Quantitative methods for small data

DAMON WISCHIK

RSP unit OU28
Reference: lecture notes for IB Data Science

Who's still working with small data?

HCI, social science, medicine

- Small number of human subjects

Natural language processing (NLP)

- Small number of corpora

A typical small-data HCI experiment

SubjectID	Device	HitRate
1	touchpad	0.939
2	touchpad	0.975
3	button	0.940
4	button	1.000
5	button	0.915
⋮	⋮	⋮

Subjects played a game in which they have to shoot at a moving UFO.

- For firing, some subjects were told to tap a touchpad, and others were asked to press a button.
- Subjects have one shot per UFO. Their hit rate over a 3-minute game was measured.

Sense of Agency and User Experience: Is There a Link? (Bergström, Knibbe, Pohl, Hornbæk. ACM Trans. HCI. 2022)

SubjectID	Device	HitRate
1	touchpad	0.939
2	touchpad	0.975
3	button	0.940
4	button	1.000
5	button	0.915
⋮	⋮	⋮

response /
outcome metric /
dependent variable

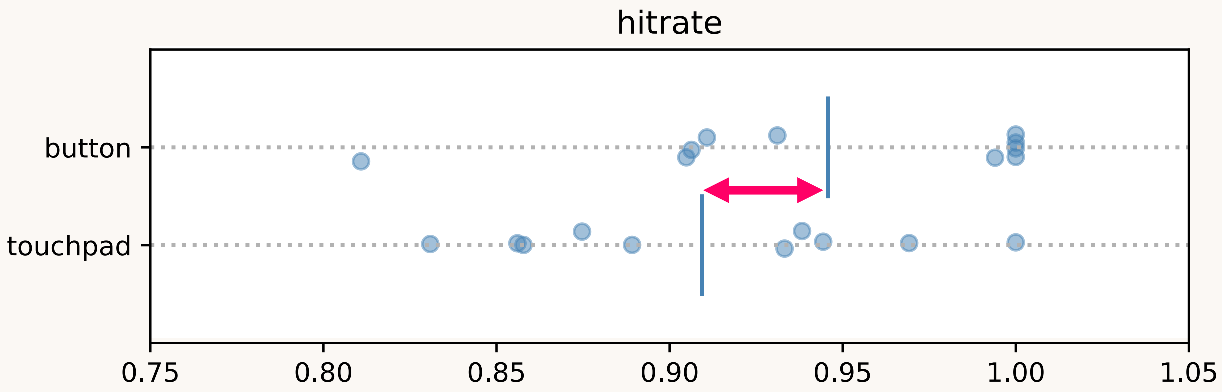
condition /
independent variable

experimental unit

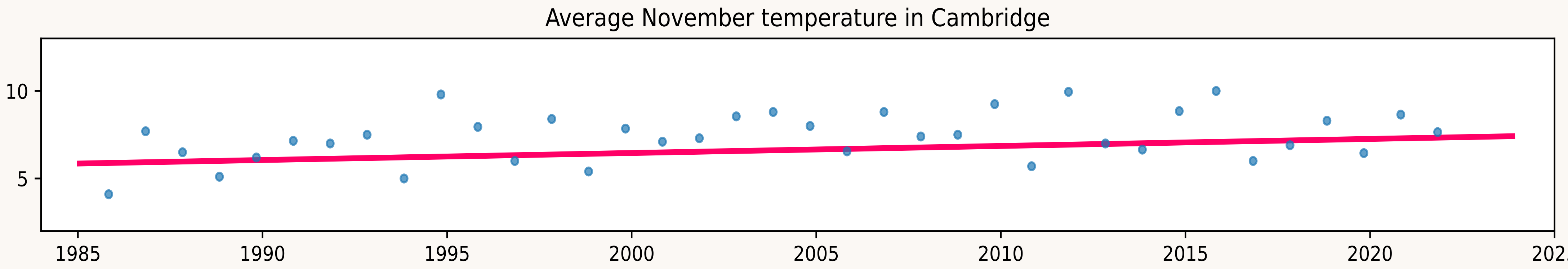
We want to learn

“How does the response depend on the condition?”

