Quantitative methods for small data

DAMON WISCHIK

RSP unit OU28
Reference: lecture notes for IB Data Science

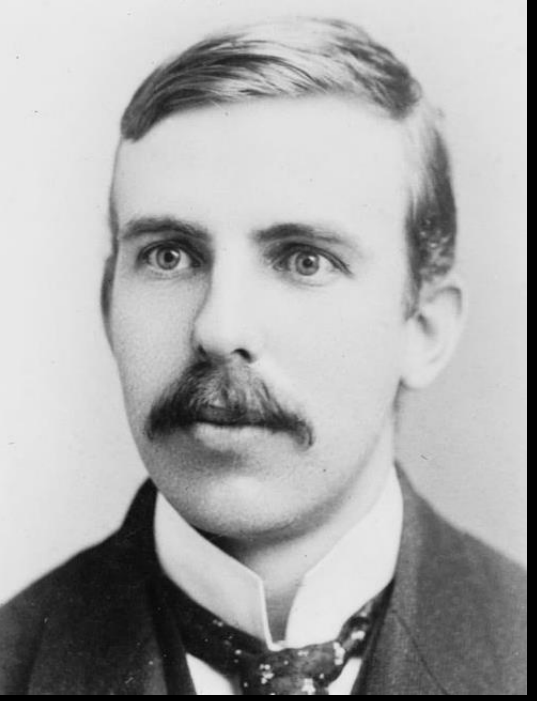
Who’s still working with small data?

HCI, social science, medicine

- Small number of human subjects
- “Does my experimental intervention affect the outcome?”

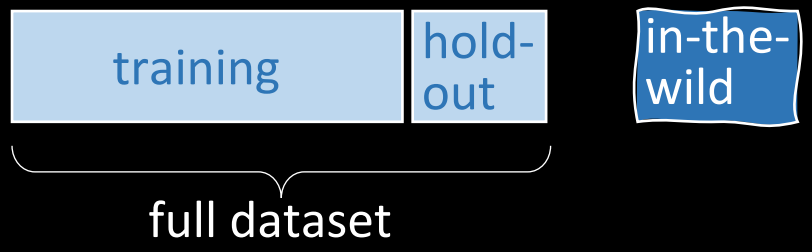
NLP

- Small number of corpora
- “Is my new algorithm better than the state-of-the-art?”



"All science is either physics or stamp-collecting."

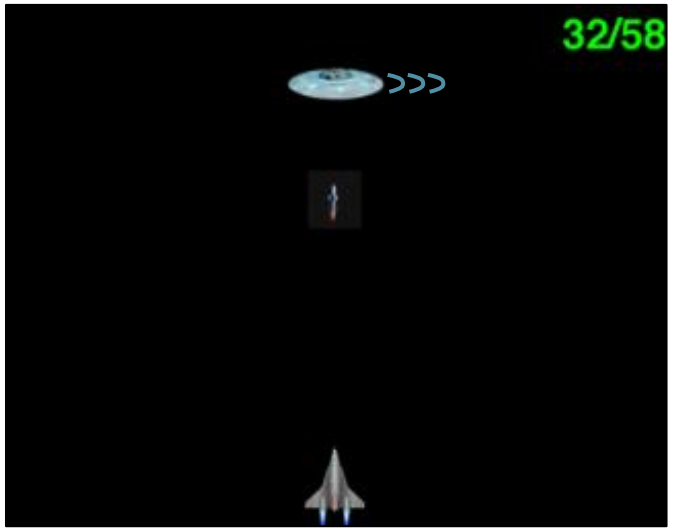
Ernest Rutherford (1871-1937)



Will my conclusion still hold for in-the-wild data? The best way to test this is to see if it holds across a variety of different corpora.

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. They have one shot per UFO. Each UFO travels at a constant speed, though the speed varies from UFO to UFO. Each game lasts 3 minutes.

