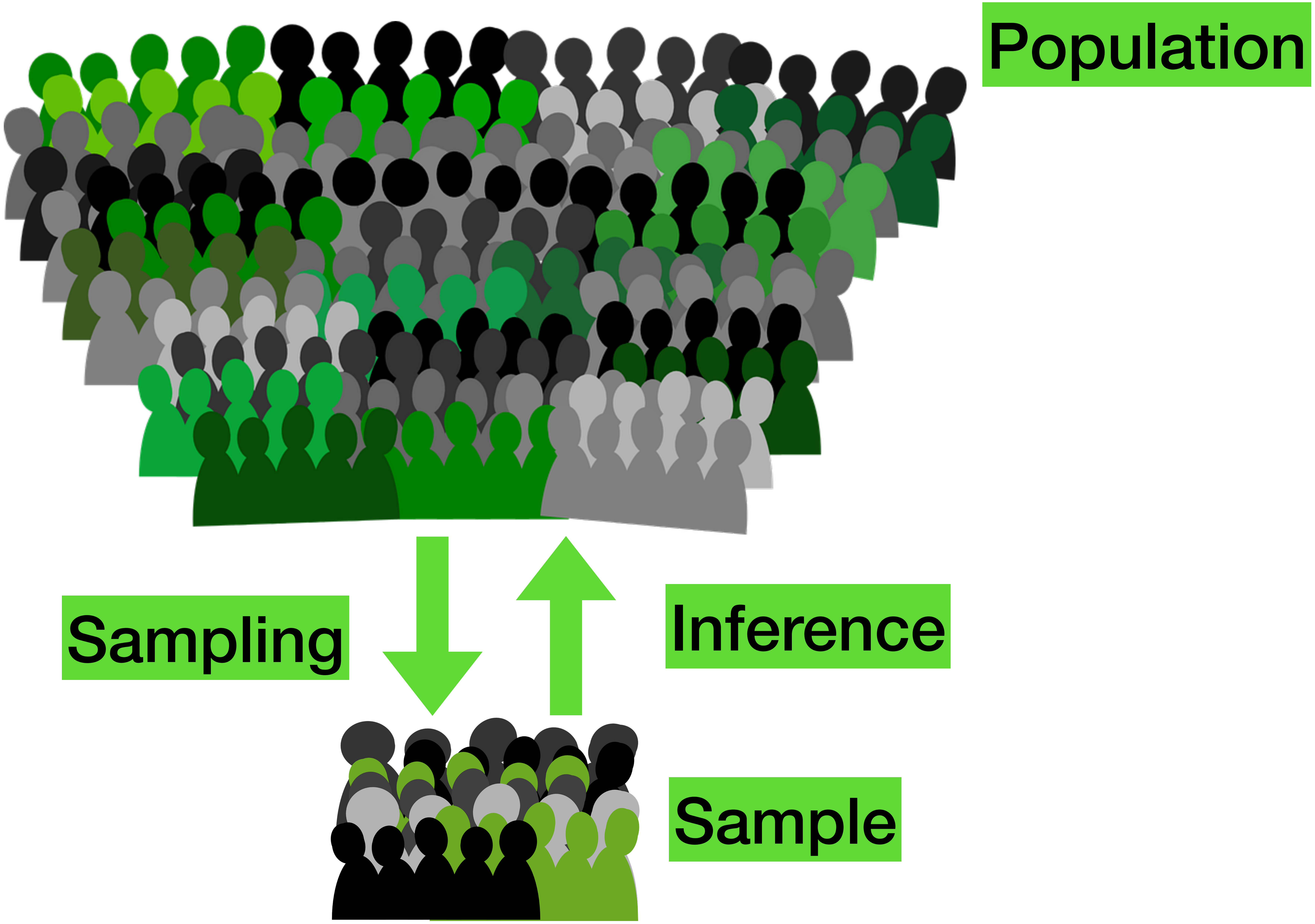
<u>Upcoming due dates</u>
Fri Oct 24th Discussion Lab 3
Mon Oct 27th Quiz 4

Repo invites: Click accept before it expires next week!

Statistical inference I

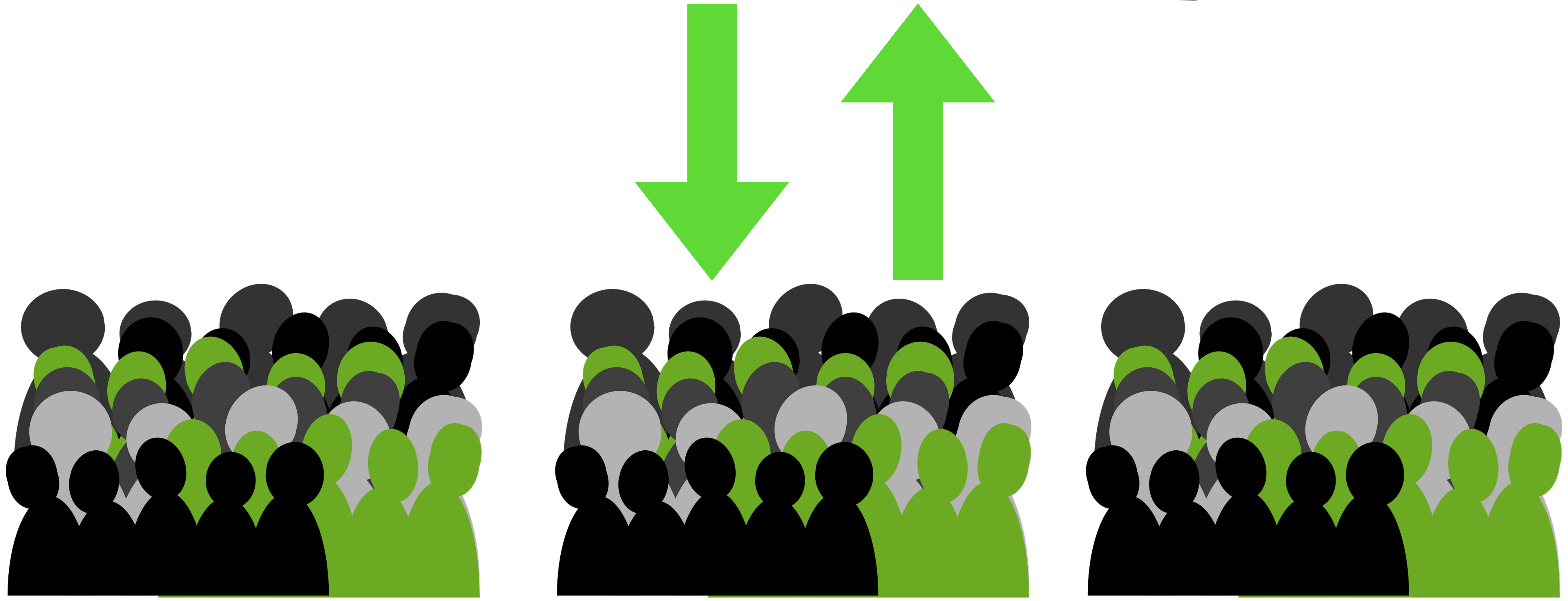
Data Science in Practice

Jason G. Fleischer, PhD **Dept. of Cognitive Science** **UC San Diego** **https://jgfleischer.com**



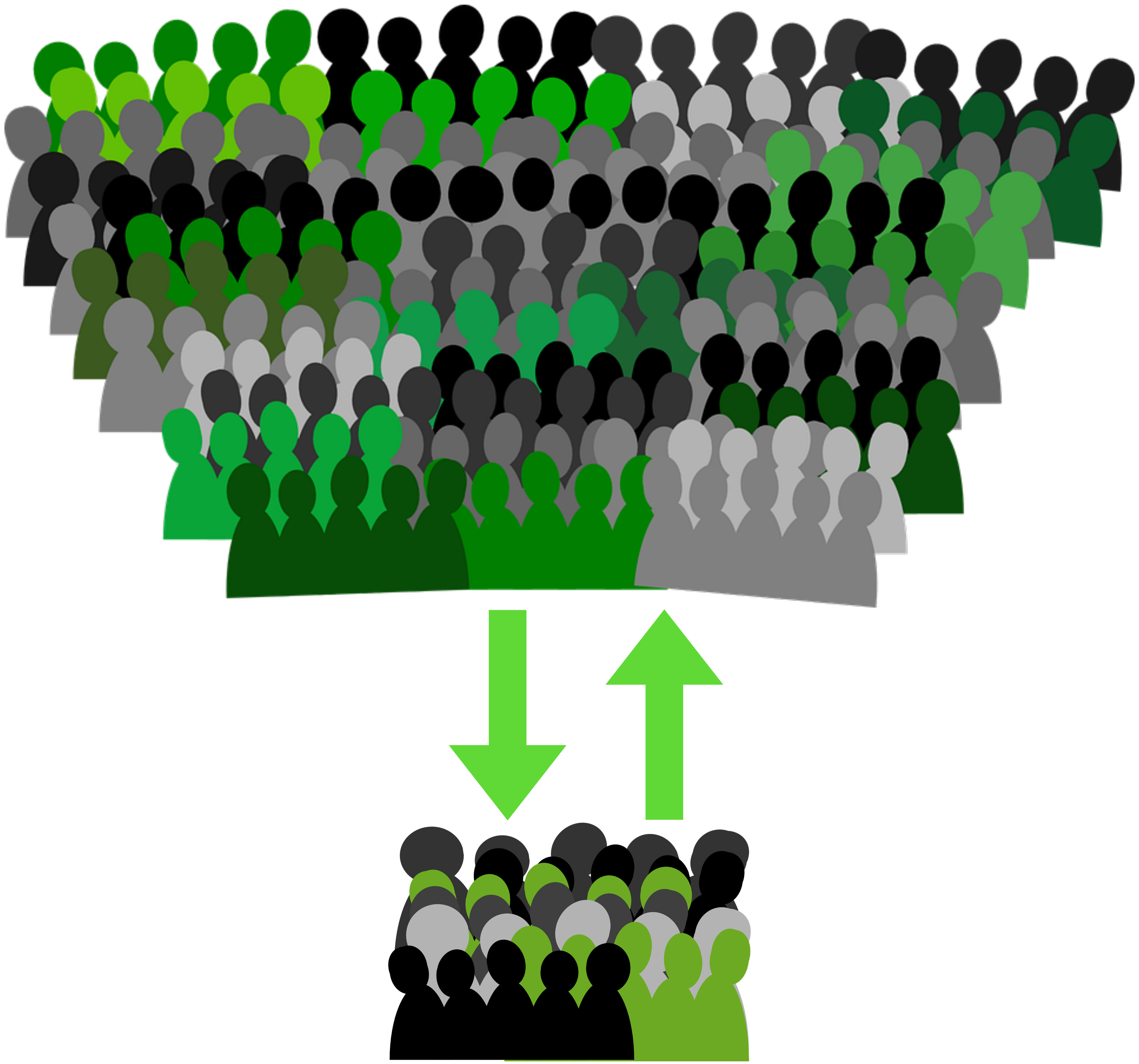
$$\bar{x} = 17.4$$

Random samples are random!
They differ from each other and the population!



