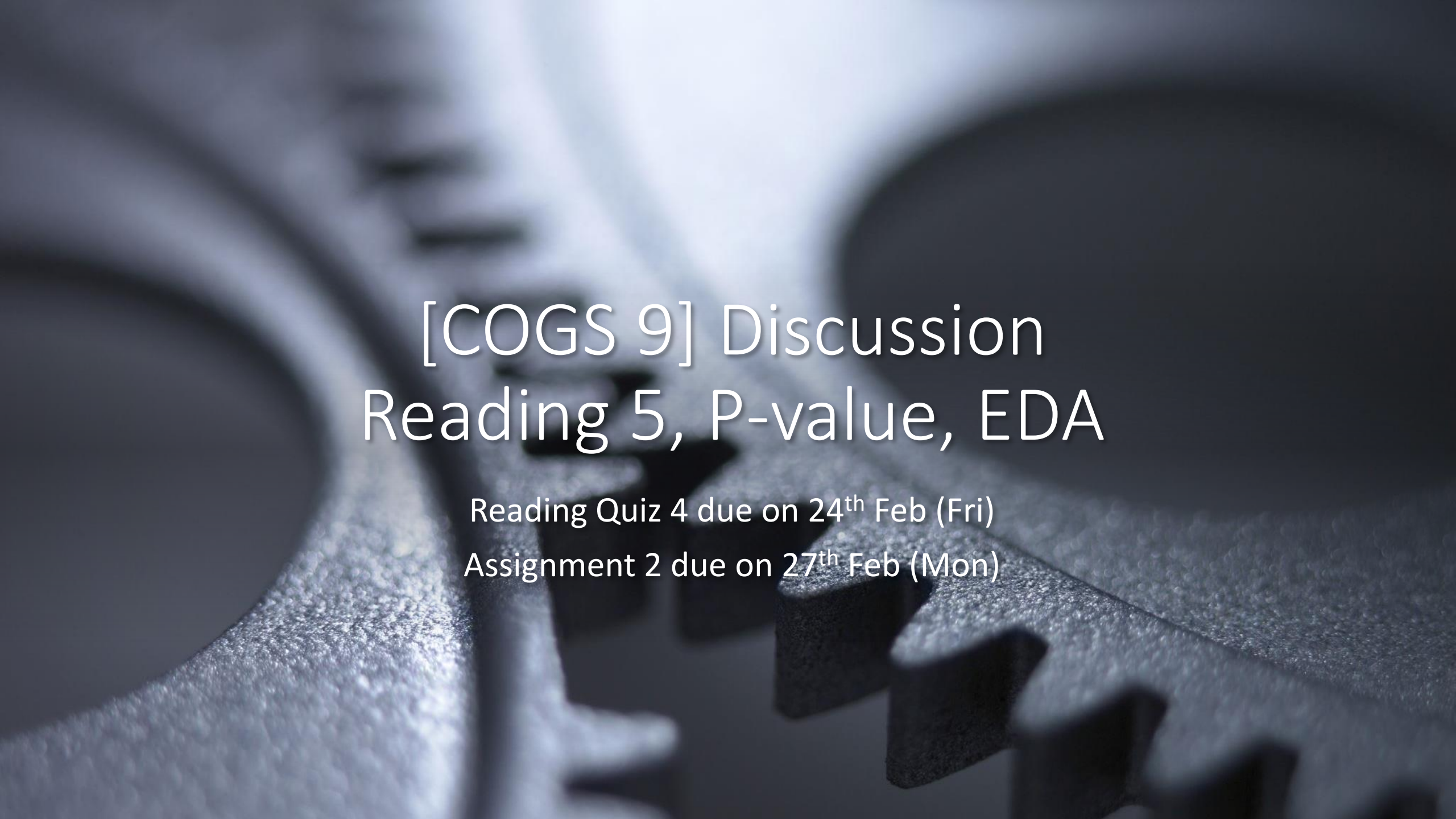


# [COGS 9] Discussion Reading 5, P-value, EDA

Reading Quiz 4 due on 24th Feb (Fri)

