1495 UNIVERSITY OF ABERDEEN | CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

A/B Testing 2

Data Mining & Visualisation

Lecture 8

2025

Recap

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

- Relationships in Data
  - Correlation

Today…

- Null Hypothesis Statistical Testing

Variants of NHST

Variants of NHST

Let’s say e.g. we have two versions of a UI, and we want to determine which version leads to more revenue.

We know that we want to use NHST… so what now?

Variants of NHST

There are lots of variants of NHST.

Determining which one you should select in any given context could be an entire course on its own.

Variants of NHST

In any scenario, the exact test that you should run depends on what you're comparing and why.

- What are you looking for?
- What type(s) of data do you have?
- How many groups?
- Etc…

Variants of NHST

Note: This is a simplification, for illustrative purposes only. <u>You will not be expected to memorise this.</u>

Source: https://ctil.dundee.ac.uk/kb/stats-bites-choosing-your-analysis/

Example: Independent Samples T-test

Independent T-test

Let’s start with a simple example, where we have two versions of a user interface: Version A and Version B.

Does version A lead to more revenue than version B?

Independent T-test

Based on their User ID, our users are assigned either:

Version A

user_id	Revenue (£)
1	33
3	39
5	30
7	37
9	36

or

Version B

user_id	Revenue (£)
2	38
4	41
6	43
8	36
10	44

Note that our two ‘groups’ contain ‘independent samples’ from different users); the revenue values of one group do not influence the revenue values of the other.

Independent T-test

user_id	Revenue (£)
1	33
3	39
5	30
7	37
9	36

user_id	Revenue (£)
2	38
4	41
6	43
8	36
10	44

	Version A	Version B
Sum:	175	202
Sample mean:	35	40.4
Std Dev:	3.54	3.36

Independent T-test

Version B has a higher total revenue, and a higher mean.

But recall from the last lecture that NHST is concerned with whether this is ‘statistically significant’.

So let’s see what happens if we run a NHST: The independent t-test.

Independent T-test

The formula for an independent t-test is:

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

---> difference between our sample means

---> Standard error of our samples (i.e. a measurement of variance)

Note: This is included for informational purposes. You will not be expected to know this formula, or calculate anything with it, for the exam.

Independent T-test

The formula for an independent t-test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{35 - 40.4}{\sqrt{\frac{3.54^2}{5} + \frac{3.36^2}{5}}} = -2.47$$

This gives us the ‘t statistic’, which we can then use to look up the corresponding p value using a ‘t table’.

Note: This is included for informational purposes. You will not be expected to know this formula, or calculate anything with it, for the exam.

Independent T-test

In practice, we're more likely to calculate this using e.g. python:

```
ver_a = [33, 39, 30, 37, 36]
ver_b = [38, 41, 43, 36, 44]

stats.ttest_ind(ver_a, ver_b)
```

✓ 0.0s

```
Ttest_indResult(statistic=-2.47508594197617, pvalue=0.038404727172098745)
```

p value is 0.038, which is < 0.05. Therefore, the difference between the groups is statistically significant.

In other words, if the null hypothesis were true, it would be unlikely (3.8%) that we'd observe results at least as large as that observed.

Variants of NHST

So far, we have seen one variant of NHST: The independent t-test.

However, there are lots of variants of NHST, and of the family of different t-tests.

We will start off by talking about some of the high-level concepts that are important to understand.

Dependent vs Independent Samples

Variants of NHST

In the previous example, our samples were independent. But what if this was not the case?

Variants of NHST

Let's say, instead of having two distinct groups of users, each user is either shown Version A or Version B randomly per session.

user_id	Version A Revenue (£)	Version B Revenue (£)
1	33	44
2	39	41
3	30	29
4	37	39
5	36	35

Since each user has a revenue value for both Version A and Version B, the groups are *not independent*. They are *paired*.

Variants of NHST

In this case, we cannot use the independent t-test (since our samples are not independent).