$\bar{x}_1 = 17.4$ $\bar{x}_2 = 17.8$ $\bar{x}_3 = 17.1$

Confidence interval : a range of values calculated from a sample statistic, such that there is a specified probability that the value of the true value of the population (parameter) lies within it.



$$\bar{x} = 17.3 \pm 0.4$$

Statistical Inference

Using “frequentist” tools

- You have
  - data
  - a model (usually a null hypothesis)

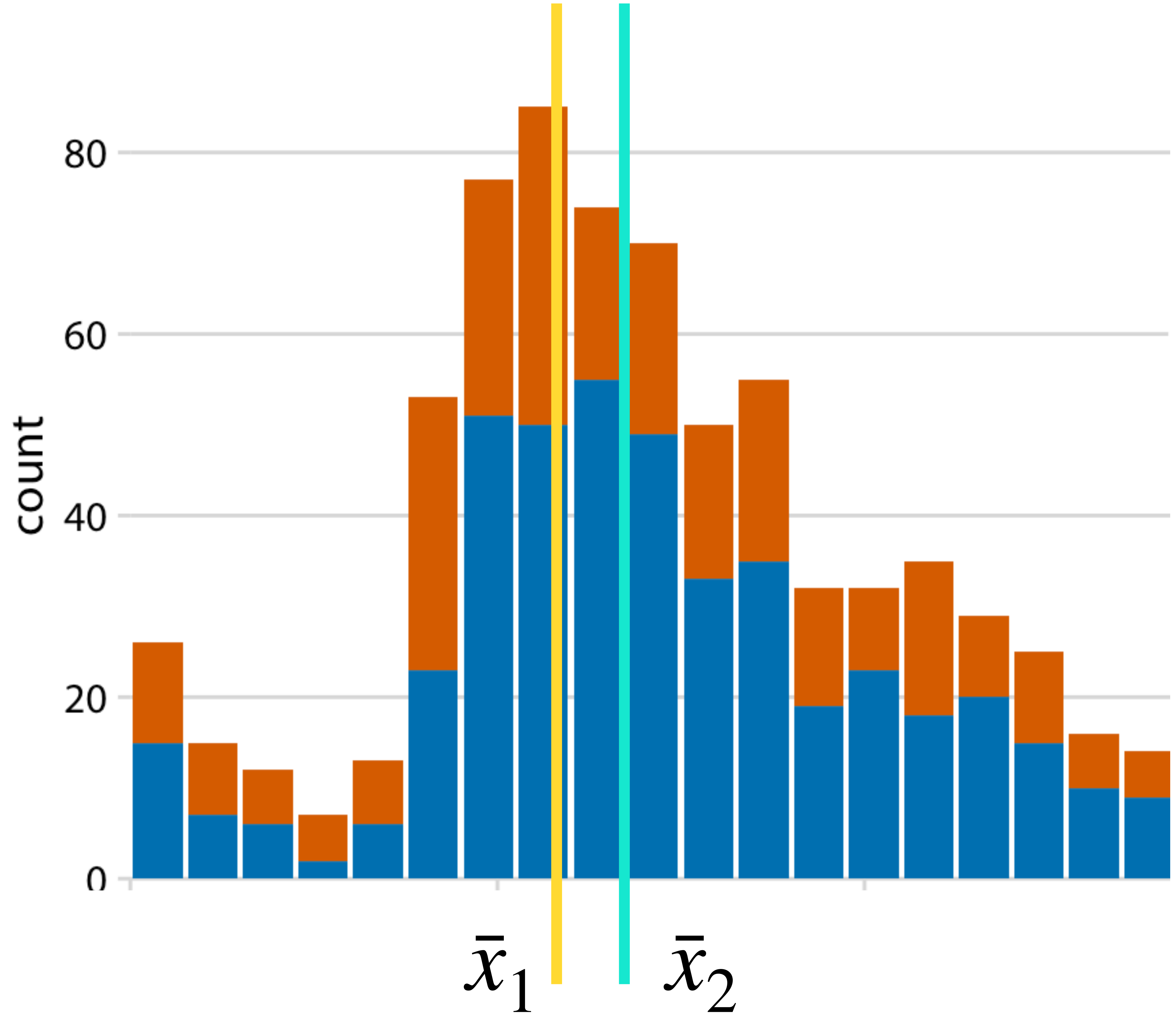
Null hypothesis: the assumption that there is no effect of the variable of interest. For example is no difference between control and treatment groups; or there is no relationship between variables x and y.

- You calculate the probability of observing the data given the null hypothesis is true: p-value”
- NOTE: this is p(D|H) not p(H|D)

Student’s t-test

Basic example of statistical inference

- Is there a real difference in mean value between two groups in the data?



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 Phillpotts). 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 Roaf 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 brewer's 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

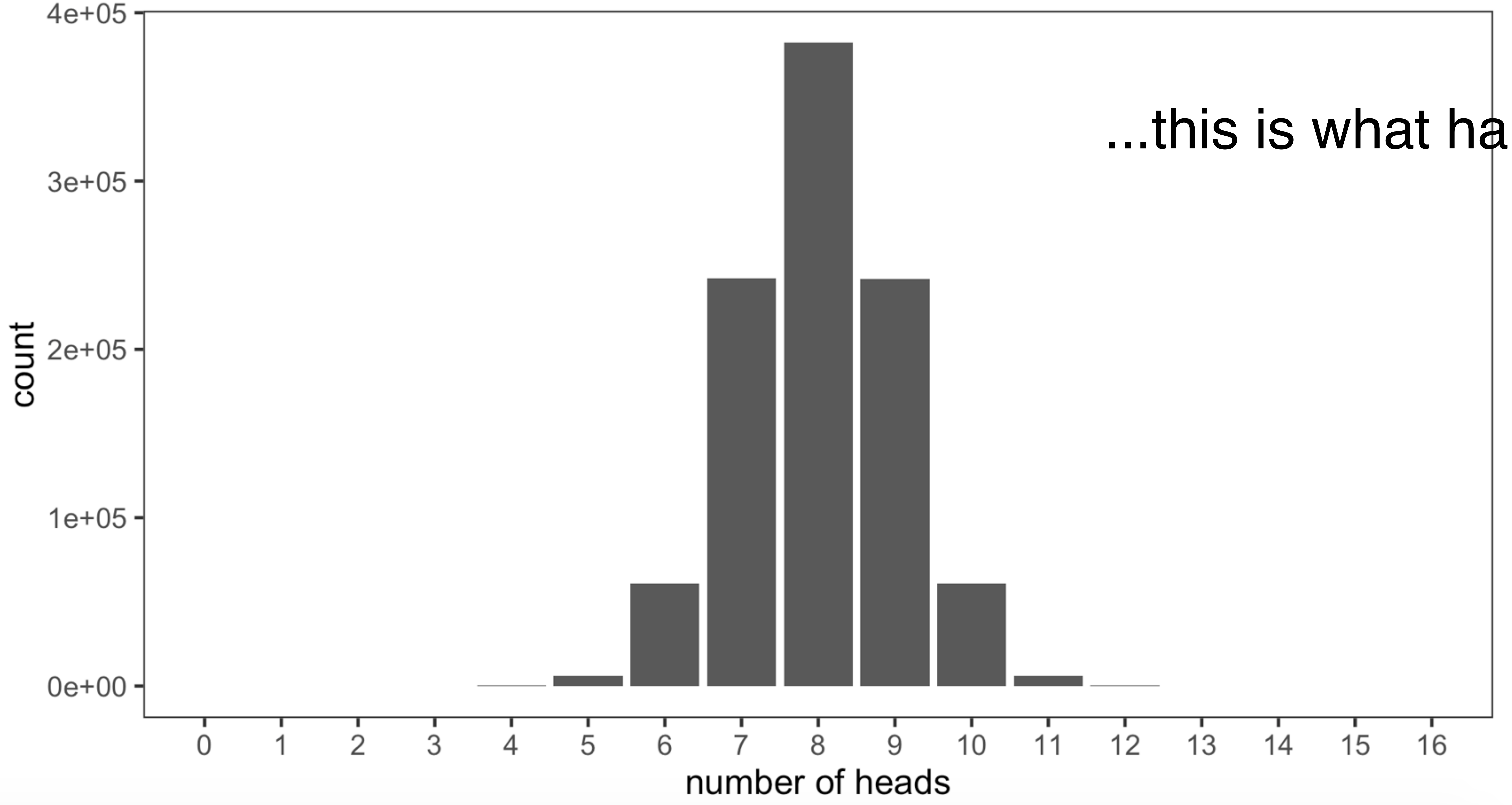
p-value : the probability under the null hypothesis of getting measurements as extreme as the observed results by chance alone