With small datasets, it's hard to untangle signal from noise

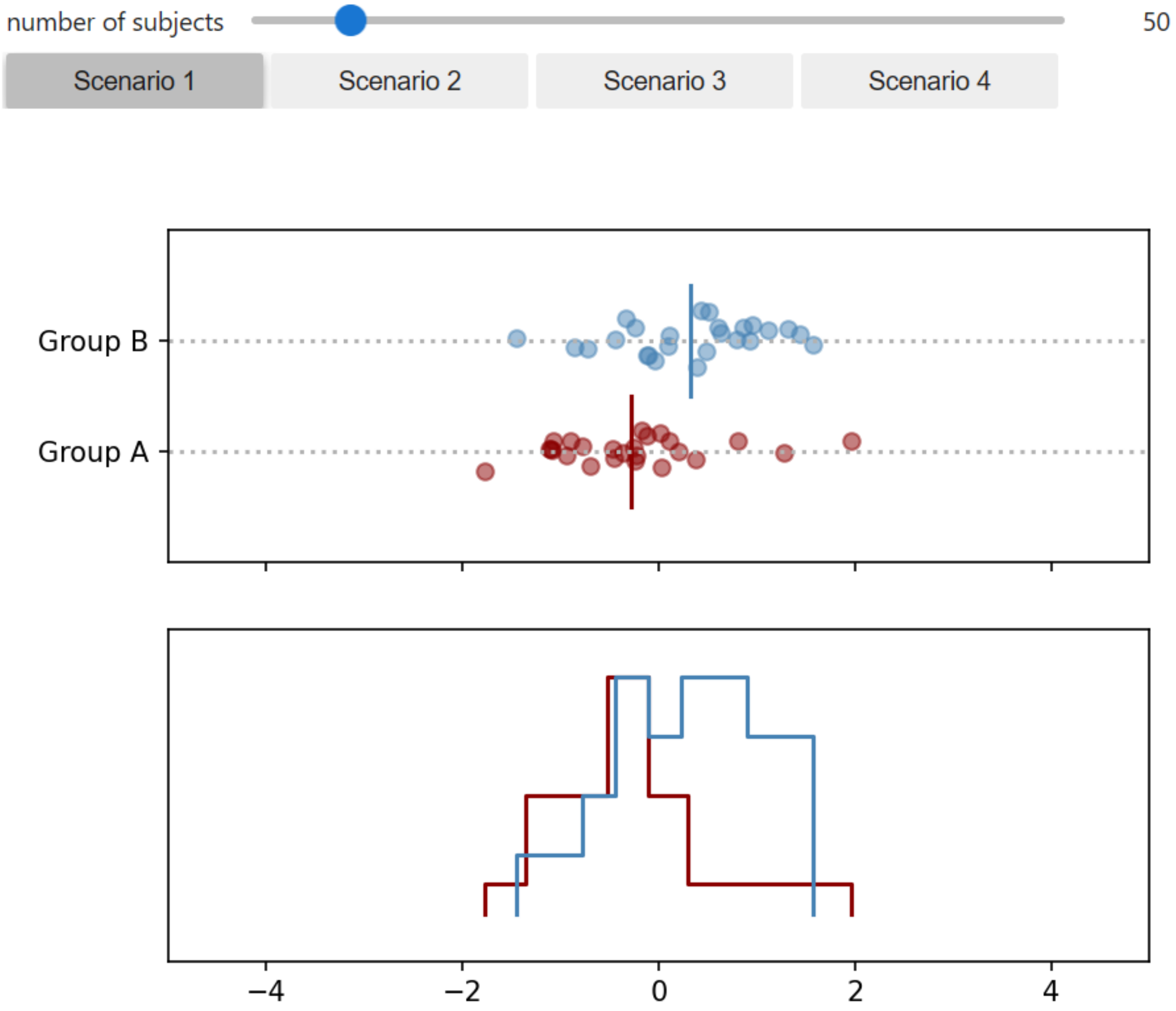


Button-users are 0.036 percentage points more accurate, on average.
But is this "real", or is it just noise?



Temperatures are increasing by 0.046°C per year. But is this "real", or is it just noise?

The *p*-value is a way to measure how confident we can be that the signal is real.



SubjectID	Device	HitRate
1	touchpad	0.939
2	touchpad	0.975
3	button	0.940
4	button	1.000
5	button	0.915
⋮	⋮	⋮