However, we can use another member of the t-test family: a *paired samples t-test*.

Paired T-test

The formula for a paired t-test is:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

---> The mean of differences between our samples

---> The standard deviation of the differences between our samples

---> The (square root of the) number of paired samples

Note: This is included for informational purposes. You will not be expected to know this formula, or calculate anything with it, for the exam.

Paired T-test

user_id	Version A Revenue (£)	Version B Revenue (£)	Difference
1	33	44	11
2	39	41	2
3	30	29	-1
4	37	39	2
5	36	35	-1

The formula for a paired t-test is:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{2.6}{\frac{4.93}{\sqrt{5}}} = -1.18$$

mean = 2.6
stdev = 4.93
n = 5

Note: This is included for informational purposes. You will not be expected to know this formula, or calculate anything with it, for the exam.

Paired T-test

In python:

```
ver_a = [33, 39, 30, 37, 36]
ver_b = [44, 41, 29, 39, 35]

stats.ttest_rel(ver_a, ver_b)
```

✓ 0.0s

```
Ttest_relResult(statistic=-1.1793839502289376, pvalue=0.3036016411350295)
```

p value is 0.30, which is > 0.05. These results are not statistically significant.

In other words, if we assume that the null hypothesis is true (that there is no real difference between the revenues per version), the probability of observing results at least as extreme as these is around 30%.

We cannot rule out that these results might simply be down to chance.

Number of Groups

Variants of NHST

In the previous example, we had two groups.

But what if we had 3 or more versions of our UI?

We cannot compare 3+ groups with an independent or paired t-test.

Variants of NHST

user_id	Revenue (£)
1	33
4	39
7	30
10	37
13	36

Version A

user_id	Revenue (£)
2	38
5	41
8	43
11	36
14	44

Version B

user_id	Revenue (£)
3	38
6	41
9	43
12	36
15	44

Version C

ANOVA

The next ‘family’ of tests we’ll talk about are referred to as ANOVA (Analysis of Variance), which allow us to compare 3+ groups.

There are several variants of ANOVA which are suitable for different situations (including versions for more complex experimental designs).

We’re just going to focus on the one-way independent ANOVA, and the one-way repeated measures ANOVA. These correspond to the independent and paired t-tests respectively.

ANOVA

ANOVAs return a significant p value if there is a statistical difference between at least one pair of groups.

However, an ANOVA will not tell us which groups are statistically different.

For that, we need to run follow-up t-tests between each combination of groups (pairwise comparisons).

ANOVA

user_id	Revenue (£)
1	33
4	39
7	30
10	37
13	36

Version A

user_id	Revenue (£)
2	38
5	41
8	43
11	36
14	44

Version B

user_id	Revenue (£)
3	38
6	41
9	43
12	36
15	44

Version C

E.g., if a one-way independent ANOVA is significant, we can then run pairwise comparisons:

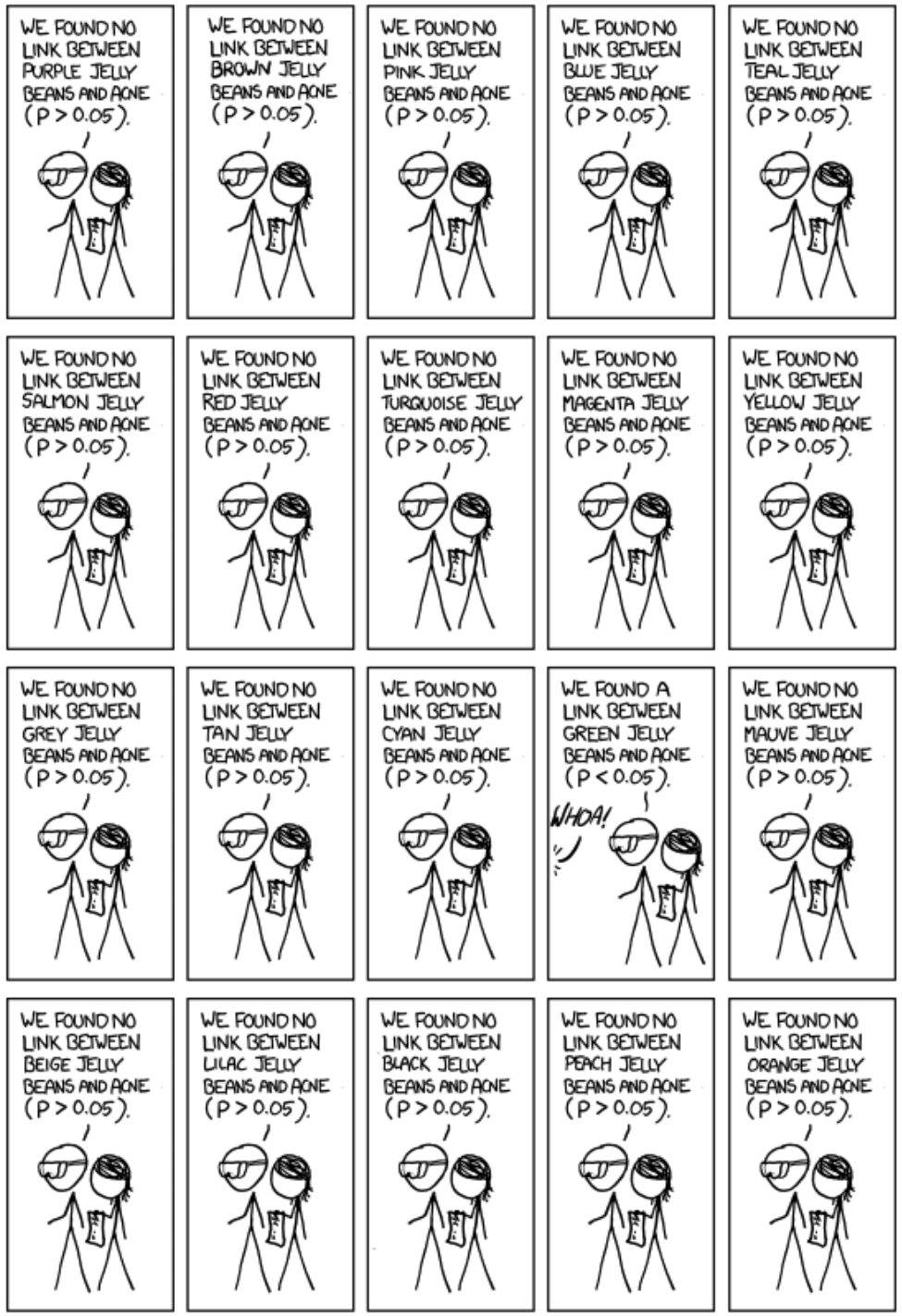
- an independent t-test between **Version A** and **Version B**
- an independent t-test between **Version A** and **Version C**
- an independent t-test between **Version B** and **Version C**

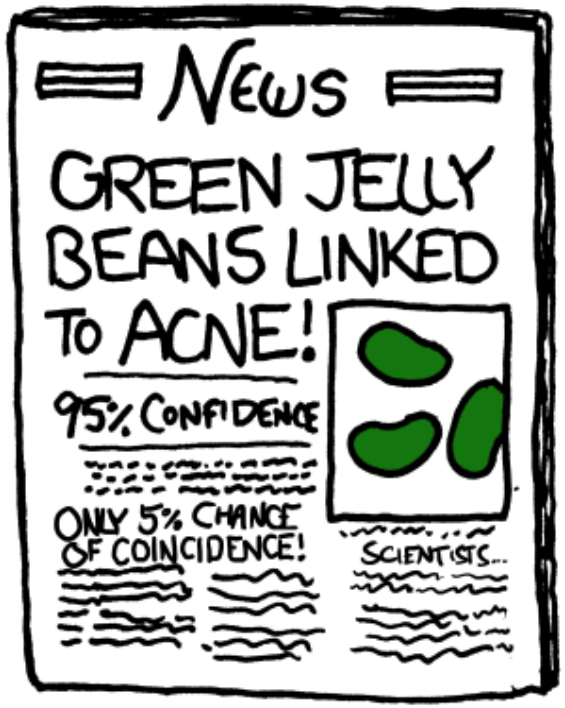
Family-Wise Error Rate

But doing this demonstrates a new quirk of statistical testing: The family-wise error rate.