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



The easy case

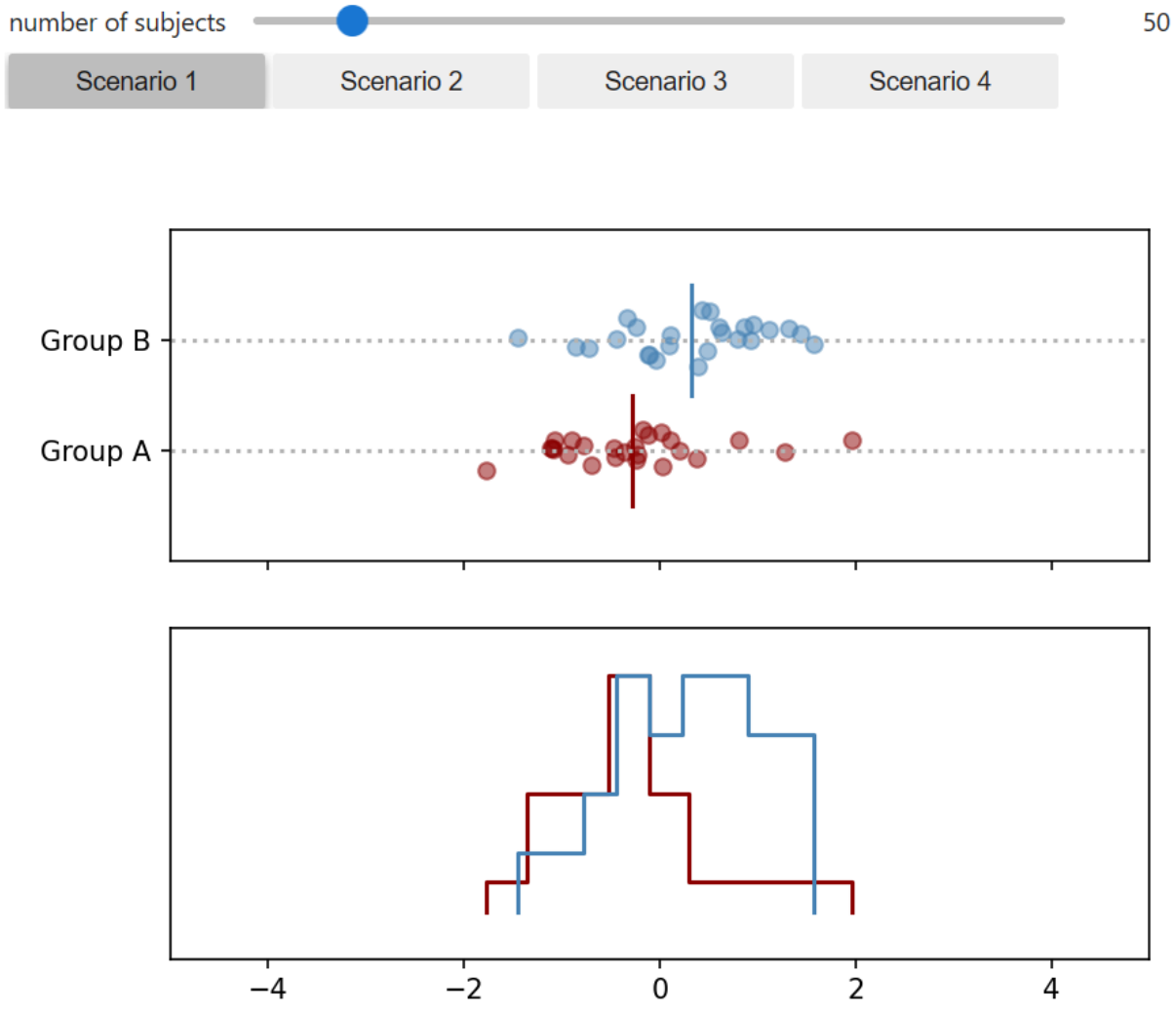
SubjectID	Condition	HitRate
1	touchpad	0.939
2	touchpad	0.975
3	button	0.940
4	button	1.000
5	button	0.915
⋮	⋮	⋮

numerical outcome measure

two test conditions

each subject experiences one test condition

Is there a difference between the outcomes for two groups?



The easy case

numerical outcome measure

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

two test conditions

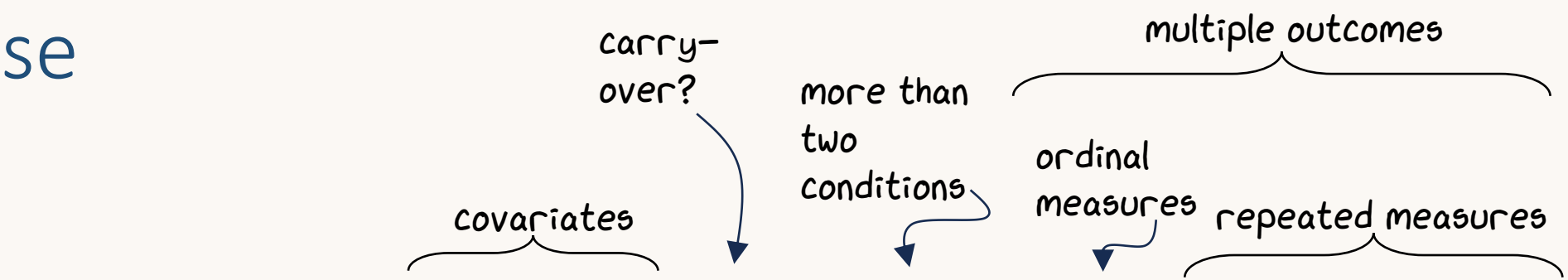
each subject experiences one test condition