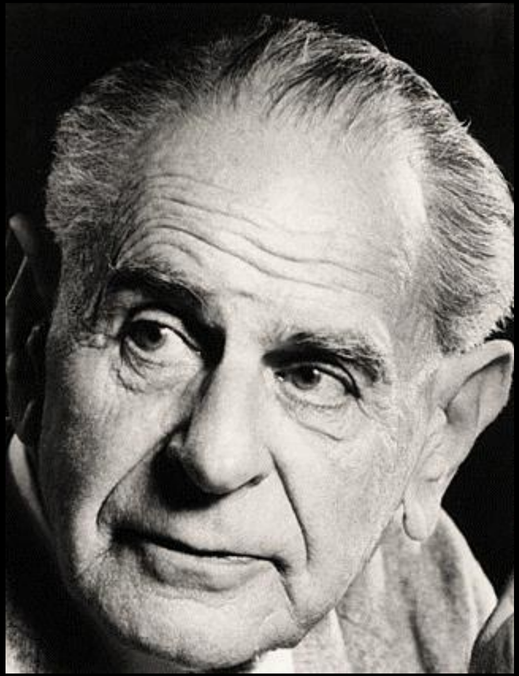
~~“The two groups have significantly different HitRate (t-test, $p = 0.020$).”~~

- Don’t confuse *significant* with *meaningful*
- Don’t use the word *significant* in any other context!

- With only two groups, it’s more informative to report a confidence interval rather than a *p*-value

The conceptual foundation of hypothesis testing

or
what type of statement am I making
when I report a *p*-value?



“Every genuine scientific theory must be falsifiable.

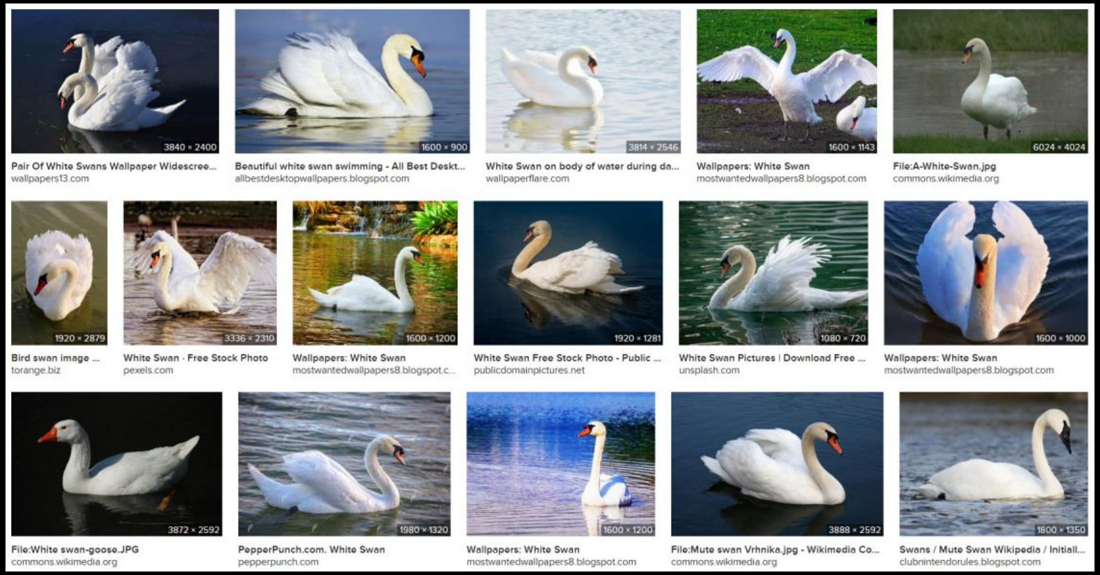
“It is easy to obtain evidence in support of virtually any theory; the evidence only counts if it is the positive result of a genuinely risky prediction.”

Karl Popper (1902–1994)

Why doesn't Popper believe in supporting evidence?

HYPOTHESIS
All swans are white, i.e.

$$\forall x \;\; \mathrm{IsSwan}(x) \Rightarrow \mathrm{IsWhite}(x)$$



ANALYSIS

The hypothesis is logically equivalent to

$$\forall x \;\; \neg\mathrm{IsWhite}(x) \Rightarrow \neg\mathrm{IsSwan}(x)$$

SUPPORTING EVIDENCE
This pot-plant isn't white, and it isn't a swan.

The hypothetico-deductive method

I wish to consider a hypothesis H_0

- If H_0 were true, we would expect …
- What really happened is …

The p-value is the probability of seeing what actually happened, assuming H_0 is true.

Is reality consistent with H_0?

$p < 5\%$ No

Yes $p > 5\%$

I shall

~~We must~~ reject H_0

QUESTION. Can we conclude that H_0 is true?