THIS IS NOT TYPE 1 ERROR RATE

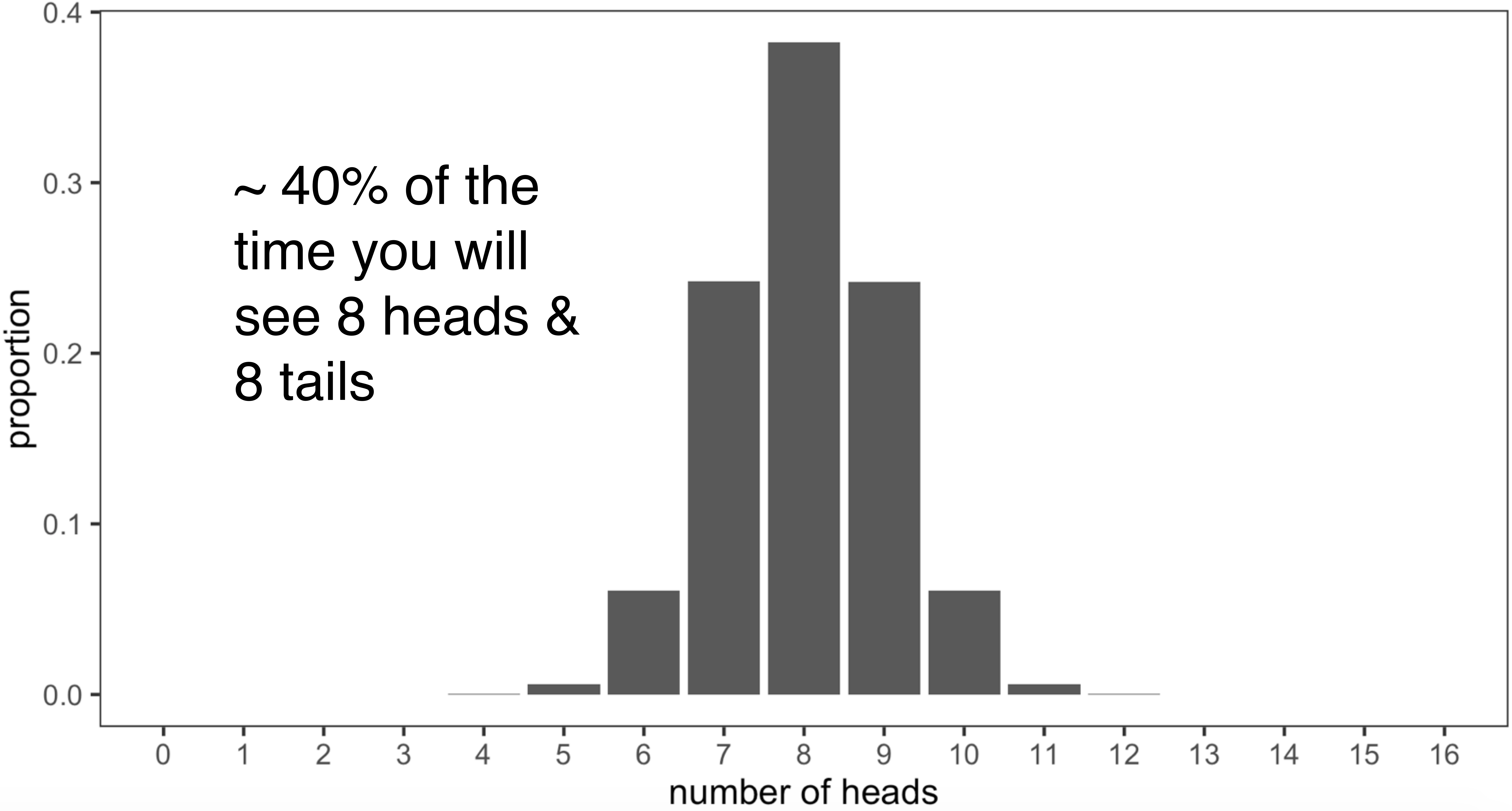
https://forms.gle/6MCyp7qFsaHgGKi5A

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

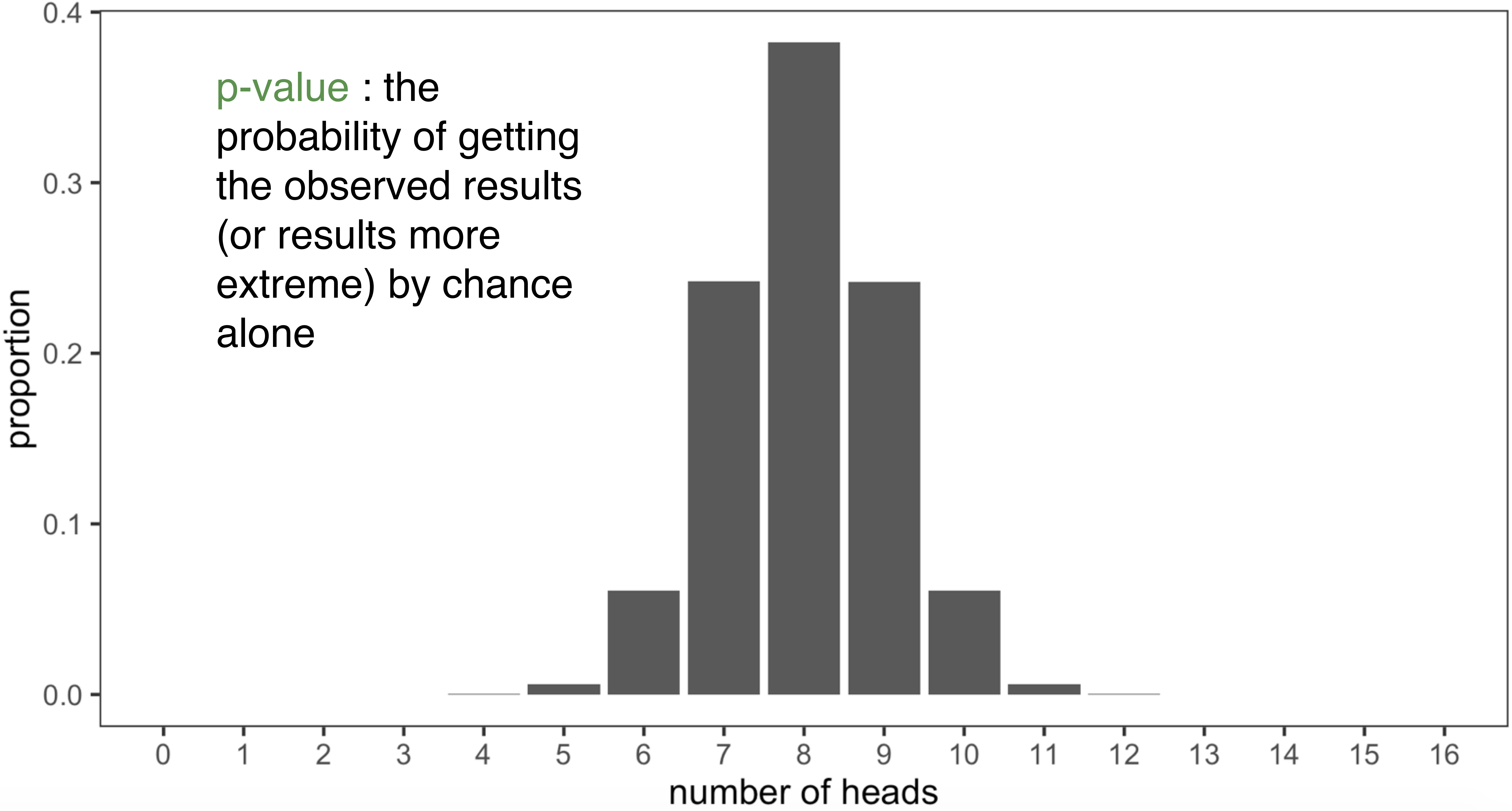
….and then do that 1M times

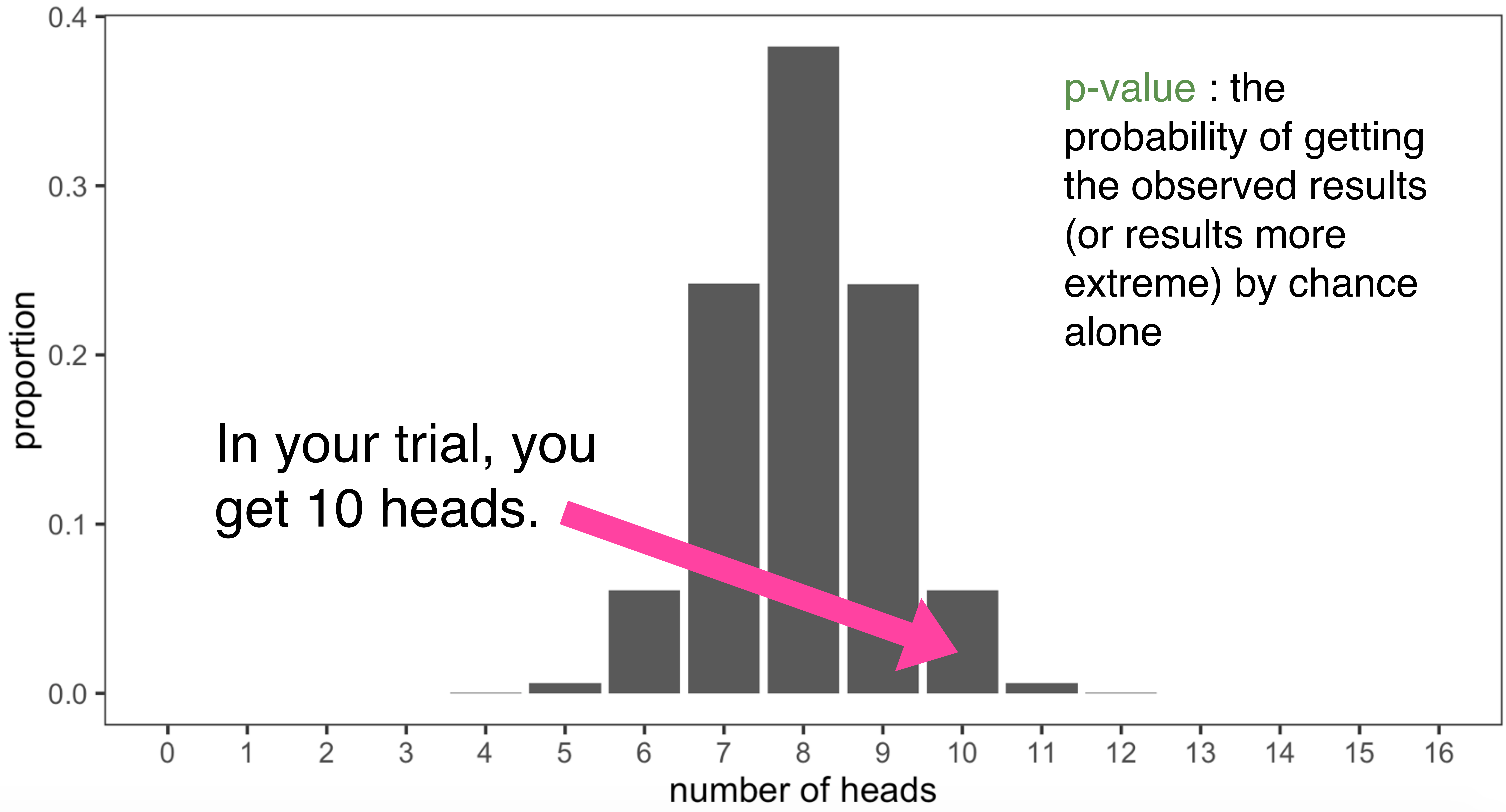


...this is what happens by chance alone.

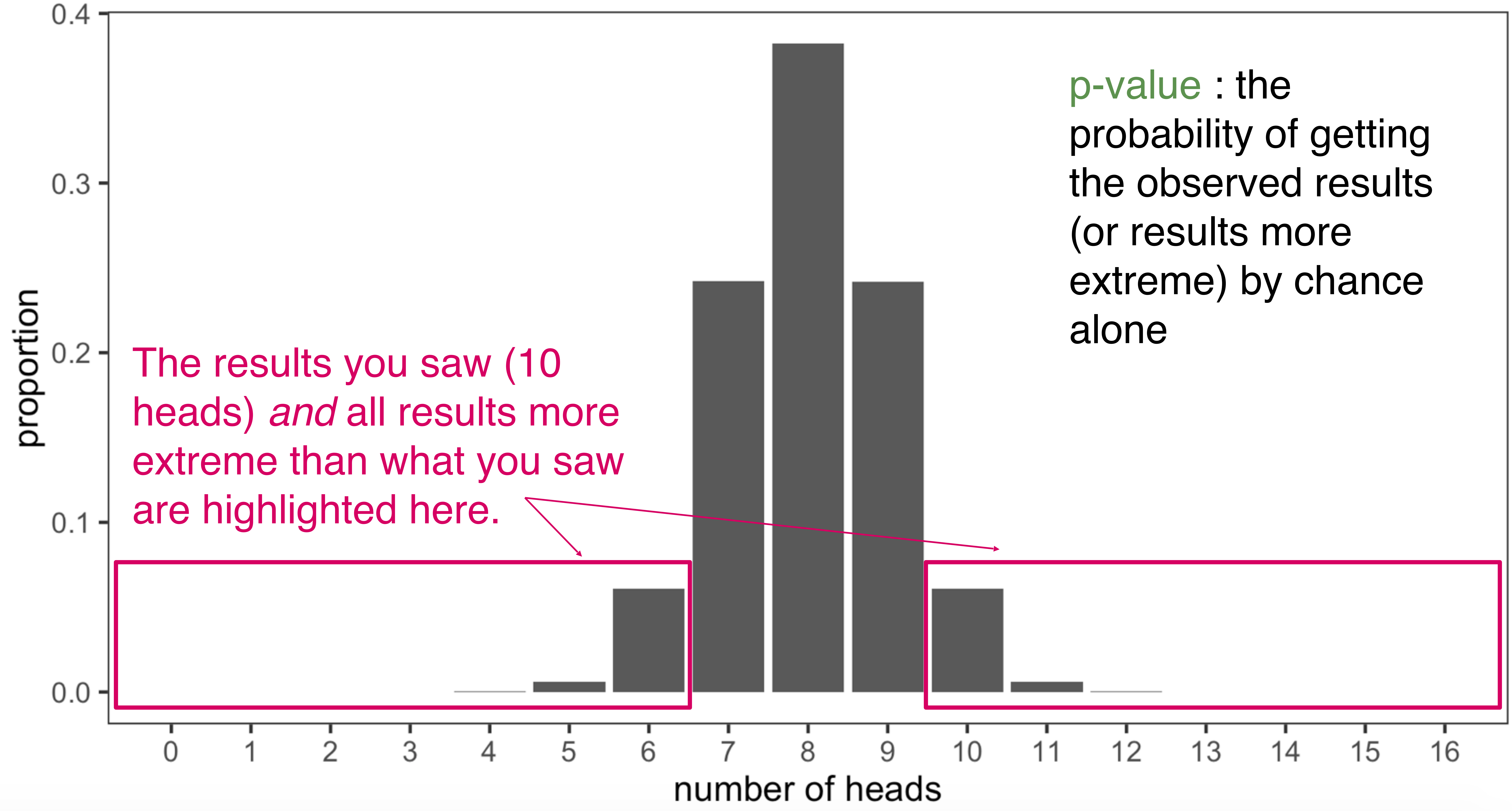


~ 40% of the time you will see 8 heads & 8 tails
proportion
0.0
0.1
0.2
0.3
0.4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