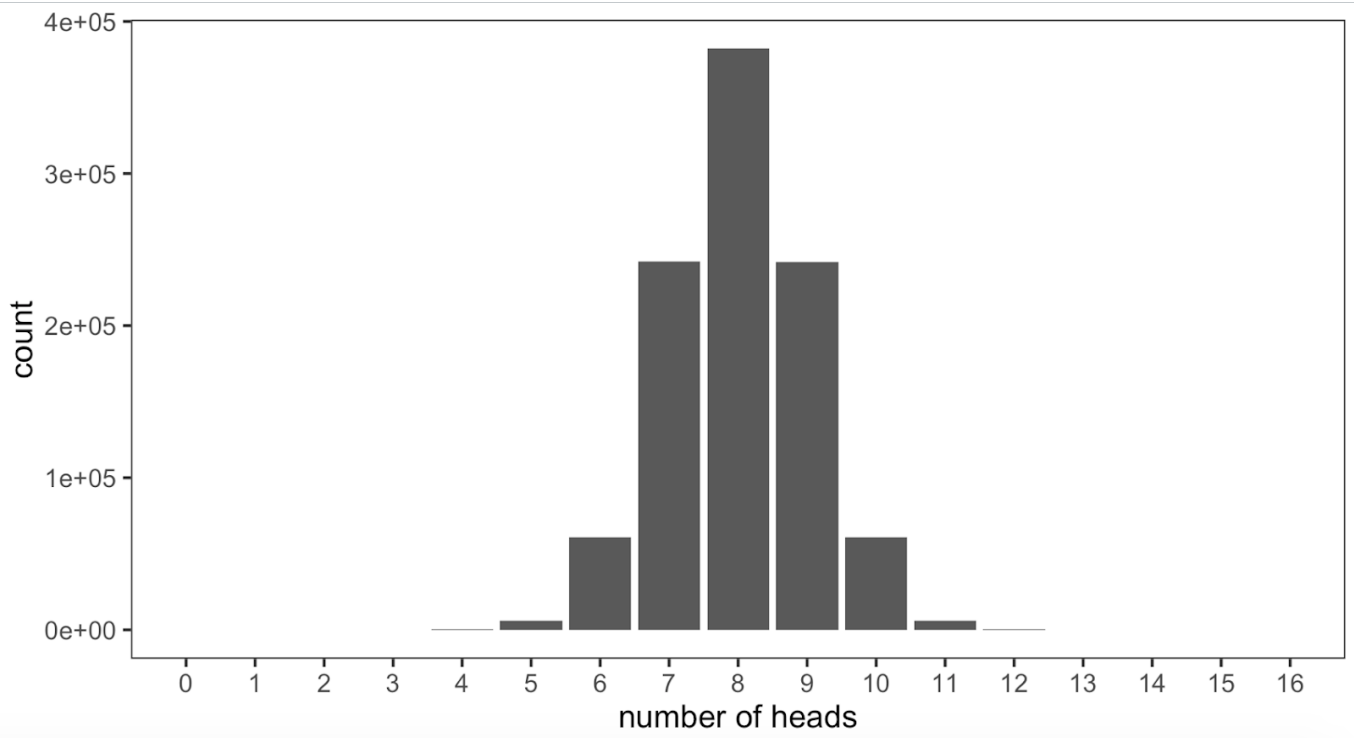
Assignment 2 due on 27th Feb (Mon)

