Quantitative methods for small data

RSP unit OU28

DAMON WISCHIK

Who’s still working with small data?

HCI, social science, medicine

- Small number of human subjects

Natural language processing (NLP)

- Small number of corpora

Causal machine learning (fit a model across data from multiple domains)

- Small number of domains

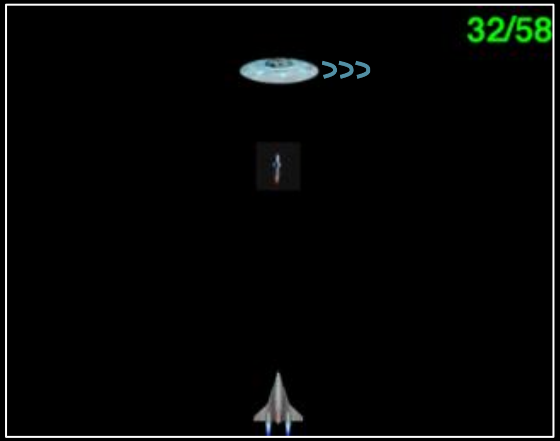
A typical small-data HCI experiment

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)



response /
outcome metric /
dependent variable

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

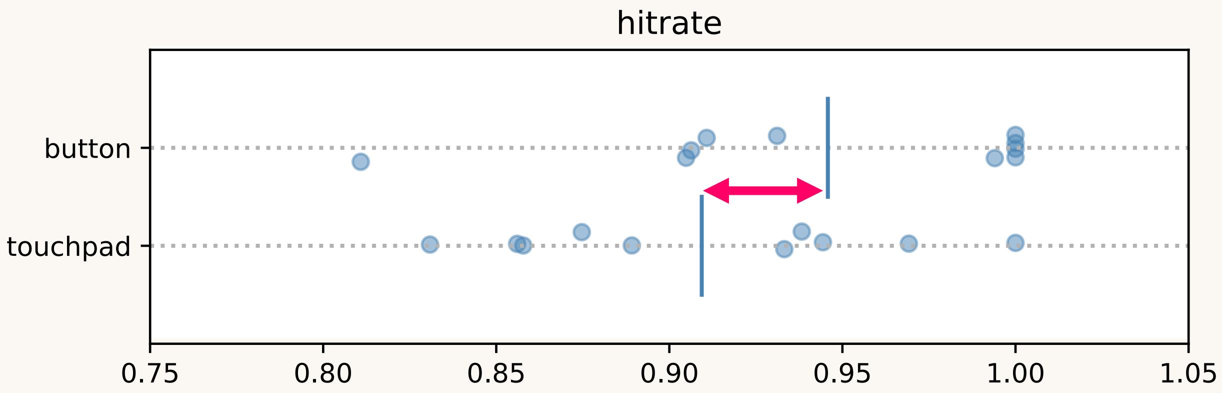
condition /
independent variable