"The two groups have significantly different HitRate (t-test, $p = 0.029$)."

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

- The t-test is only appropriate if the outcome is Gaussian
- With two groups, it's more informative to report a confidence interval

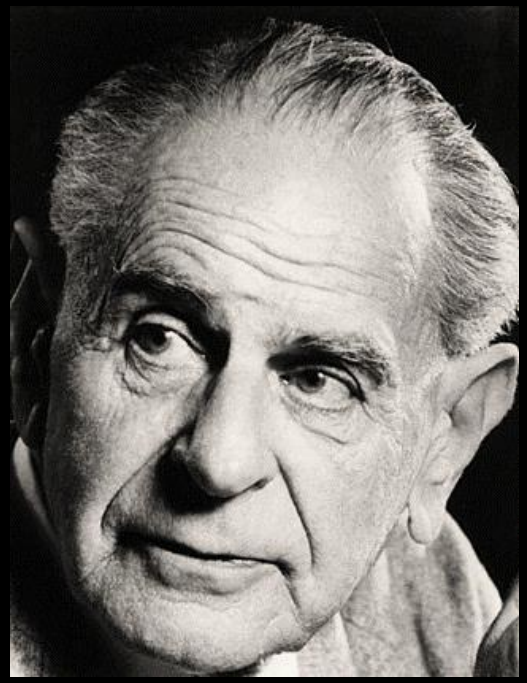
The tricky case



panel data

SubjectID	Age	Gender	Trial	Condition	FeelsLike Body	HitRate1	HitRate2
1	23	female	1	touchpad	weak agree	0.939	0.950
			2	armtap	strong agree	0.914	1.000
			3	button	neutral	1.000	0.965
2	22	male	1	armtap	agree	0.988	0.931
			2	touchpad	weak disagree	0.975	0.947
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The conceptual foundation of hypothesis testing



"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."

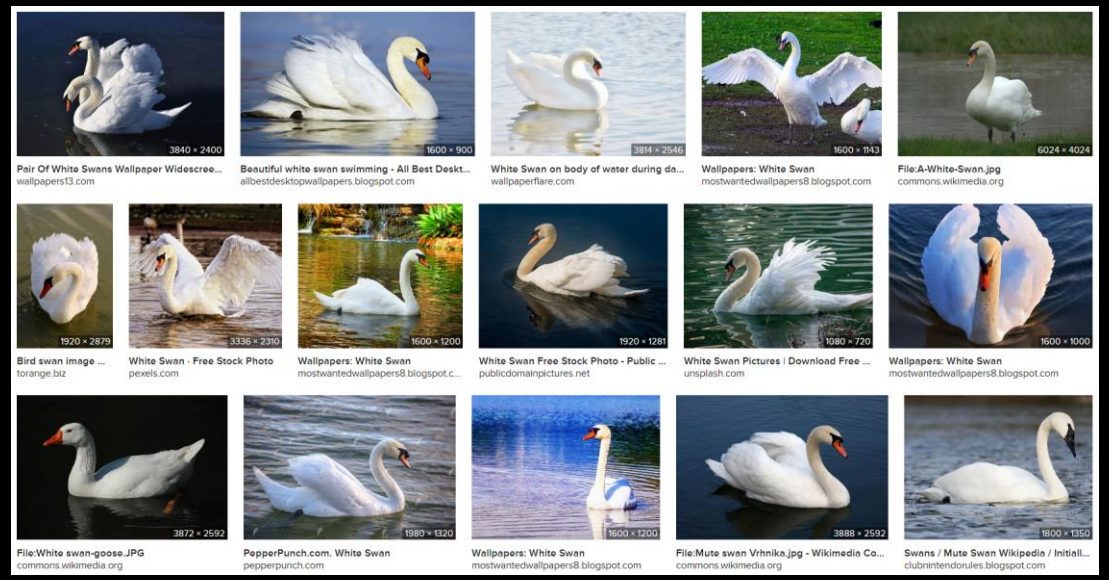
Karl Popper (1902–1994)