The mechanics of hypothesis testing

[explained fully in IB Data Science videos & lecture notes]

1. Decide on your null hypothesis, H_0

2. Choose a test statistic t,
   e.g. "t = average difference between group A and group B"

3. Assuming H_0 to be true, what distribution would I expect to see for t?

The p-value is defined to be $p = \mathbb{P}(t \text{ as extreme or more so than } t_{\text{real}} \mid H_0)$

the value of t that we actually saw

Expected histogram of t, if H_0 were true

Multiple testing

Four metrics

#tests = 112

Eight algorithms

	R-1	R-2	R-L	R-SU4
Abstract generation from propositions				
OurAbs (A)	0.364	0.088	0.340	0.131
Sentence extraction with compression				
X + Cl	0.361	0.090	0.335	0.132
X + Co	0.340	0.074	0.321	0.113
L + Cl	0.356	0.077	0.325	0.126
L + Co	0.336	0.067	0.314	0.110
Sentence extraction				
OurExt (X)	0.376	0.122	0.345	0.154
LexRank (L)	0.349	0.087	0.316	0.129
Token extraction for propositions				
OurTok (T)	0.356	0.088	0.336	0.130

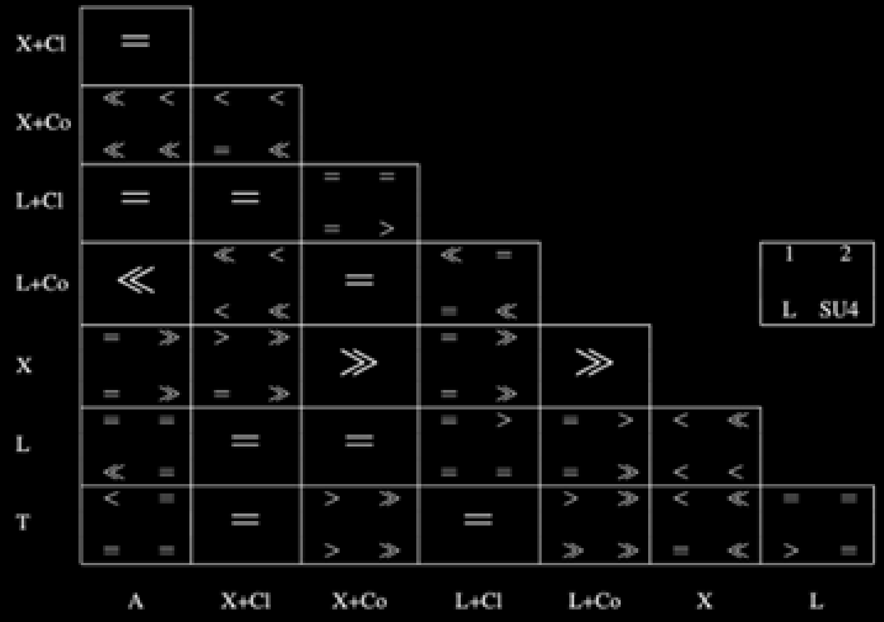


Table 2: ROUGE F-scores and statistical significance of the differences. The four positions in the significance table correspond to ROUGE-1, 2, L and SU4, respectively. “≫” means row statistically outperforms column at $p < 0.01$ significance level; “>” at $p < 0.05$ significance level, and “=” means no statistical difference detected.

Can I do multiple tests, for example on multiple outcomes?

It depends. Why are you doing hypothesis tests in the first place? Exploratory, or rhetorical?

EXPLORATORY

"I want to develop the best model I can for my dataset"

- A hypothesis test is how I ask "Is my current model good enough to explain my dataset?"
- I'll **try lots of tests**, to discover any area where I need to improve my modelling

RHETORICAL

"I want to present a hypothetico-deductive conclusion to my audience"

- **There should be one p-value to quantify a conclusion**
- If there are multiple tests then (to avoid cherry-picking) one should present a single overall p-value, and

$$p_{\text{overall}} \leq \#\text{tests} \times \min_{i \in \text{tests}} p_i$$

QUESTION. Which of these is a correct interpretation of the p-value?

1. "The probability that H_0 is true is p."

Stupid! There is no "probability of a hypothesis".

2. "I tested H_0 against an alternative, H_1. Since $p <$ MAGIC_CONST, we should accept H_1."

Stupid! "There is no alternative" fallacy.

No! It's not logic, it's policy that we reject H_0.

3. "Since $p <$ MAGIC_CONST we should reject H_0."

4. "Since $p <$ MAGIC_CONST I shall reject H_0."

This is reasonable; but I still prefer (5) because it doesn't have the arbitrary policy choice.

5. "The chance of seeing data as extreme as what I saw, assuming H_0, is p."

☺

Attendance question

What question strikes fear into the heart of a simple-minded experimentalist?

[illegible]

And if they're bold enough to answer you, follow it up with

"Have you corrected for multiple testing?"