experimental unit

We want to learn

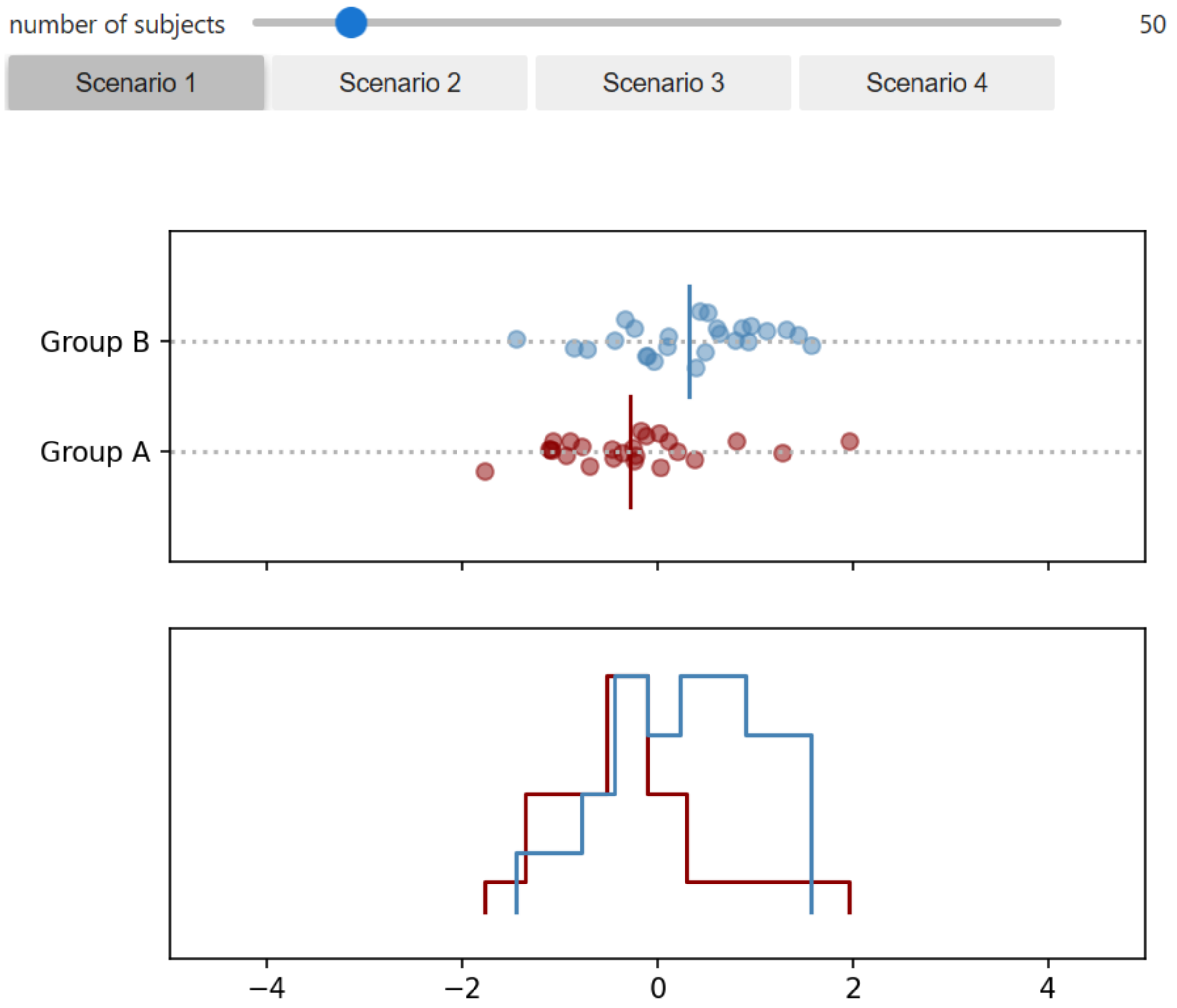
“How does the response depend on the condition?”

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?

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



number of subjects 50

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!

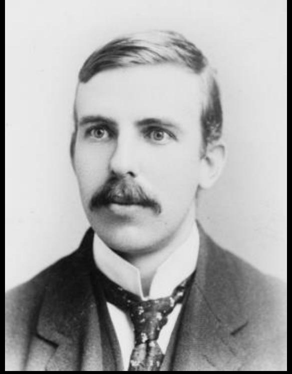
(With only two groups I think it’s more helpful to report a confidence interval for the difference, rather than a *p*-value.)

The conceptual foundation of hypothesis testing

or

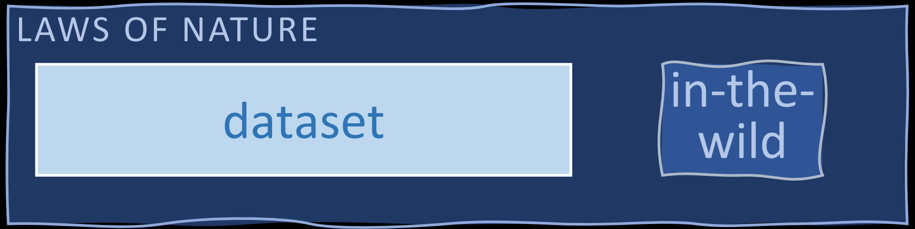
what type of statement am I making when I report a *p*-value?