# P-value

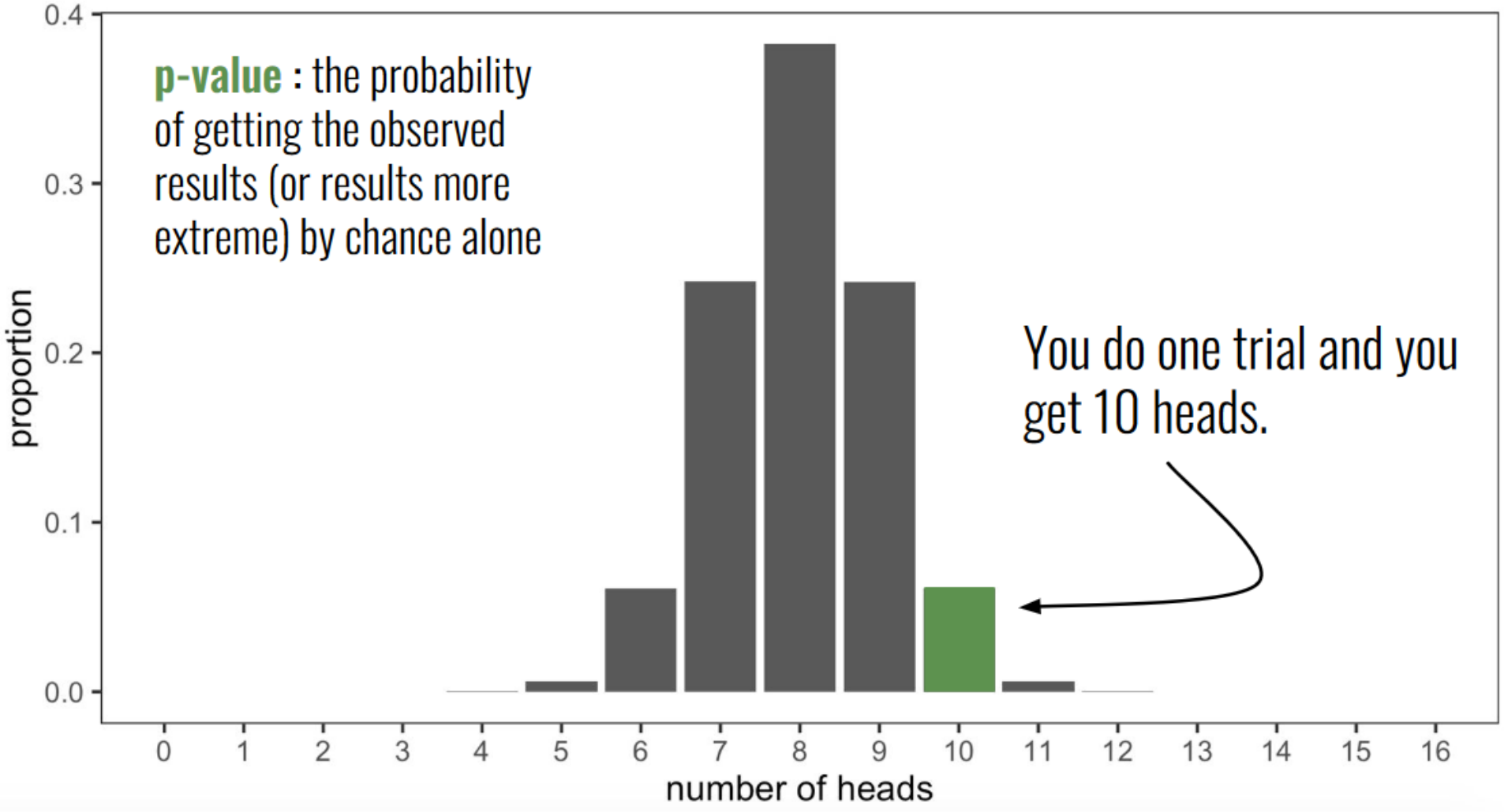
P-value = Probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct

Assuming a fair coin

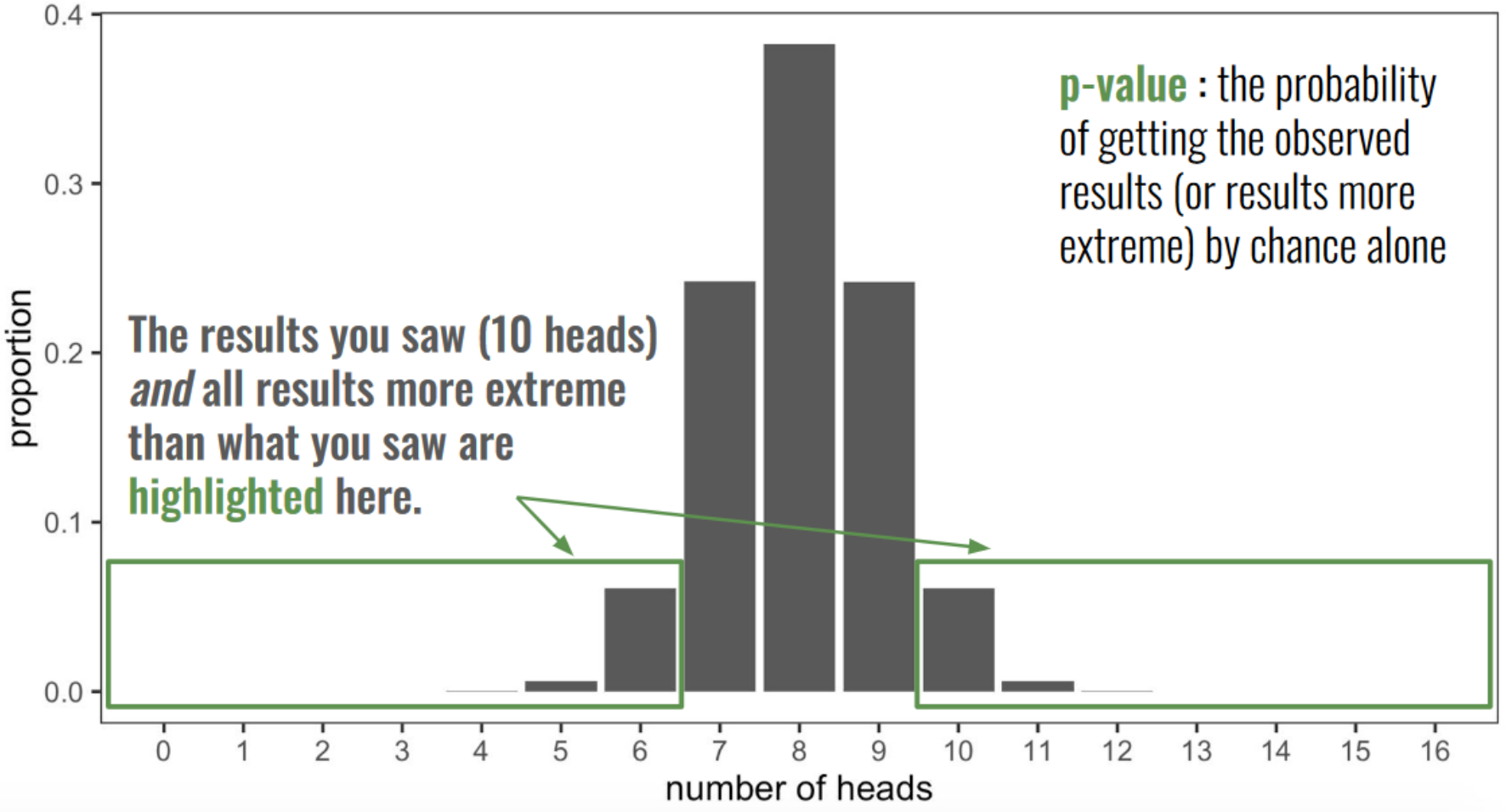
If we flip a coin 16 times and record the number of heads and then repeat this 16 flip trial 1 million times



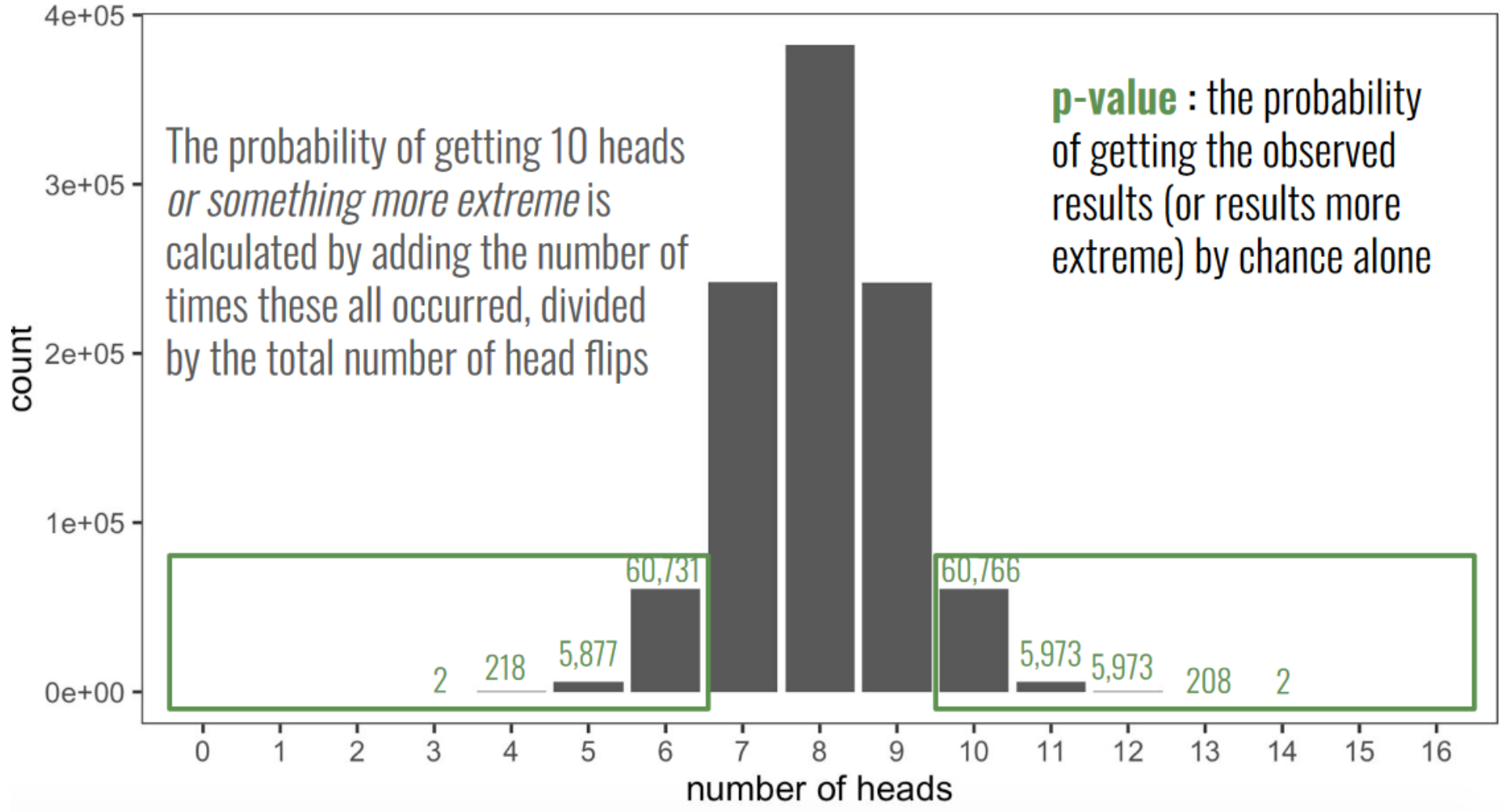
# P-value


p-value : the probability of getting the observed results (or results more extreme) by chance alone
You do one trial and you get 10 heads.
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

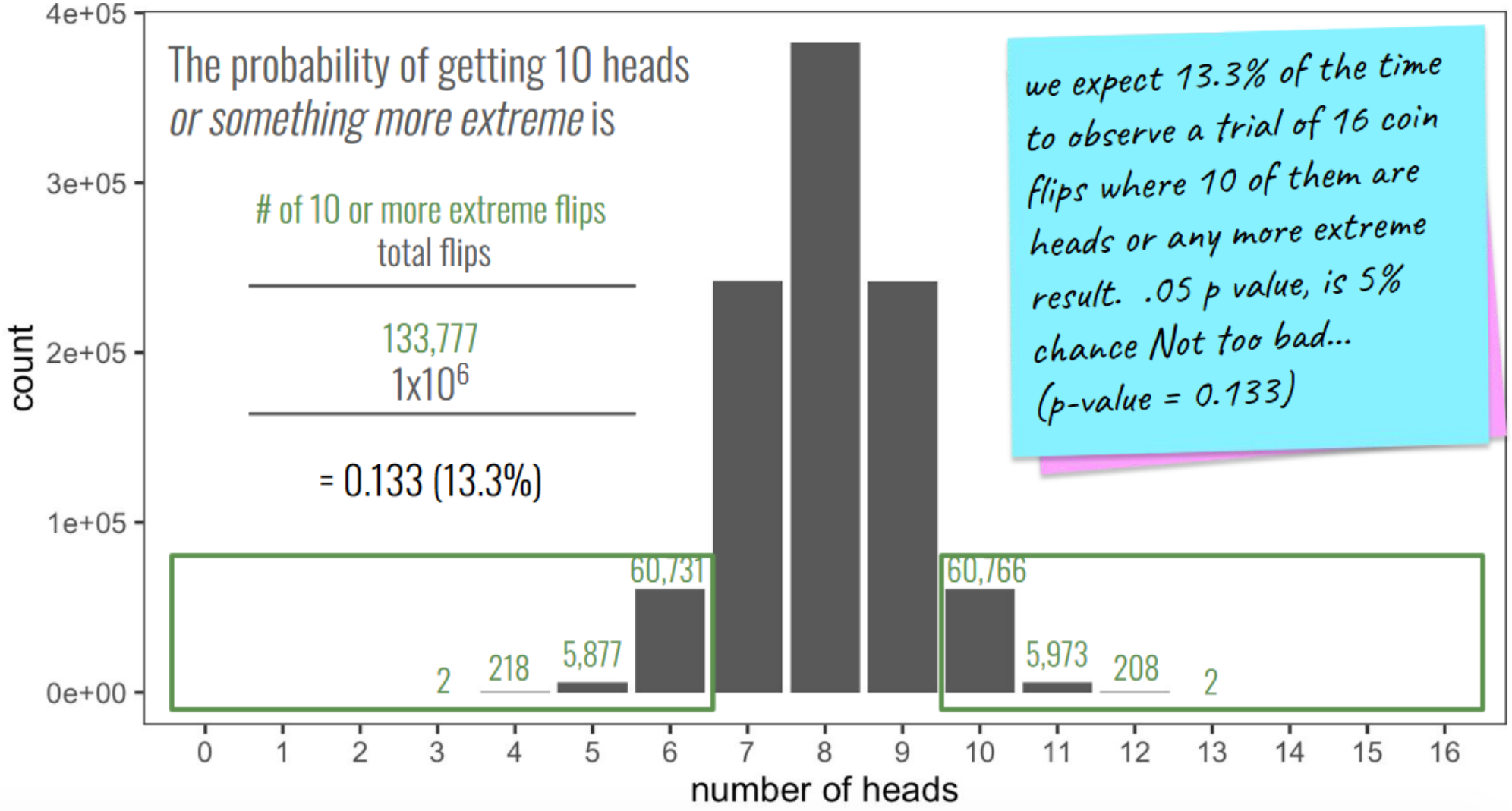
# P-value


0.4
0.3
0.2
0.1
0.0
proportion
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.
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
number of heads

# P-value


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
2
218
5,877
60,731
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

# P-value


The probability of getting 10 heads
*or something more extreme* is
# of 10 or more extreme flips
total flips
133,777
1x10^6
= 0.133 (13.3%)
we expect 13.3% of the time to observe a trial of 16 coin flips where 10 of them are heads or any more extreme result. .05 p value, is 5% chance Not too bad... (p-value = 0.133)
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

# Exploratory Data Analysis

Data -> Descriptive Analysis -> Exploratory Analysis -> Product

Exploratory: Inferential, Predictive, Causal, Mechanistic

- Inferential: Statistics, Frequentist, Bayesian, Text & Geospatial analysis
- Predictive: Statistical Learning/ML, Deep, Reinforcement Learning
- Causal: How variable X correlates to Y
- Mechanistic: How much does variable X affect Y

Univariate, Bivariate, Multivariate

Explanatory (Independent) vs Response (Dependent) variables

Source of data (Zipcode vs hometown), explore missing data

Don’t do EDA to give you the result you want

# Exploratory Data Analysis

## Checklist of things to do during EDA

- ❑ Investigate missing values
- ❑ Understand outliers
- ❑ Add filters, transform and scale data
- ❑ Calculate numerical summaries
- ❑ Generate plots to explore relationships
- ❑ Handle proportions correctly
- ❑ Use tables to scan data
- ❑ Search for patterns

# EDA
# Demo