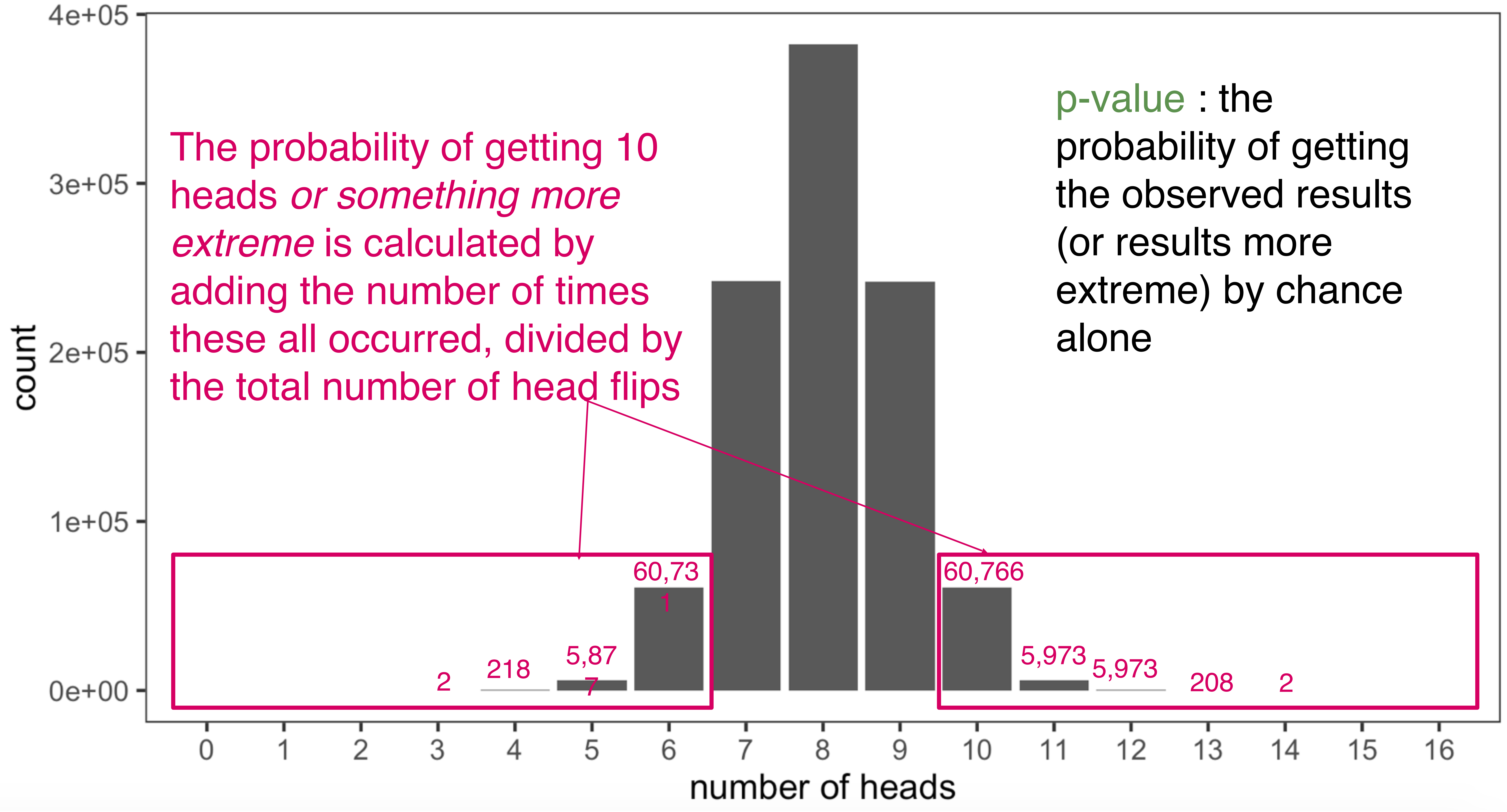

p-value : the probability of getting the observed results (or results more extreme) by chance alone
proportion
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
number of heads



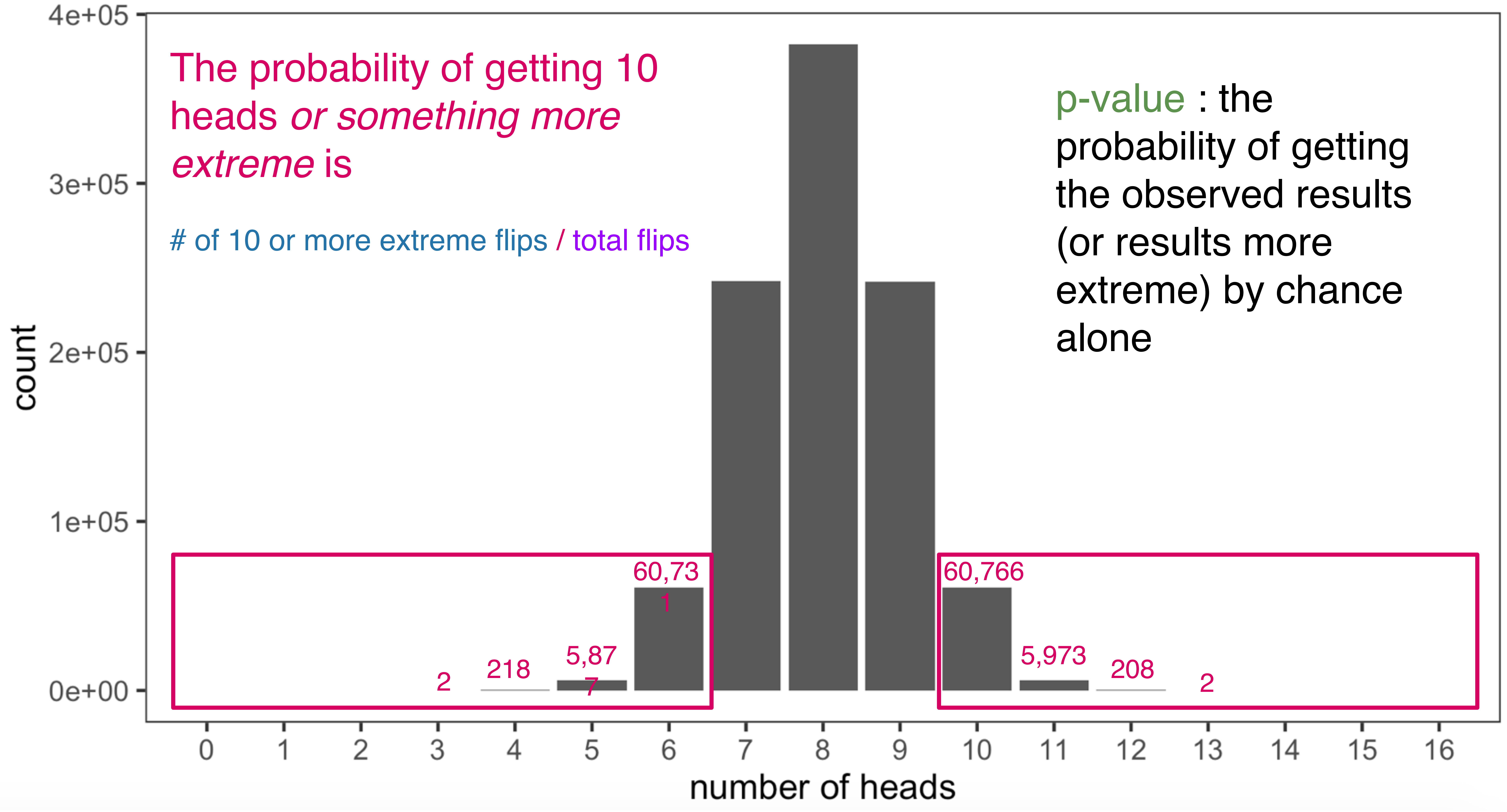
p-value : the probability of getting the observed results (or results more extreme) by chance alone