GENERALIZATION



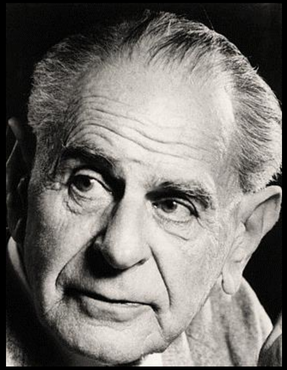
> “All science is either physics or stamp-collecting.”

Ernest Rutherford (1871–1937)



I gathered a dataset and I modelled it.
What can I usefully say about future data?
i.e. about Nature?

FALSIFICATION

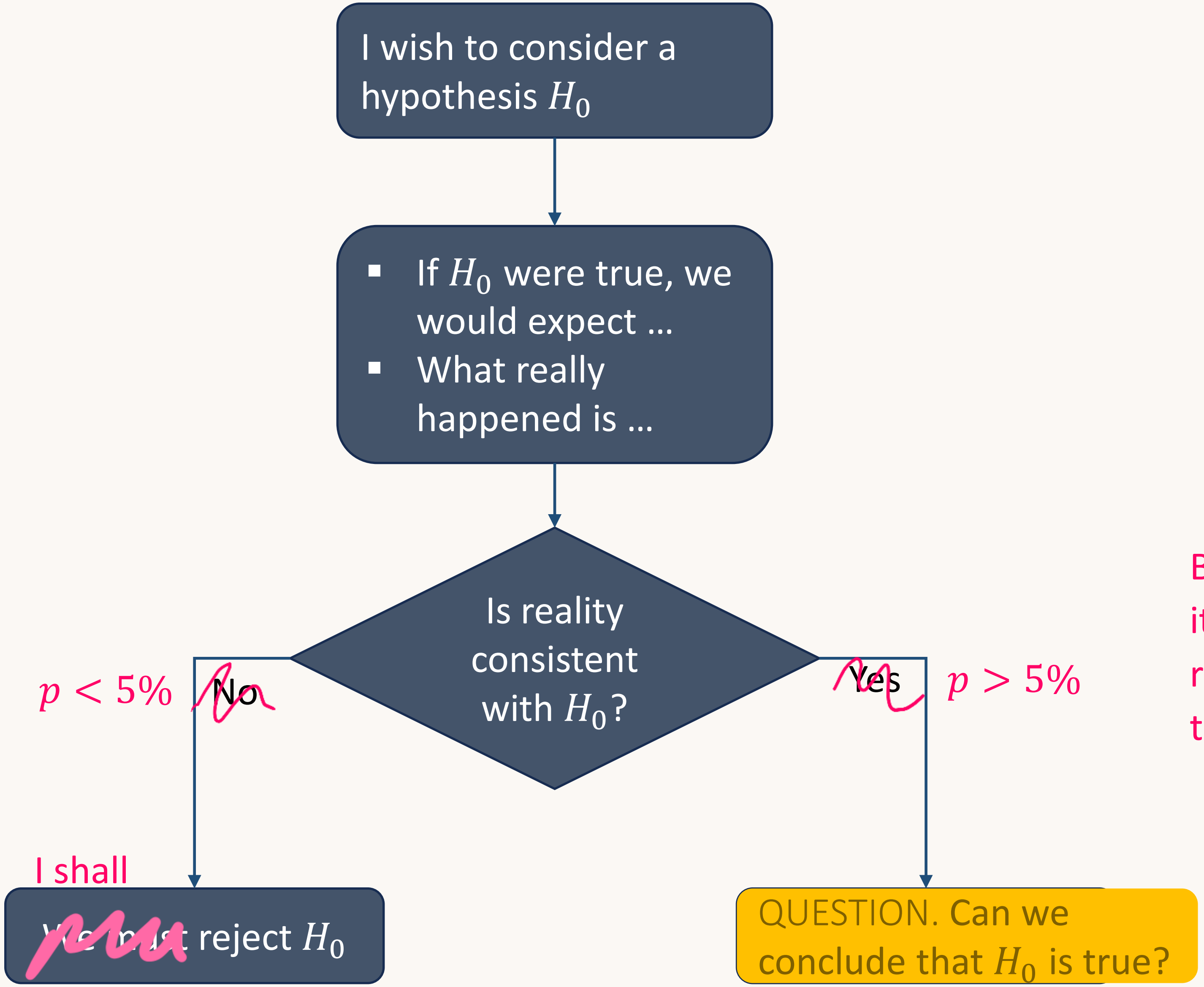


> “Every genuine scientific theory must be falsifiable.”

Karl Popper (1902–1994)

- Scientists propose models for Nature a.k.a. hypotheses
- Data may make us reject a model, but it cannot prove a model true

Popper's hypothetico-deductive approach



I wish to consider a hypothesis H_0

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

Is reality consistent with H_0?

$p < 5\%$ ~~No~~

$p > 5\%$ ~~Yes~~

I shall ~~We must~~ reject H_0

QUESTION. Can we conclude that H_0 is true?