Choosing or designing a test that suits your data

- Whatever we want to conclude, we have to dress it up as "reject the null hypothesis" for some null hypothesis H_0

QUESTION. What might you conclude by rejecting this H_0?

Data might not be Gaussian.
The means might be different.
The variances might be different.
They might not be independent.

SubjectID	Device	HitRate
1	touchpad	0.939
2	touchpad	0.975
3	button	0.940
4	button	1.000
5	button	0.915
⋮	⋮	⋮

H_0: the readings from both group A and group B are all independent Gaussian random variables with mean μ and variance σ^2 for some μ, σ

This is the null hypothesis that is tested by the standard t-test

- Whatever we want to conclude, we have to dress it up as "reject the null hypothesis" for some null hypothesis H_0
- If our audience considers our H_0 to be non-credible *a priori*, we won't achieve anything by rejecting it

SubjectID	Device	HitRate
1	touchpad	0.939
	button	0.975
2	touchpad	0.940
	button	1.000
3	touchpad	0.915
⋮	⋮	⋮

"The touchpad and button groups have significantly different HitRate (t-test, $p = 0.020$)."

QUESTION. Is the implied H_0 credible?

QUESTION. What's a credible test?

The paired t-test

SubjectID	button	touchpad	difference
1	0.975	0.939	+0.036
2	1.000	0.940	+0.060
3	0.905	0.915	-0.010
⋮	⋮	⋮	⋮

Null hypothesis: the within-subject differences are independent Normal$(0, \sigma^2)$ for some σ (hence there's no difference between button and touchpad)

Test statistic: let t be the average of within-subject differences

There are competing goals in choosing a test

Choose a detailed and explicit H_0 that models everything about my dataset, to make full use of my data

Choose a simple H_0 so that my audience is more likely to accept it

The sign test (doesn't assume Gaussian distributions)

TrialID	Alg1 score	Alg2 score	Which Better
1	78.5	93.2	Alg2
2	33.4	25.8	Alg1
3	65.0	64.1	Alg1
4	57.5	58.3	Alg2
5	57.6	93.2	Alg2
⋮	⋮	⋮	⋮

Null hypothesis: the two algorithms are equally as good.

Test statistic: let t be the number of trials in which Alg1 does better (out of n).

The distribution of t under H_0 is simply $\mathrm{Bin}(n, {}^1\!/_2)$.

A permutation test (doesn't assume Gaussian distributions)

PatientID	Treatment	Outcome
1	placebo	93.2
2	drug	25.8
3	drug	64.1
4	drug	58.3
5	placebo	44.2
⋮	⋮	⋮

Imagine that the office that prepared the treatment allocation list had used a different random number seed.

If H_0 is true, it'd make no difference to the outcome.

Null hypothesis: the drug has no effect

Test statistic: let t be the difference in average outcome between the two treatments

To find the distribution of t under H_0, we simply simulate many permutations of Treatment.

The messy case

To make full use of a rich dataset, we generally have to propose a detailed and explicit probability model for our H_0.

carry-over?

covariates

repeated measures

panel data

SubjectID	Age	Gender	Trial	Condition	HitRate1	HitRate2
1	23	female	1	touchpad	0.939	0.950
			2	armtap	0.914	1.000
			3	button	1.000	0.965
2	22	male	1	armtap	0.988	0.931
			2	touchpad	0.975	0.947
⋮			⋮	⋮	⋮	⋮

We introduce elaTCSF with a spatial probability summation model, which accounts for eccentricity, luminance, and area, extending the industry flicker detection standard $TCSF_{IDMS}$. We also address past controversies (a 120-year-old debate) regarding parafovea sensitivity peak.

Our elaTCSF model is fitted to and tested against 6 different datasets with both Critical Flicker Frequency (CFF) and sensitivity measurements, one of which we collected specifically to address the prominent issue of flicker in VRR displays.

elaTCSF is built on established psychophysical models, such as Watson's TCSF, or the spatial probability summation. This choice was made to avoid overfitting given the sparsity of available psychophysical data. Even if a better fit can be found with a polynomial function or a neural network, such a function is unlikely to generalize to the conditions outside the training dataset. The model comparisons are listed in Table 3.1.

The data comes from multiple datasets, each probably having multiple readings per subject. Uh oh!

FIRST DESCRIBE YOUR DATA

- List the columns
- Explain the rows
  - How are they grouped?
  - Who is included? Who is omitted?

THEN SAY HOW YOU'RE ANALYZING IT

- When fitting models or testing hypotheses, use methods that account for the grouping
  - keywords: "panel data", "repeated measures", "meta-analysis"

Where to go for help: