1495 UNIVERSITY OF ABERDEEN | CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

A/B Testing

Data Mining & Visualisation

Lecture 7

2025

Recap

- Levels of measurement
  - Nominal, ordinal, interval, ratio

- Relationships in Data
  - Correlation

Today…

- A/B Testing
- Null Hypothesis Statistical Testing

A/B Testing

A/B Testing

A/B testing, or ‘split testing’, is a widely-used approach for running controlled, randomised experiments.

In short, it involves comparing two or more variants of something (an interface/design/product etc.).

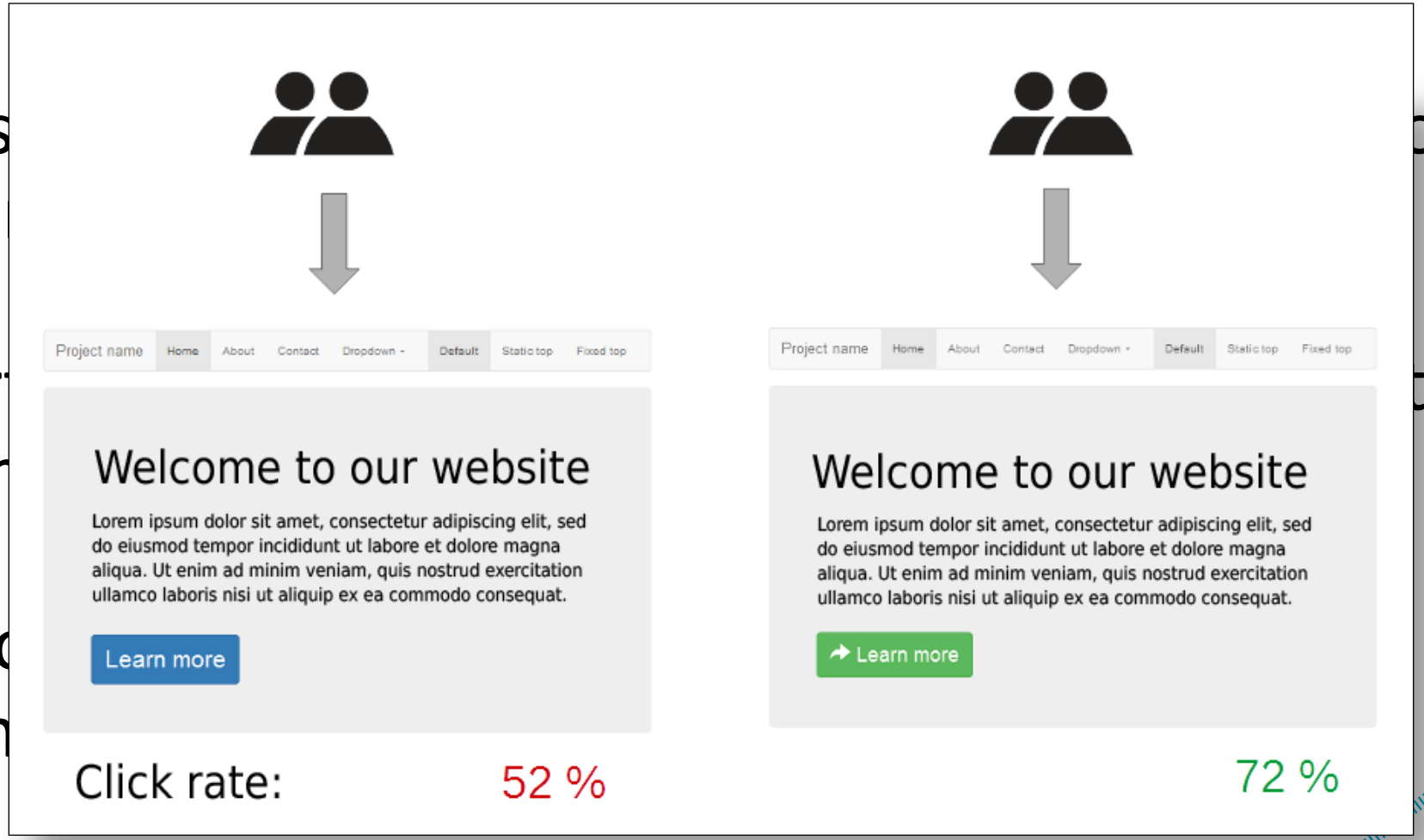
Based on some performance indicators or metrics, we determine if one variant is better than the other.

A/B Testing

A/B tes…bach for run…

In shor…ts of someth…

Based …we determ…



A/B Testing – Scenario

Let’s say you work for a popular online retail website, which sells products to thousands of customers.

The CEO wants to introduce a discount for your customers, to try to generate more profit.

They ask which discount would be better: 5% or 10%?

A/B Testing – Scenario

Let's say, broadly speaking:

5%	10%
• Less discounted products	• More discounted products
• More profit per sale	• Less profit per sale
• Less likely to make sales	• More likely to make sales

So how do we know whether one strategy is better?

A/B Testing – Scenario

One possible experimental Setup:

- Give 1,000 customers (group A) a 5% discount
- Give another 1,000 customers (group B) a 10% discount
- Wait 7 days for the experiment to run and data to materialise

A/B Testing – Scenario

Once we have waited for the experiment to run, we can analyse the data and interpret the results.

We can then determine whether one discount led to higher profits, and deploy that discount to more people.

A/B Testing

A/B testing has wide applicability across a range of contexts. It might involve testing particular:

- Interface designs
- Email campaigns
- Interventions
- Rewards
- Etc.

A/B Testing

It might involve comparing variants for their:

- Revenue
- Click-through rates
- Subscriptions
- Retention
- Etc.

A/B Testing

A/B testing will typically involve some form of statistical testing, as opposed to simply comparing descriptive statistics for the two outcomes.

But why is this the case?

A/B Testing

Importantly, we can't always reliably determine 'better' based on descriptive statistics.

	Group A	Group B
Number of Customers	9,994	10,025
Total Revenue	$78,552.84	$78,395.50
Revenue per Customer (Mean)	$7.86	$7.82
Revenue per Customer (Std. Dev.)	$5.45	$4.97

Flipping a Coin

To demonstrate this, let's say we flip a coin 10 times:

coin_flip = { H, H, T, H, H, T, H, T, T, H }

In this case, the coin toss came up **Heads** 6 times.

> Does this mean that coin tosses will always come up Heads 60% of the time?

Flipping a Coin

In reality, randomised datasets will be subject to chance:

- If you flip the coin again, maybe **Tails** comes up 6/10
- If you re-run the A/B test, maybe **Group B** has a higher mean revenue per customer

So how do we determine the extent to which chance is a factor in our findings?

Null Hypothesis Statistical Testing

ABERDEEN 2040

Null Hypothesis Statistical Testing

’Null Hypothesis Statistical Testing’ (NHST) refers to a set of statistical methods for determining whether data support a particular hypothesis.

E.g. ‘Barley variant A leads to a larger yield than variant B’.

Null Hypothesis Statistical Testing

The **null hypothesis** (H_0) is that the effect being studied does not exist.

- i.e. that variant A leads to the same yield as variant B

The **alternative hypothesis** (H_1; H_a) is what you're looking to test.

- i.e. that variant A leads to a larger yield than variant B

Null Hypothesis Statistical Testing

The idea is as follows:

- We use our statistical test to calculate a ‘p-value’, where $0 < p < 1$
- If this p value is below a certain threshold (usually .05), we might say that results are ‘statistically significant’

So given that this p-value is such a core aspect of a NHST, what does it actually mean?

The p-Value

The **p-value** is defined as:

The probability of receiving test results at least as extreme as those observed given the assumption that the null hypothesis is true.

This definition is very precise, but fairly convoluted.

So let's break it down in the context of an NHST…

The p-Value

Flipping the definition round, we're essentially saying:

If we assume that the null hypothesis (H_0) is true…

(i.e. that there is no difference between groups A and B...),

…the p-value is the probability that we would expect to observe the difference between groups A and B

(or a difference more extreme than that observed).

The p-Value

When this probability is low (< .05; 5%; 1 in 20), we say:

‘Given that the difference we observed between groups A and B is so unlikely (if the null hypothesis were indeed true), we therefore reject the null hypothesis.’

When we **reject the null hypothesis**, we say that the result of the NHST is ‘**statistically significant**’.

Effect Size

Also important is the **effect size** of what we are testing.

Effect size is the strength of relationship between the two or more variables or groups.

In other words, how large is the effect (e.g. that variant A has on the yield, compared to variant B)?

Sample Size

The outcome of a NHST will also depend on the **sample size** (i.e. how many data points we have).

Typically, the sample size will be one factor that we (as data scientists) can have some control over.

More data will mean that we have a better chance of finding statistically significant results.

Sample Size

As we increase the sample size (gather more data), we aim to more accurately reflect the population.

In other words, increasing the sample size reduces the sampling error.

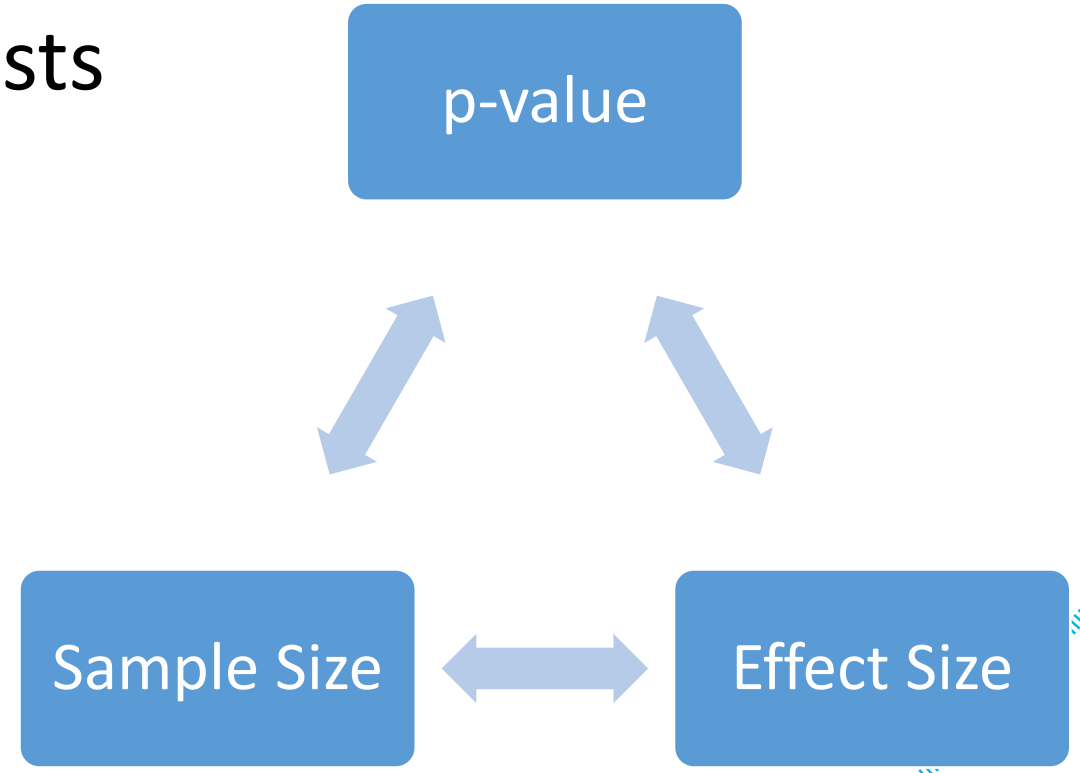
Therefore, more data points can give us more confidence that our data reflects that of the population.

One last important point…

The outcome of statistical tests will typically depend on

three inter-related factors:

- The p-value
- The sample size
- The effect size

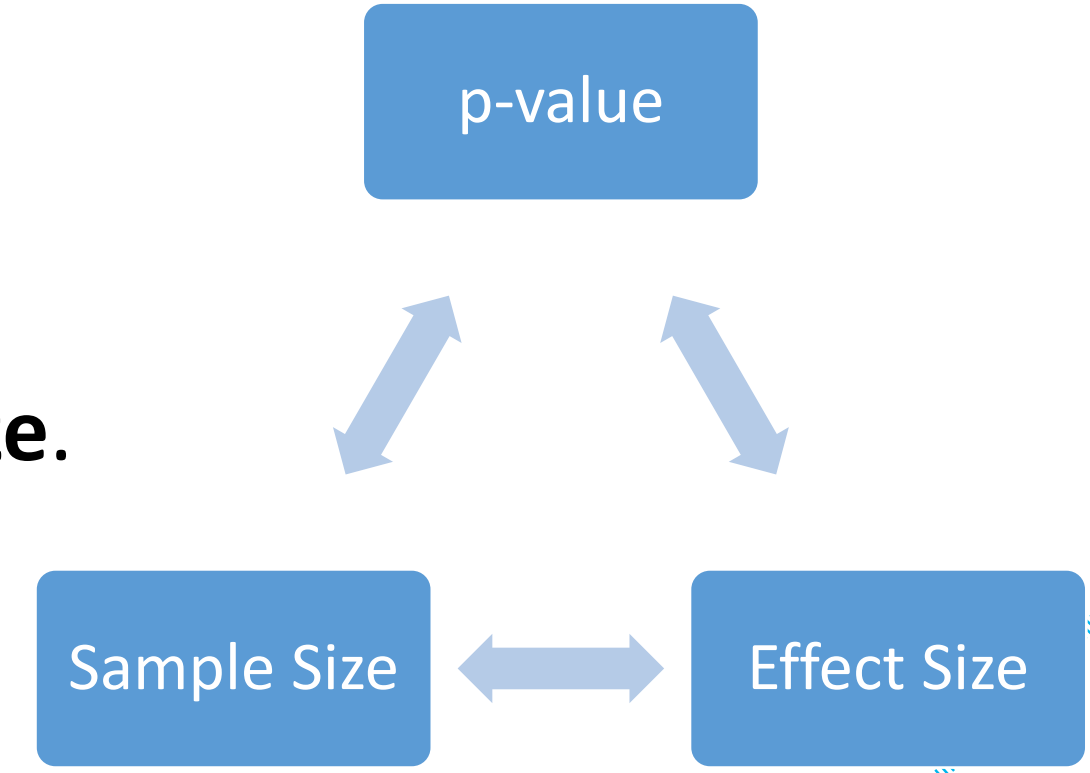


One last important point…

If the **p-value** is fixed:

A stronger **effect size** will require a smaller **sample size**.

A higher **sample size** will detect a weaker **effect**.

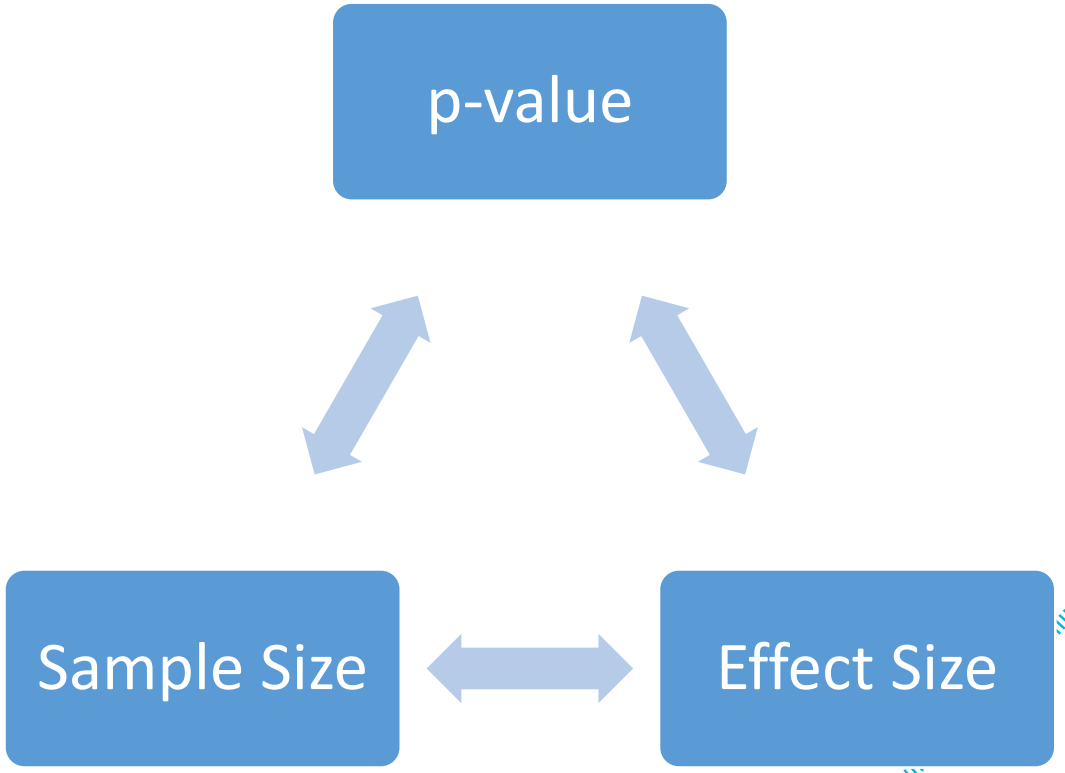


One last important point…

If the **sample size** is fixed:

A stronger **effect size** will have a lower **p-value**.

A higher **p-value** will result from a weaker **effect size**.

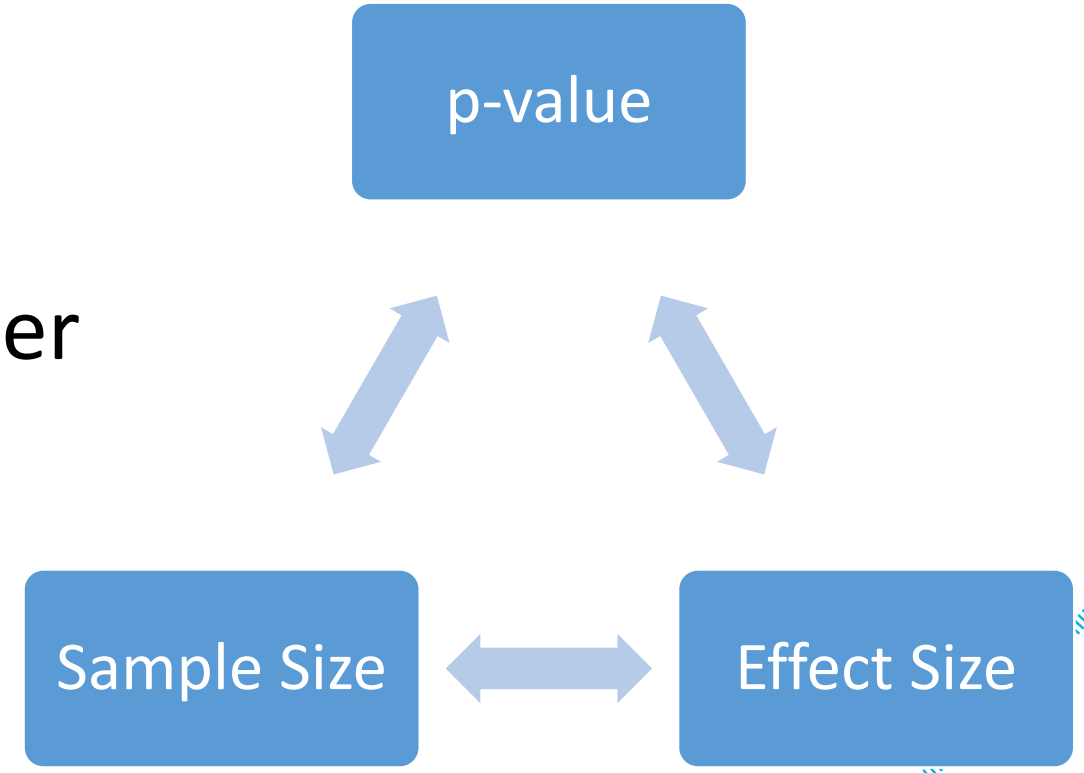


One last important point…

If the **effect size** is fixed:

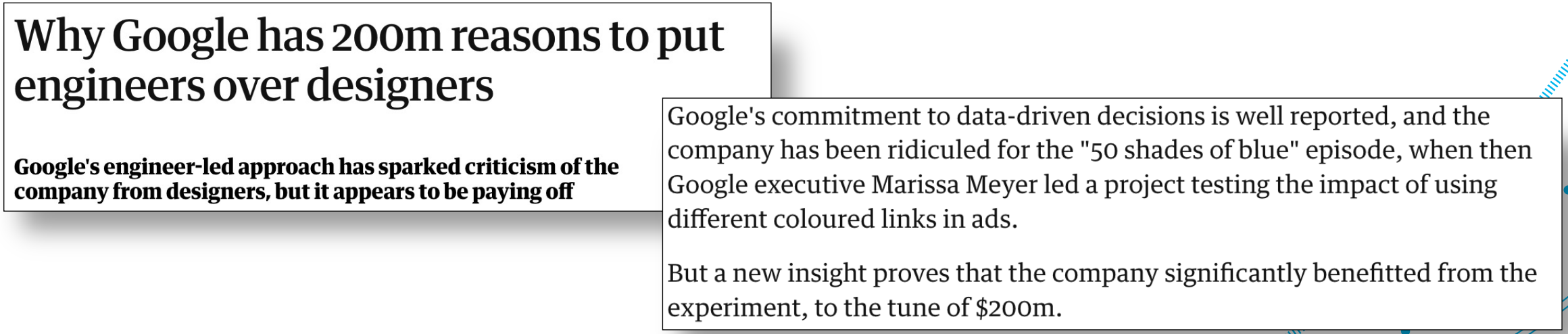
A larger **sample size** will lower the **p-value**.

A higher **p-value** will result from a lower **sample size**.



One last important point…

In other words, when $p < .05$ is our fixed cut-off, we can find statistically significant results from the smallest of effect sizes – as long as we have enough data.



Why Google has 200m reasons to put engineers over designers

Google's engineer-led approach has sparked criticism of the company from designers, but it appears to be paying off

Google's commitment to data-driven decisions is well reported, and the company has been ridiculed for the "50 shades of blue" episode, when then Google executive Marissa Meyer led a project testing the impact of using different coloured links in ads.

But a new insight proves that the company significantly benefitted from the experiment, to the tune of $200m.