Because of noise, it's not yes/no, it's a question of *how* consistent reality is with H_0. We measure this with the p-value.

The mechanics of hypothesis testing

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

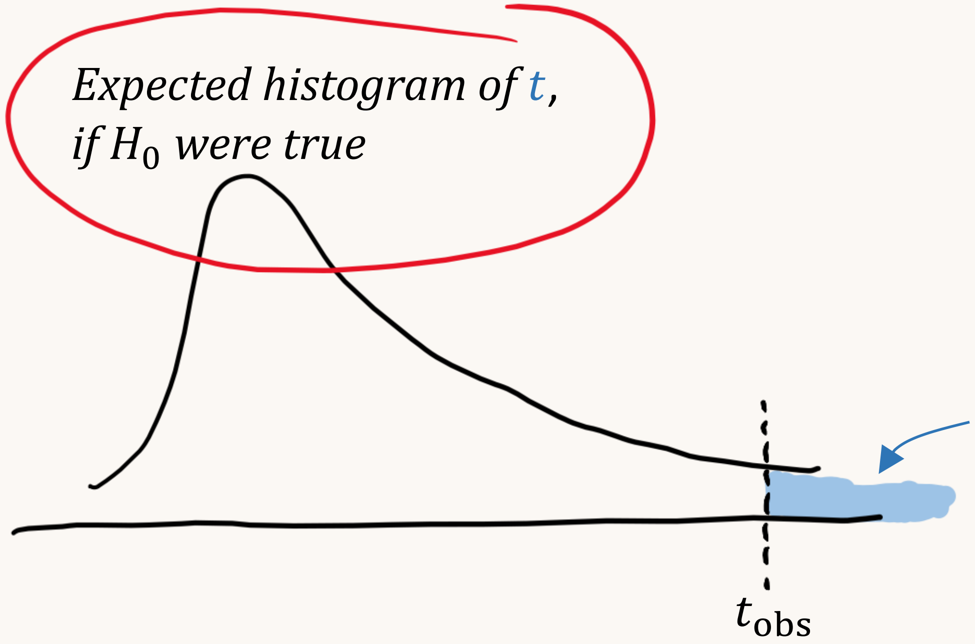
1. Decide on your null hypothesis, H_0

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 $p = \mathbb{P}(t \text{ as extreme or more so than } t_{\text{obs}} \mid H_0)$

the value of t that we actually saw



Expected histogram of t, if H_0 were true

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

1. “If $p \approx 0$ then H_0 is super-duper unlikely, and if $p \approx 1$ then it’s likely.” ✗

p measures the likelihood of the data, not of the hypothesis.

There is no such thing as “probability that a hypothesis is true”

Even if $p \approx 1$, H_0 may be false.

e.g. very small amount of data.