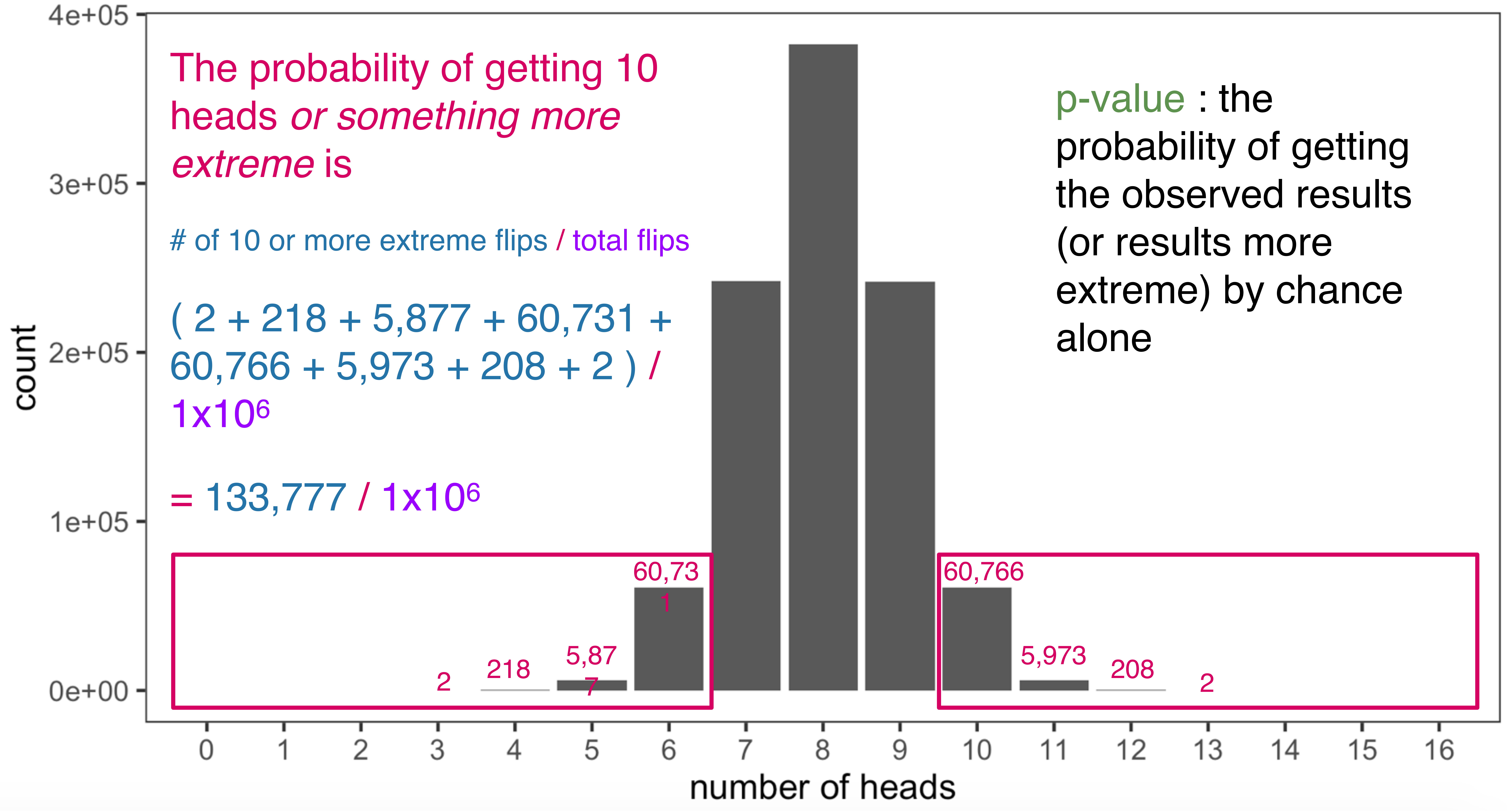


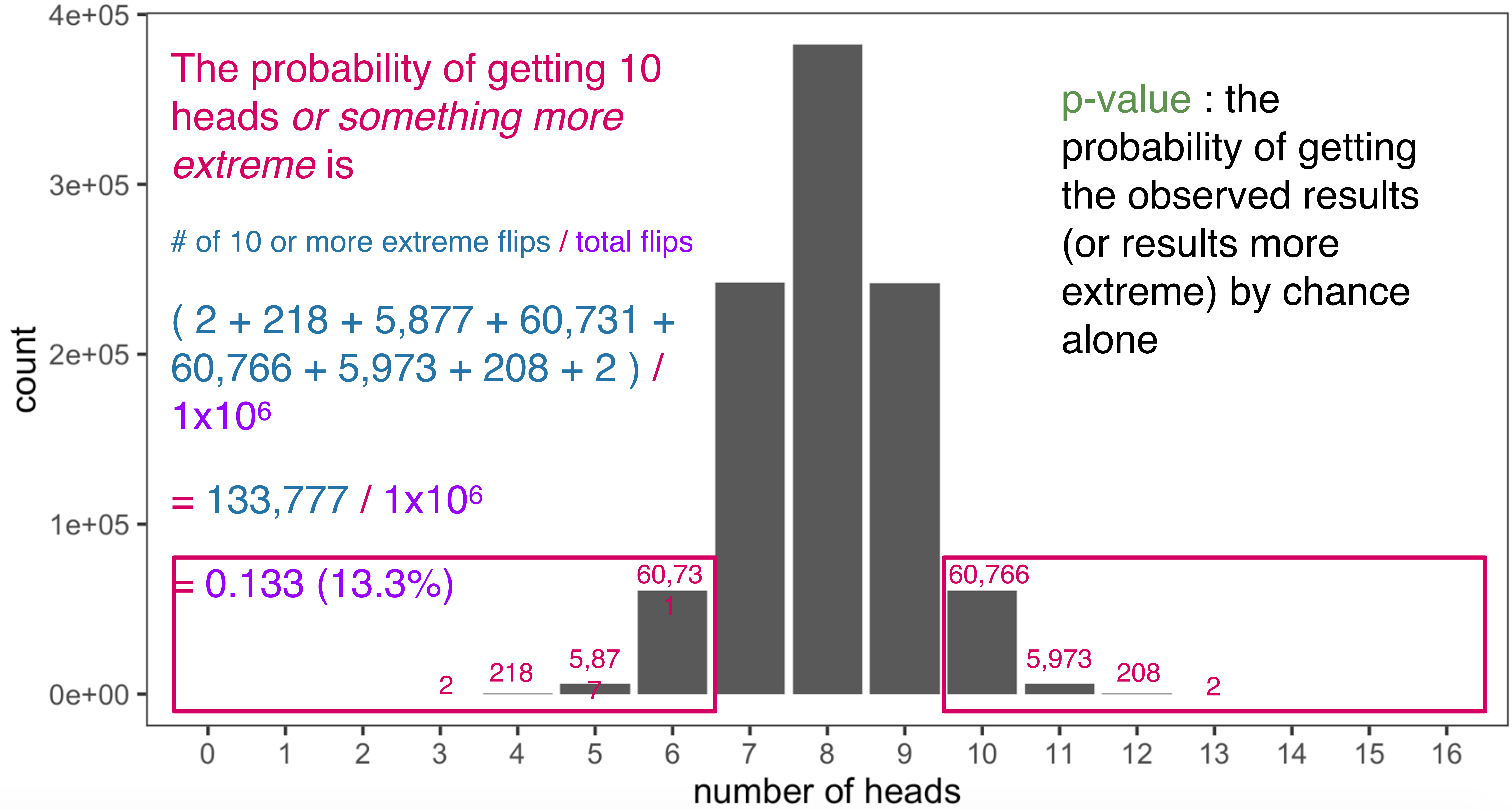
p-value : the probability of getting the observed results (or results more extreme) by chance alone
The results you saw (10 heads) *and* all results more extreme than what you saw are highlighted here.
proportion
0.0
0.1
0.2
0.3
0.4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

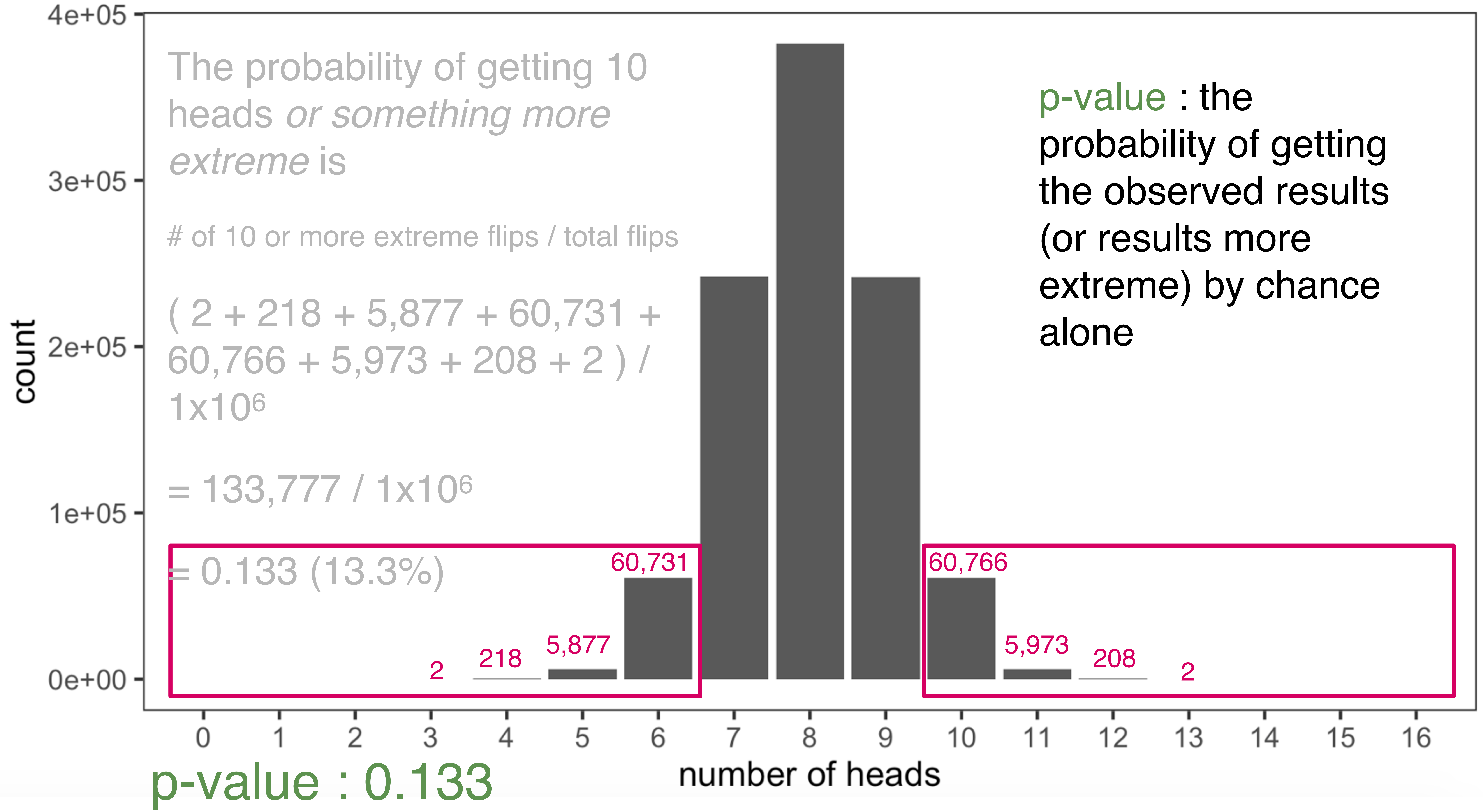


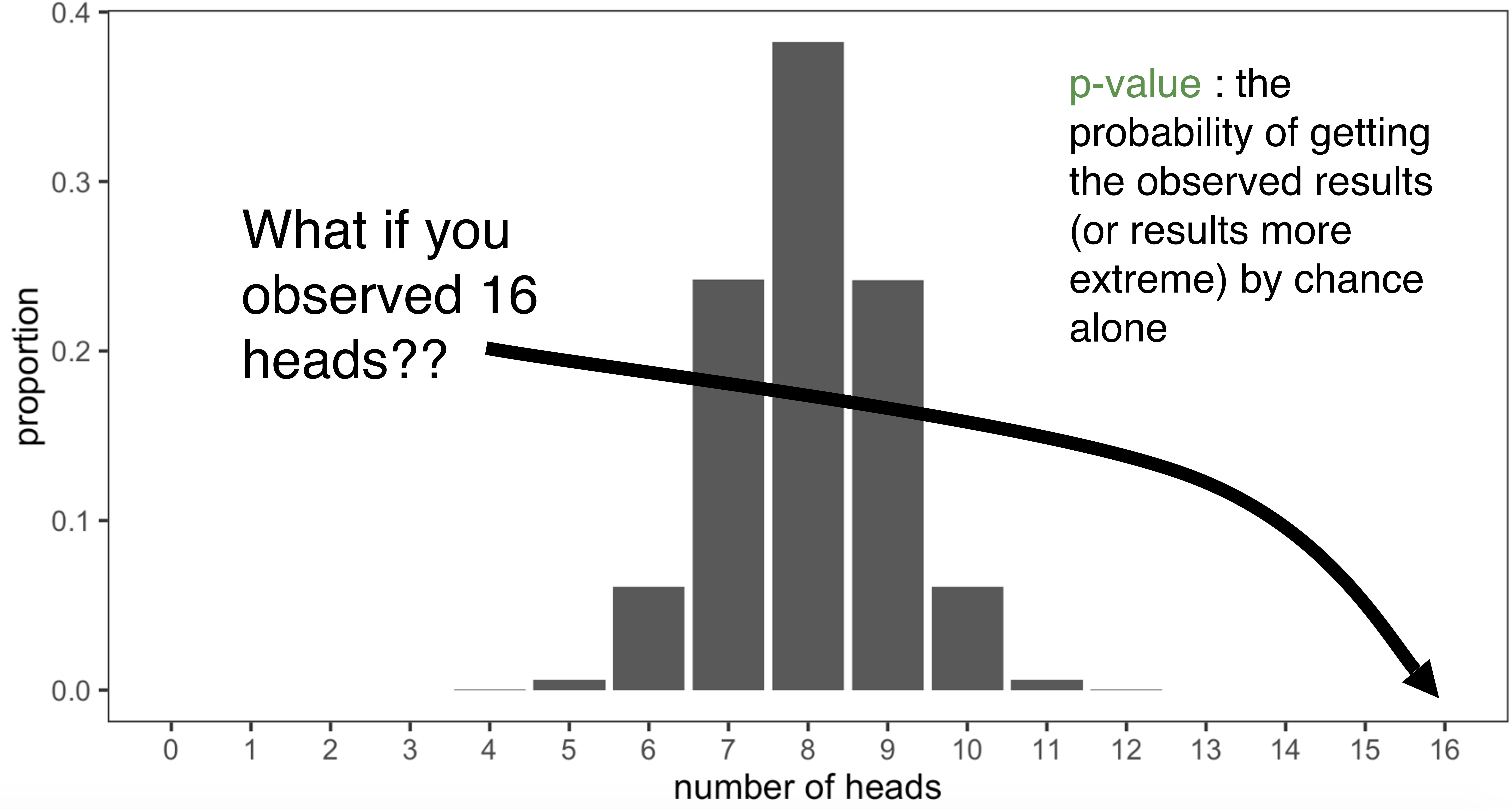
4e+05
3e+05
2e+05
1e+05
0e+00
count
The probability of getting 10 heads *or something more extreme* is calculated by adding the number of times these all occurred, divided by the total number of head flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone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

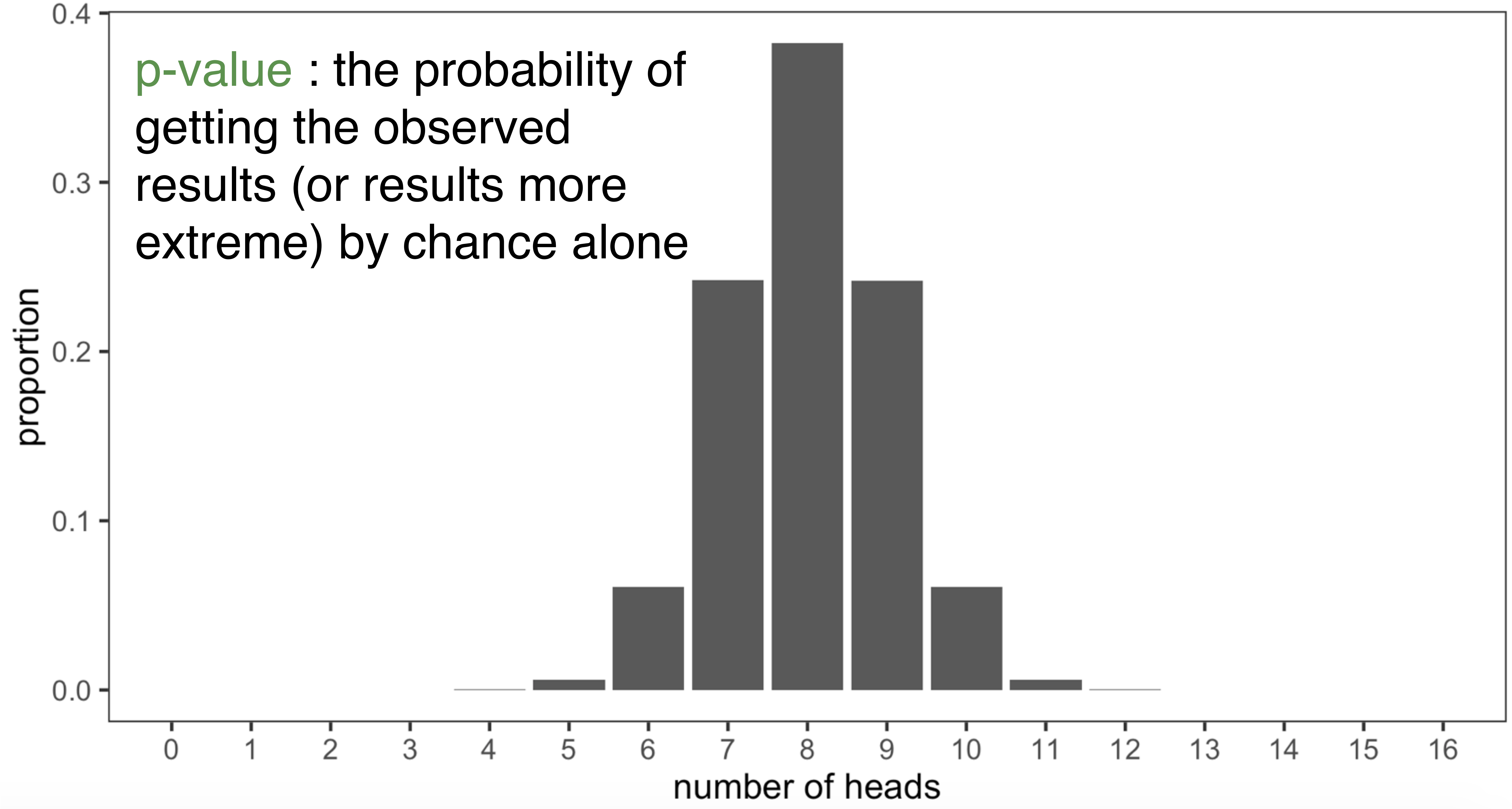

The probability of getting 10 heads *or something more extreme* is
# of 10 or more extreme flips / total flips
p-value : the probability of getting the observed results (or results more extreme) by chance alone
4e+05
3e+05
2e+05
1e+05
0e+00
count
2
218
5,87
7
60,73
1
60,766
5,973
208
2
0 1 2 3 4 5 6 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
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
5,877
60,731
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
2
218
5,877
60,731
60,766
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
(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
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
p-value : 0.133


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
What if you observed 16 heads??
p-value : the probability of getting the observed results (or results more extreme) by chance alone



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

A
< 0.13

B
> 0.13

This is how all statistical inference works

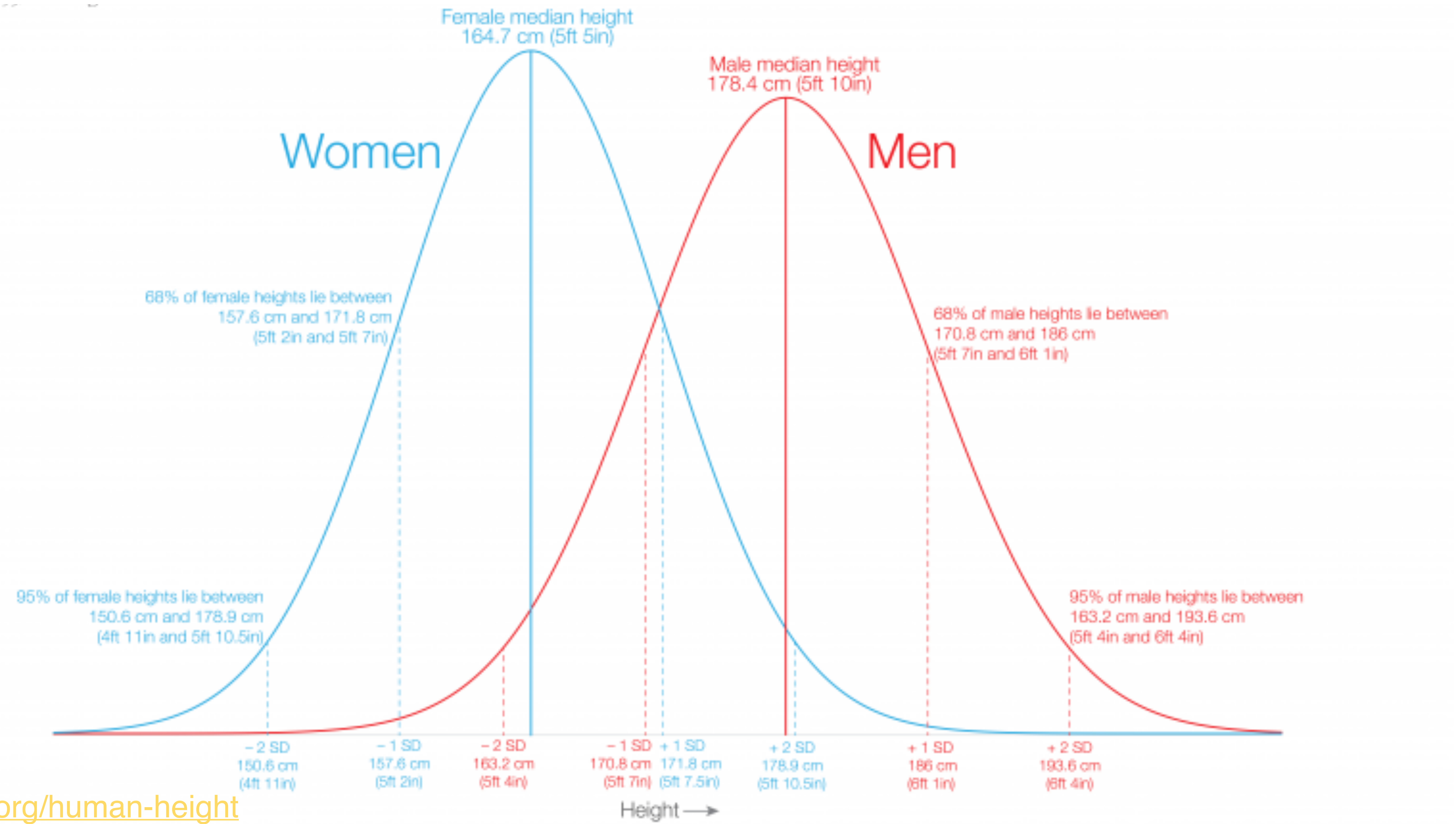
Including Student's t test

- A test defines a statistic with a known distribution of probabilities under a null hypothesis H_0
    - Remember, null is NO DIFFERENCE.
    - Different kinds of tests define different forms of…
        - the null hypothesis
        - the test statistic and its distributional form
- For a given set of data, calculate the probability to get results this weird or weirder assuming H_0 is true

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)