Consider this: Say we have a p value cut-off of 0.05 (i.e. assuming that the null hypothesis is true, there is less than 5% probability that we'd see results at least as extreme as those observed).

But what happens if we just run lots of tests (e.g. 20+)?





Source: https://xkcd.com/882/

Family-Wise Error Rate

The family-wise error rate (FWER) refers to the following concept:

As you increase the number of NHSTs that you run, you increase the probability of making at least one false discovery (‘type one error’).

It is worth keeping this in mind when carrying out repeated NHSTs.

Family-Wise Error Rate

Whenever we do multiple NHSTs (like with our ANOVA example), we should correct for this error rate. There are several methods for this.

One example is the ‘Bonferroni correction’, where you divide the p value cut-off by the number of tests that you run.

E.g. when running our three pairwise independent t-tests, we might instead use a p value cut-off of 0.05/3 (= 0.0167).

Thus, repeated tests are less likely to be considered statistically significant.

Parametric vs Non-Parametric

Parametric vs Non-Parametric

Lastly, specific NHSTs are typically classified as either ‘parametric’ or ‘non-parametric’ tests.

Parametric tests make assumptions, including about the shape and distribution of underlying populations, and about the types of data.

Using a parametric tests on datasets that violate these assumptions can lead to misleading and inaccurate test results.

Therefore, always check that your data meets these assumptions before running the tests!

Parametric vs Non-Parametric

The assumptions for parametric tests will often relate to the following concepts:

- Normally distributed data
- Continuous (i.e. interval- or scale-level) data
- Homogeneity of variance (the variance of each group should be roughly equal)
- Observations should be independent

There are several ways to check that these assumptions are met, including some separate tests that you can run (e.g. Levene's test checks for homogeneity of variance).

Note that each parametric test will have its own specific assumptions, and the above assumptions will mean different things in different contexts (i.e. for different tests).

Parametric vs Non-Parametric

Non-parametric tests do not make the same assumptions about the underlying shape and distribution of the data.

So far, all of the tests we have discussed (independent and paired t-test, independent and repeated measures ANOVA) are parametric tests.

So when our data violates these assumptions, there are typically non-parametric equivalents of these tests that we can use instead.

Parametric vs Non-Parametric

Parametric Test	Non-Parametric Test Equivalent
Independent t-test	Mann-Whitney U test
Paired t-test	Wilcoxon signed rank test
independent measures ANOVA	Kruskal-Wallis test
repeated measures ANOVA	Friedman test

Getting Back to A/B Testing

ABERDEEN 2040

Getting Back to A/B Testing

So, we've gathered data from our A/B test. What now?

In short, we probably want to:

- Identify some variant of NHST that we want to run
- Check that our data meets any underlying or parametric assumptions
- Run that test (or run a non-parametric equivalent if assumptions are not met)
- Obtain a p-value
- Use that p value to determine whether the difference between the groups is 'statistically significant' (if $p<.05$)

Getting Back to A/B Testing

NHST is quite often the cornerstone of A/B testing.

Furthermore, p-values are used widely across data analysis and machine learning methods.

You may have come across them before and not known what they are (hopefully now you do!).

Getting Back to A/B Testing

By understanding the role of NHSTs, and by understanding how to determine which test to run, and how to analyse their results, you should now be able to design and interpret A/B tests.

These are commonplace across the field of data science (industry), and in academia (graduation thesis, MSc programmes, further research).