2. “The p-value lets me select between models. I’ll test H_0 against an alternative, H_1. Since $p <$ MAGIC_CONST, H_1 is better.” ✗

The p-value is only for evaluating a SINGLE hypothesis.

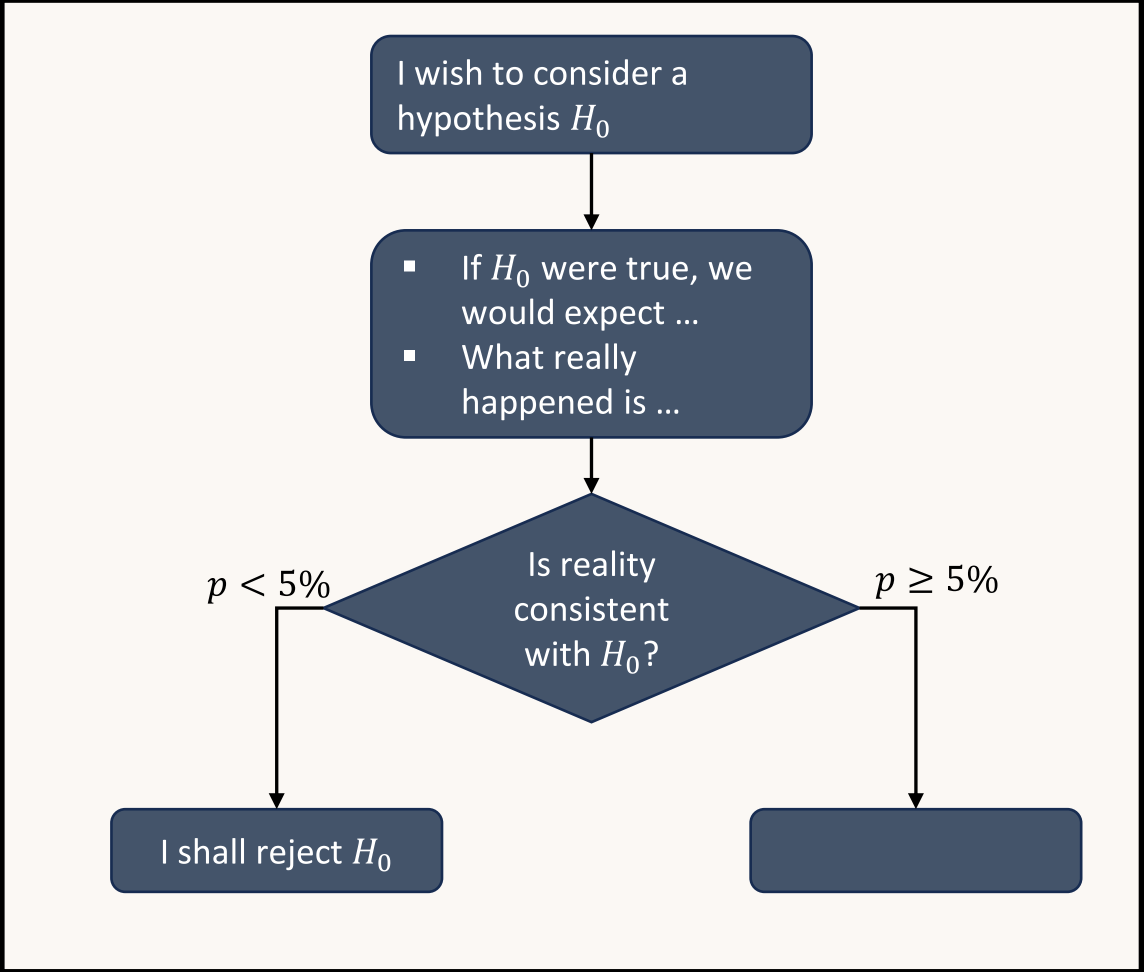
Holdout set evaluation is THE best way to choose between models.

In small-data world, use AIC ≡ leave-one-out cross-validation.