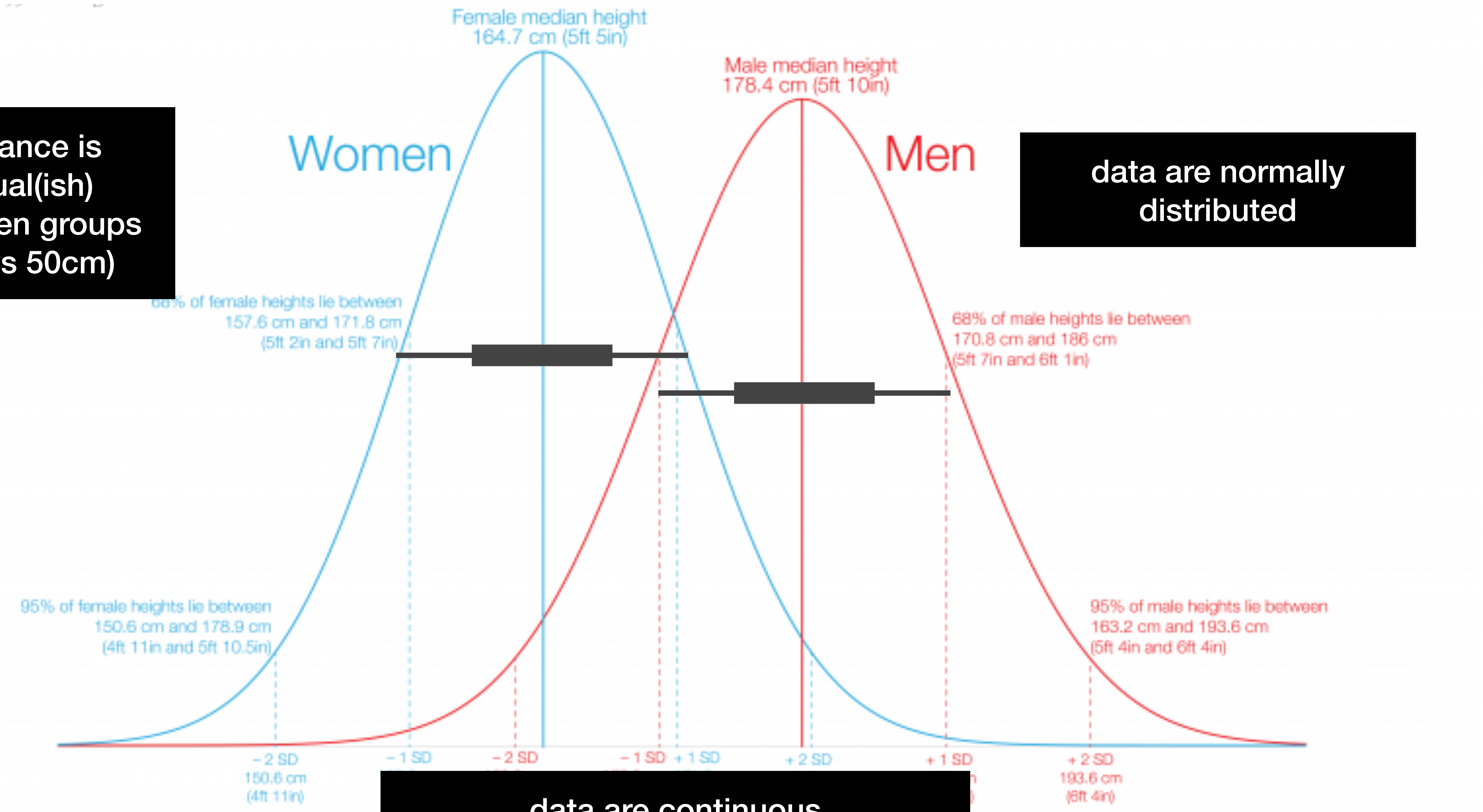
Do the heights between males and females differ?



N=10,000

https://ourworldindata.org/human-height

Do the heights between males and females differ?



variance is equal(ish) between groups (57 vs 50cm)

data are normally distributed

sample size affects statistic

N=10,000

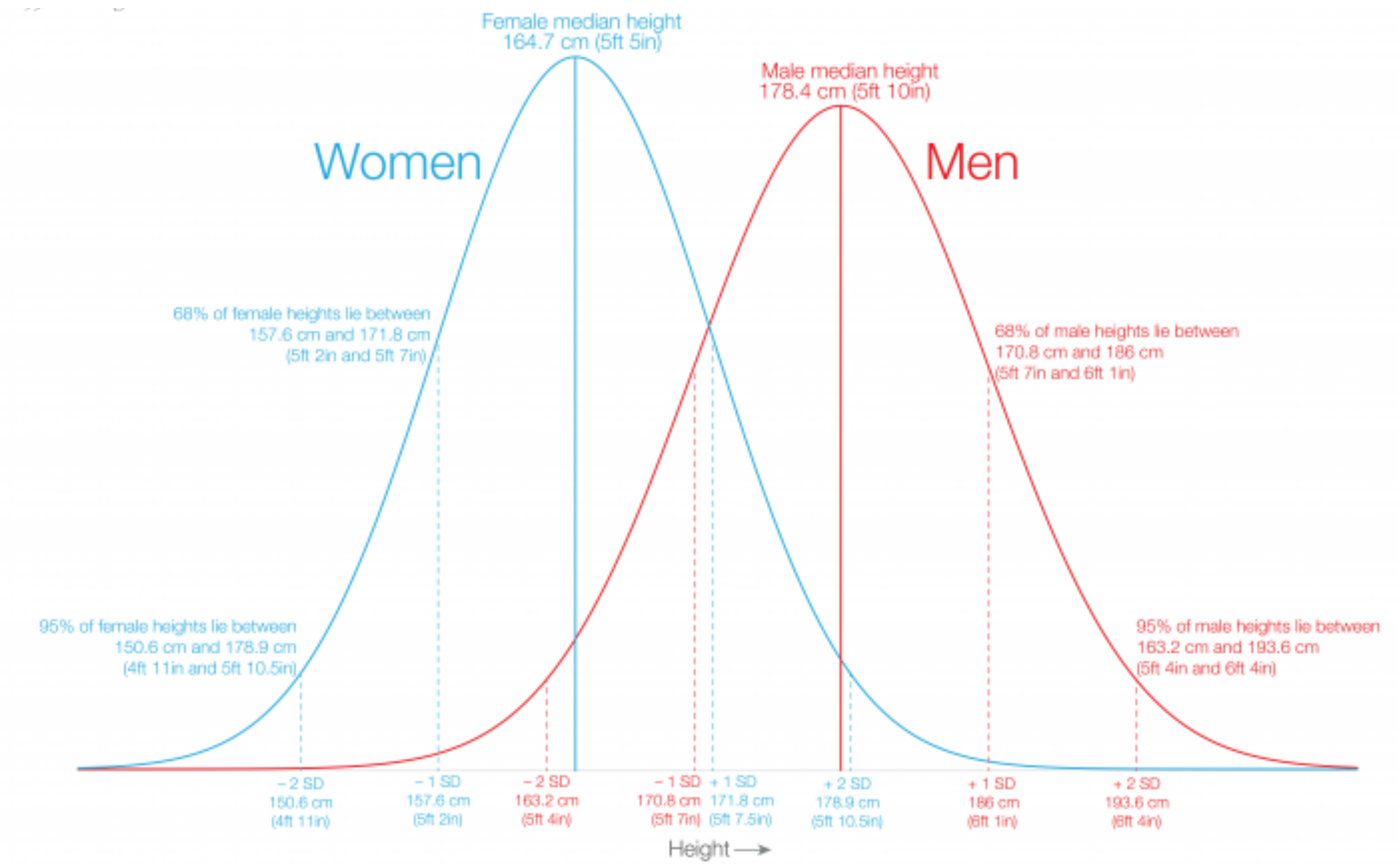
data are continuous

Do the heights between males and females differ?

Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample



t-statistic: -95.6

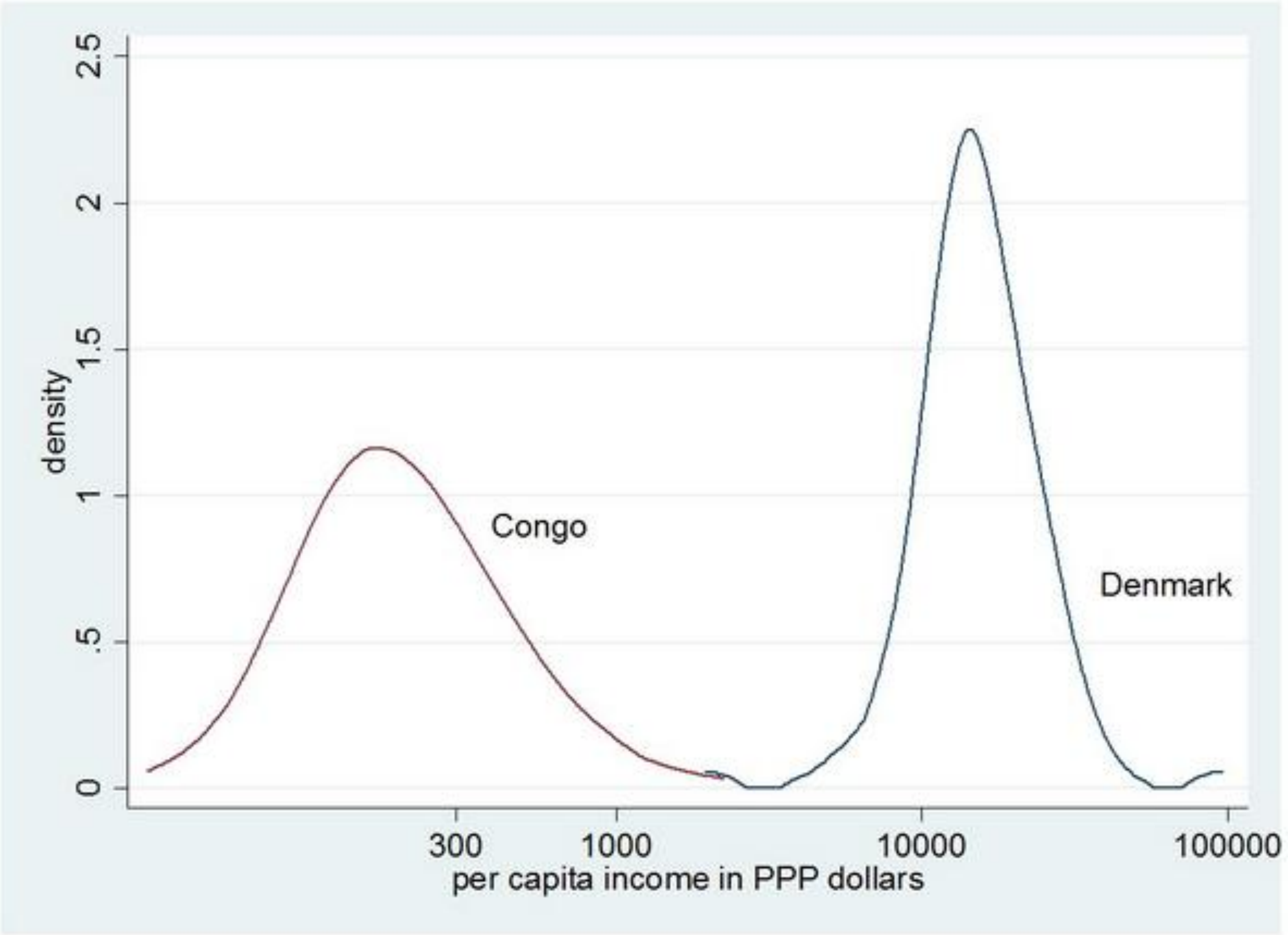
p-value << 0.001

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

```
import scipy.stats as stats

p, t = stats.ttest_ind(data1, data2)
```

Yes.