3. “Since $p <$ MAGIC_CONST ~~we should~~ reject H_0.”

I shall

What makes a good hypothesis test?


I wish to consider a hypothesis H_0
- If H_0 were true, we would expect …
- What really happened is …
Is reality consistent with H_0?
$p < 5\%$
$p \geq 5\%$
I shall reject H_0

RHETORICAL ANALYSIS

We only get a definite publishable conclusion if we reject H_0.

Anything we want to argue, we have to phrase it as “reject H_0” for a suitable H_0.

Our H_0 should match the research question we want to answer, and not bring in contentious subclaims.

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

Composite hypothesis

H_0: the readings from both groups are all independent Gaussian random variables with identical mean and variance

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

means not identical

OR variances not identical

OR distributions might not be Gaussian

OR readings might not be independent.

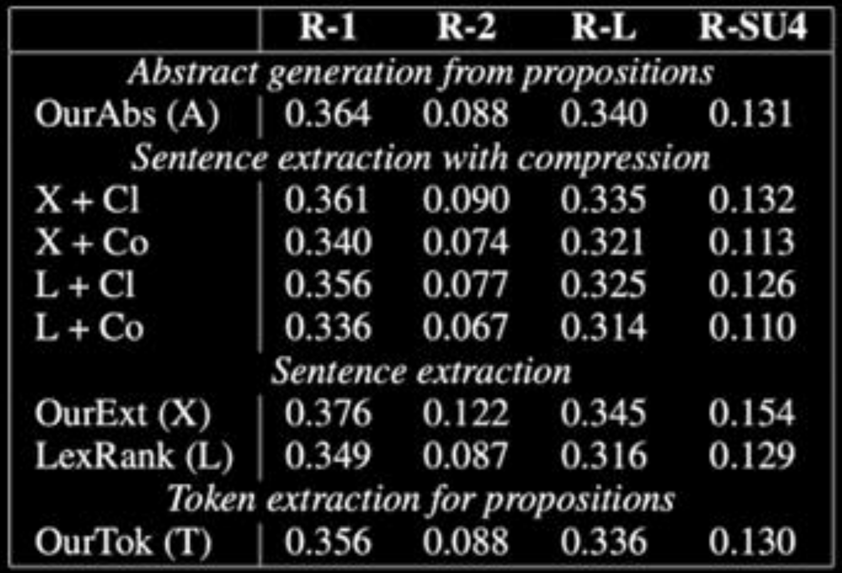
QUESTION. What might you conclude by rejecting this H_0?

Multiple testing

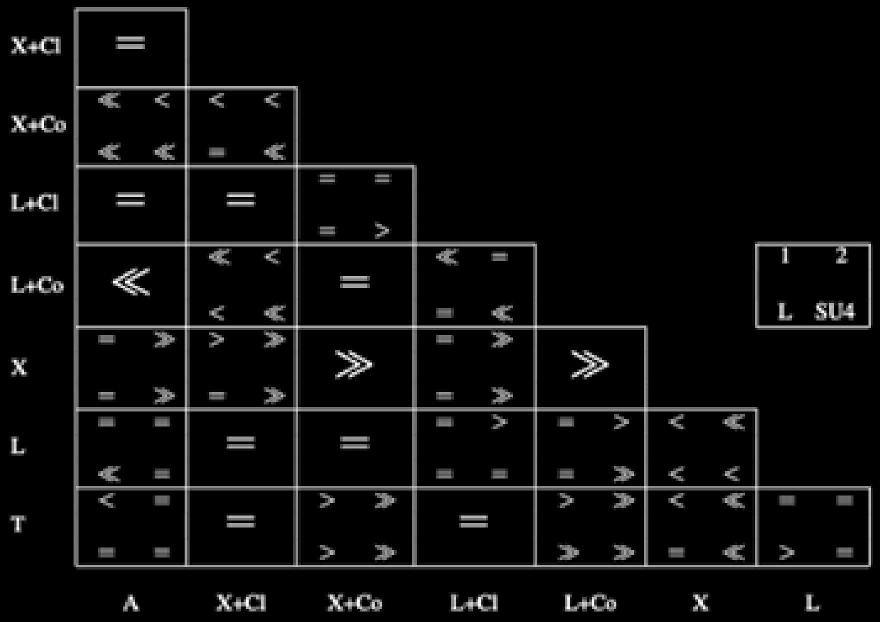
Four metrics

#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



	A	X+Cl	X+Co	L+Cl	L+Co	X	L
X+Cl	=						
X+Co	≪ < ≪ ≪	< < = ≪					
L+Cl	=	=	= = = >				
L+Co	≪	≪ < < ≪	=	≪ = = ≪			
X	= ≫ = ≫	> ≫ = ≫	≫	= ≫ = ≫	≫		
L	= = ≪ =	=	=	= > = =	= > = ≫	< ≪ < <	
T	< = = =	=	> ≫ > ≫	=	> ≫ ≫ ≫	< ≪ = ≪	= = > =

1	2
L	SU4

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.

QUESTION. What H_0 do you think the authors have in mind?

H_0: "All models are equally good across all 4 metrics"

p-value $\le$ # tests × $\min_i p_i$ ← min p-value across all tests.

Attendance question

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

[redacted]

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

"Have you corrected for multiple testing?"

Our H_0 should be credible to our audience. If we propose a non-credible H_0 and then reject it — who cares?

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. What's the implied H_0, and is it credible?

Not credible. H_0 includes "readings are independent" and no one will believe that.

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

I want to test if the results are higher with the button than with the touchpad.

How can I account for the grouping structure in my dataset?

There’s an art to designing tests that make minimal assumptions.
Such tests are highly credible.
But they often involve condensing the data.

SubjectID	button	touchpad	difference	1[button better]
1	0.975	0.939	+0.036	1
2	1.000	0.940	+0.060	1
3	0.905	0.915	-0.010	0
⋮	⋮	⋮	⋮	
n subjects				

PAIRED t-TEST

$\boldsymbol{H_0}$: the within-subject differences are independent Normal$(0, \sigma^2)$ for some σ

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

[If H_0 is true then $t \sim N(0, \hat{\sigma}^2/n)$ and we can calculate the p-value on this basis.]

SIGN TEST

$\boldsymbol{H_0}$: the two devices are equally as good

Test statistic: let t be the number of trials in which button is better

[If H_0 is true then $t \sim \mathrm{Bin}(n, {}^1\!/_2)$ and we can calculate the p-value on this basis.]

Grouped data

To make full use of a rich dataset, we typically have to propose an “anti-minimal” detailed probability model for H_0 that incorporates all the grouping structure.

And the covariates too.

When you describe your data and tests, be very clear about the grouping structure. It has a huge impact on the analysis.

carry-over?

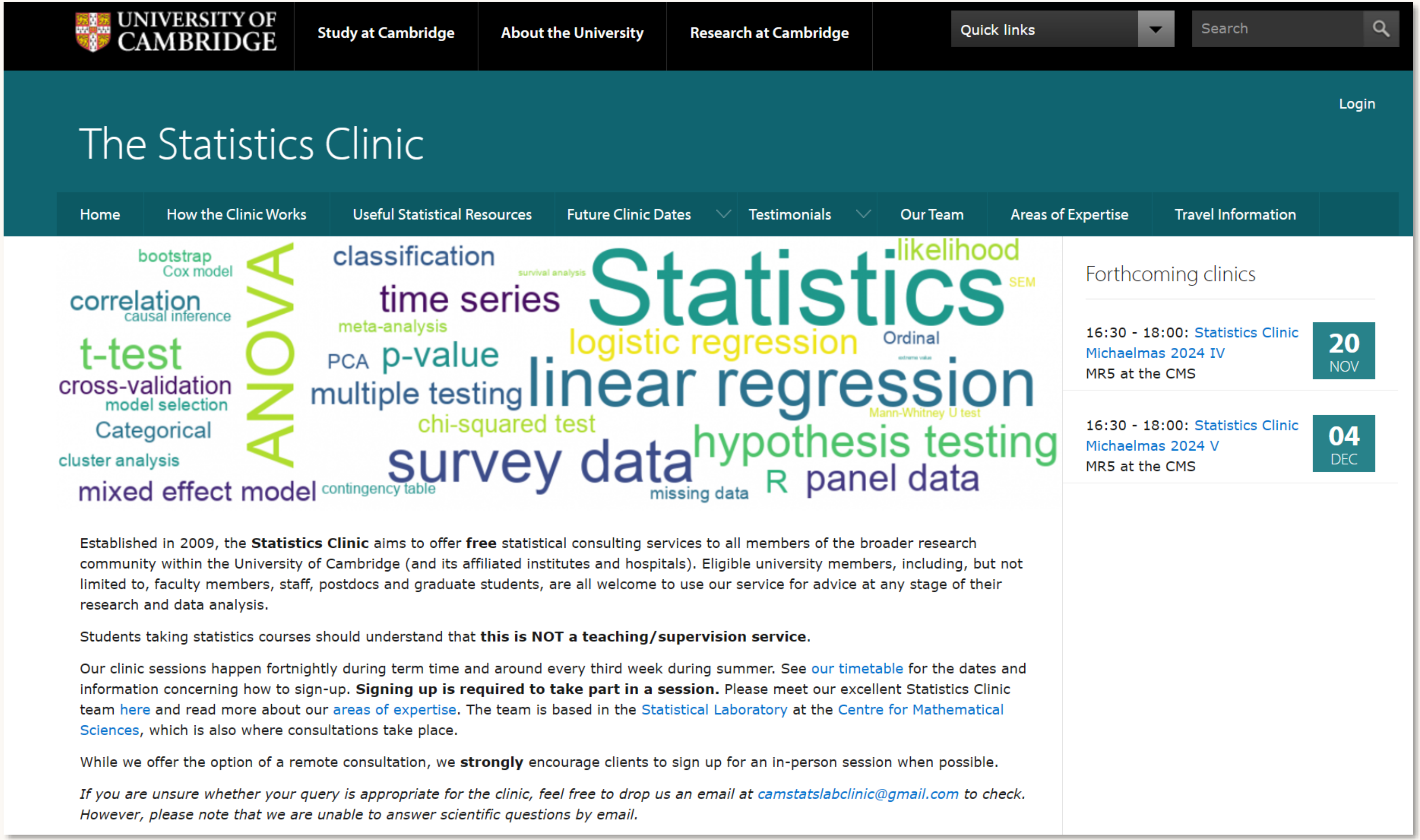
covariates

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
⋮			⋮	⋮	⋮	⋮

Where to go for help:



Forthcoming clinics

16:30 - 18:00: Statistics Clinic Michaelmas 2024 IV
MR5 at the CMS
20 NOV

16:30 - 18:00: Statistics Clinic Michaelmas 2024 V
MR5 at the CMS
04 DEC

Established in 2009, the **Statistics Clinic** aims to offer **free** statistical consulting services to all members of the broader research community within the University of Cambridge (and its affiliated institutes and hospitals). Eligible university members, including, but not limited to, faculty members, staff, postdocs and graduate students, are all welcome to use our service for advice at any stage of their research and data analysis.

Students taking statistics courses should understand that **this is NOT a teaching/supervision service**.

Our clinic sessions happen fortnightly during term time and around every third week during summer. See our timetable for the dates and information concerning how to sign-up. **Signing up is required to take part in a session.** Please meet our excellent Statistics Clinic team here and read more about our areas of expertise. The team is based in the Statistical Laboratory at the Centre for Mathematical Sciences, which is also where consultations take place.

While we offer the option of a remote consultation, we **strongly** encourage clients to sign up for an in-person session when possible.

If you are unsure whether your query is appropriate for the clinic, feel free to drop us an email at camstatslabclinic@gmail.com to check. However, please note that we are unable to answer scientific questions by email.