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

A
t-test not appropriate

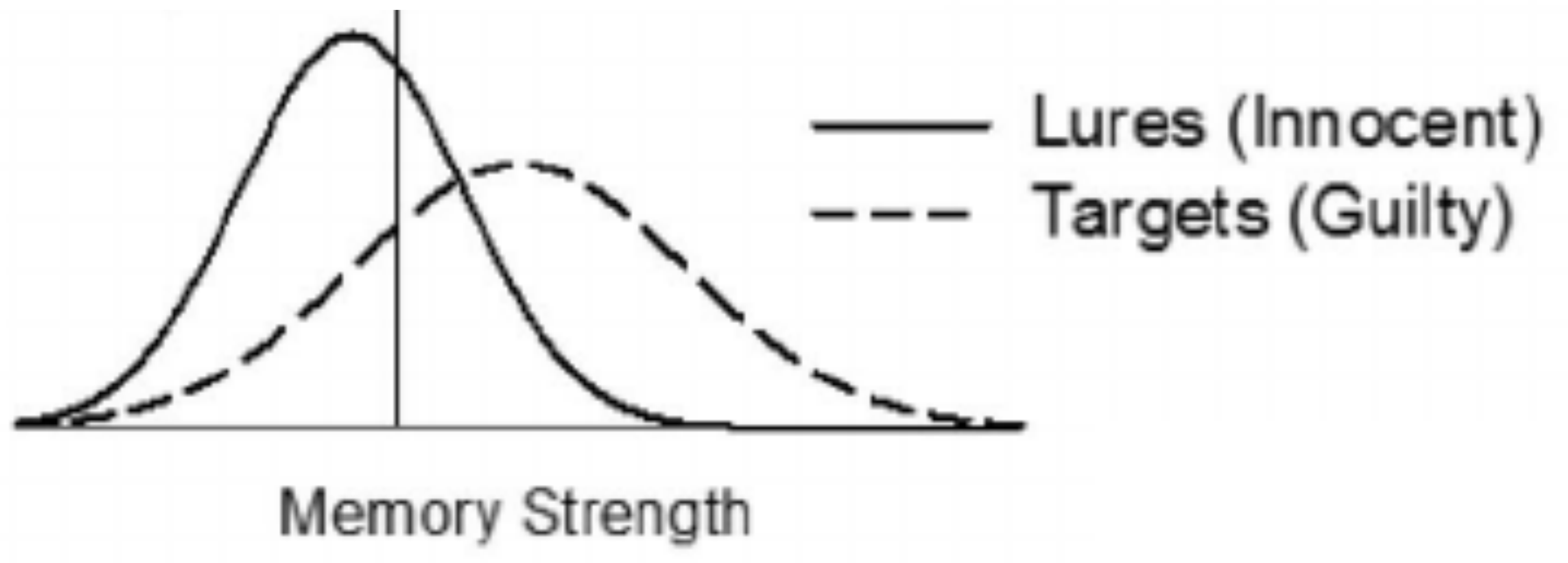
B
Yes

C
No

D
Need more information

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

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

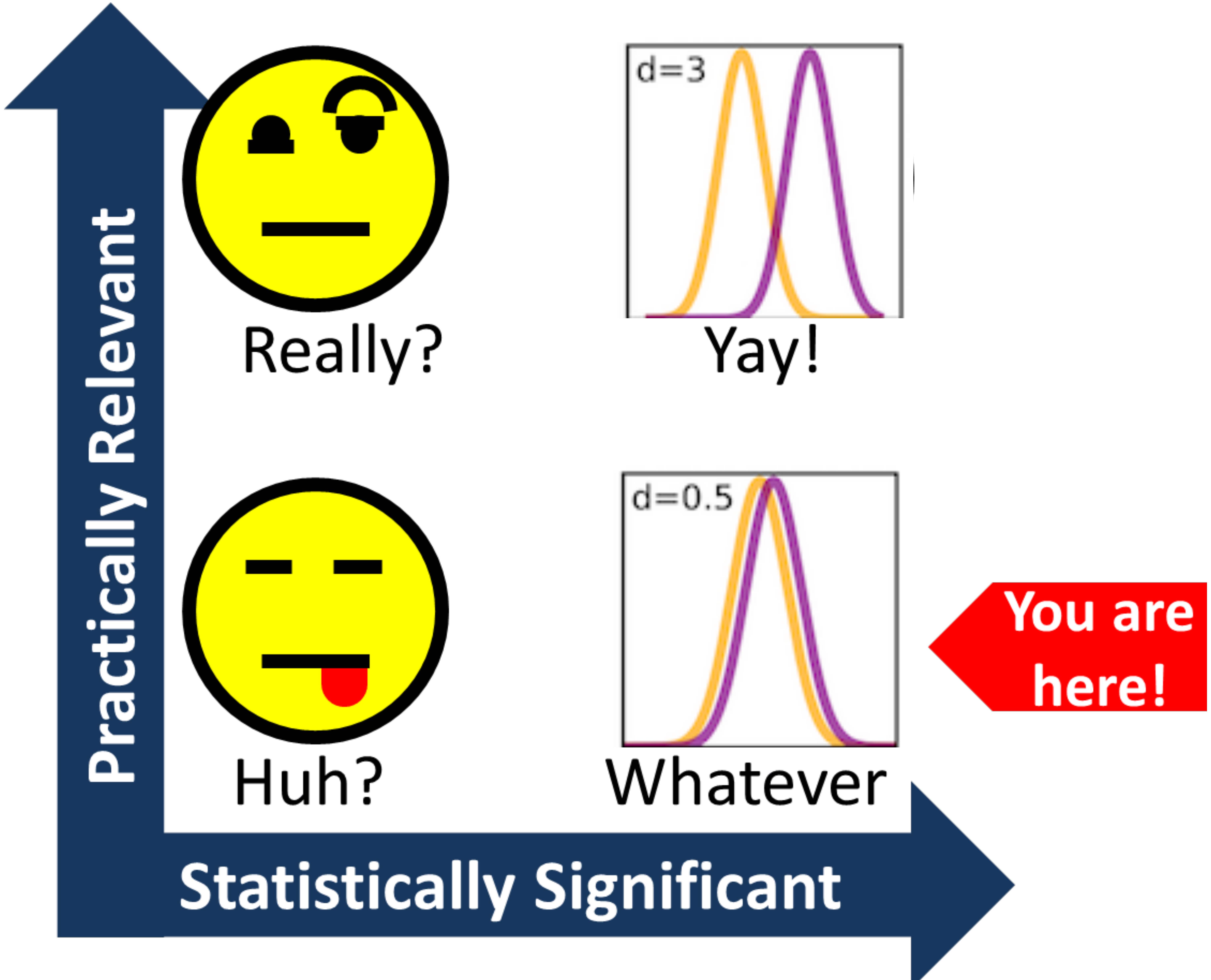
D
Data are not continuous

https://www.researchgate.net

Cohen's *d*

Cohen's *d* is defined as the difference between two means divided by a standard deviation for the data

Effect sizes!



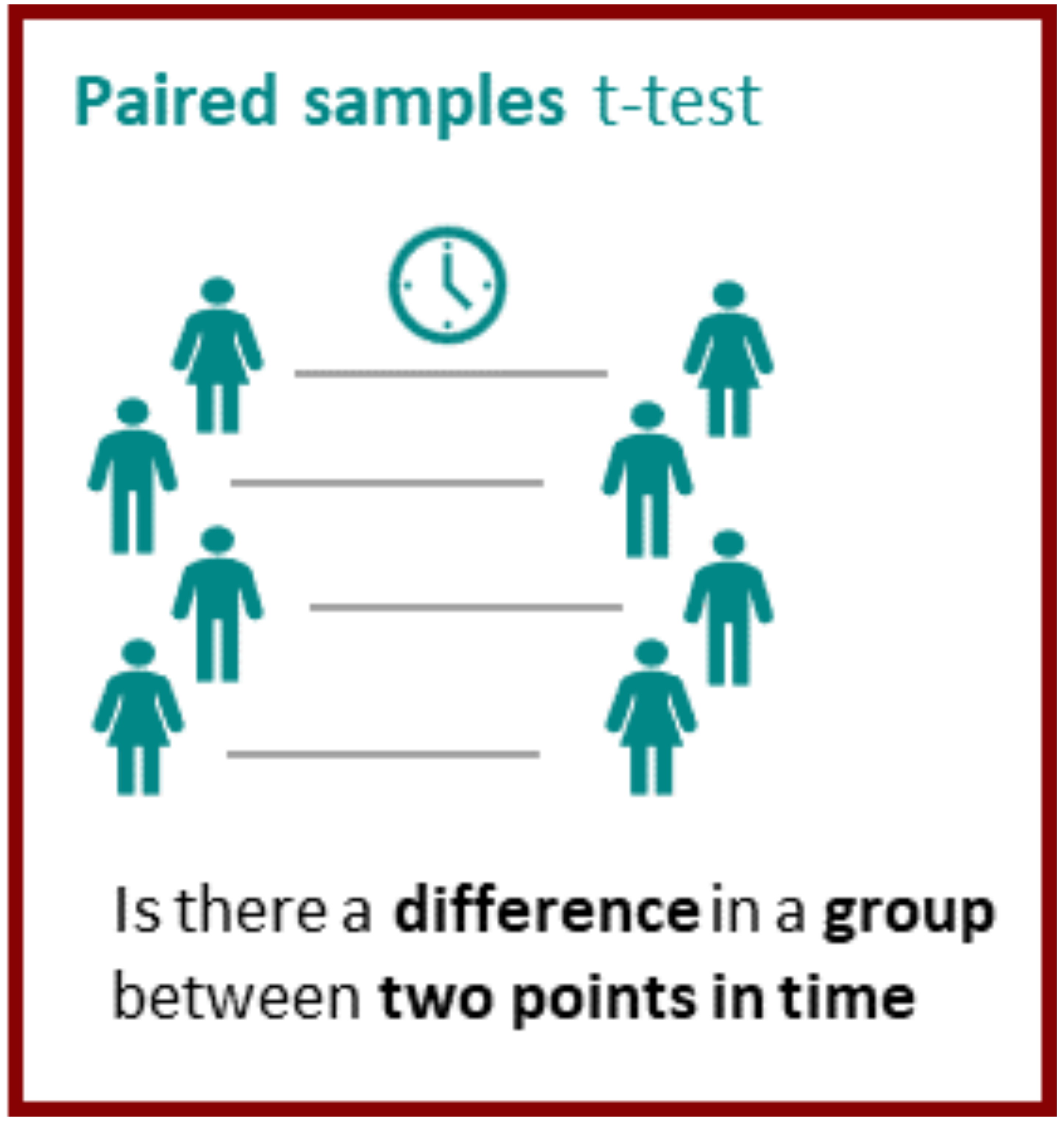
Statistical power

- A weak test cannot see the difference between Cohen's d=0.15 groups with 57 samples. A strong test can.
    - But the weaker test could with 157 samples
    - And the weaker test could do it with 57 samples for d=1.5 groups

Paired data

Default choice

More statistical power when this is the right choice



SciPy: stats.ttest_ind()

SciPy: stats.ttest_rel()

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

False positives

Saying its a real difference when its actually NOT

- Historically people take $p<0.05$ as decent evidence that the difference is real
    - 1 in 20 chance that the data would look this weird assuming H_0
    - This is BS. Some dead old guy thought it was a good rule of thumb
- So… if you test 20 different things and all of them have p just under 0.05, what do you expect to observe

https://xkcd.com/882/

    - $p_{\text{false positive}} = 1 - (p_1 p_2 \ldots p_{20})$
    - Conclusion: if we test a lot of different things and we want to be sure that we don't have a false positive result we need to correct for the multiple comparisons

The interpretation of p-values in NHST

- Even though we just talked about p-value AS IF its the probability of a false positive ITS NOT NOT NOT NOT NOT the so called Type I error rate
- People just do this. They are wrong
- P-values are $P(D \mid H_0)$… Type I error rate (false positive rate) is $P(H_0 \mid D)$
- Bayes rule: $P(H_0 \mid D) = P(D \mid H_0)P(H_0)$
  - If this is the first study ever on a topic then we know less about the truth than if this is the 36th study on a topic
- Empirically a p=0.05 is usually a **MUCH BIGGER Type 1** error rate
- Simulation / math studies tell us that we can expect a Type 1 error rate of 23 to 50% for p=0.05