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

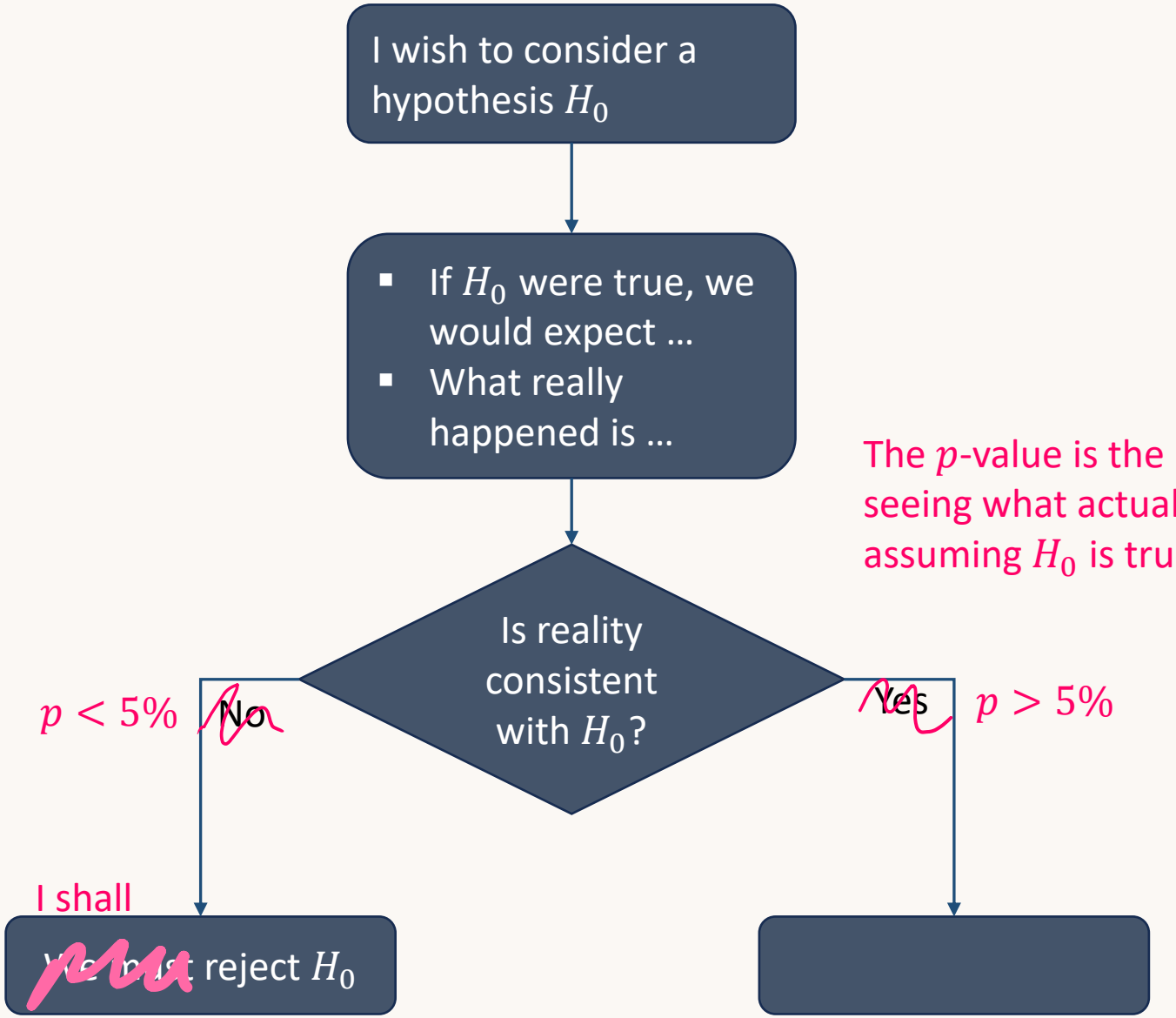
The hypothesis is logically equivalent to

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

SUPPORTING EVIDENCE

My pot plant isn’t white, and it isn’t a swan.

The hypothetico-deductive method



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

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

* And if our audience doesn't think our H_0 is credible, we won't have achieved anything!

What might you conclude by rejecting these H_0?

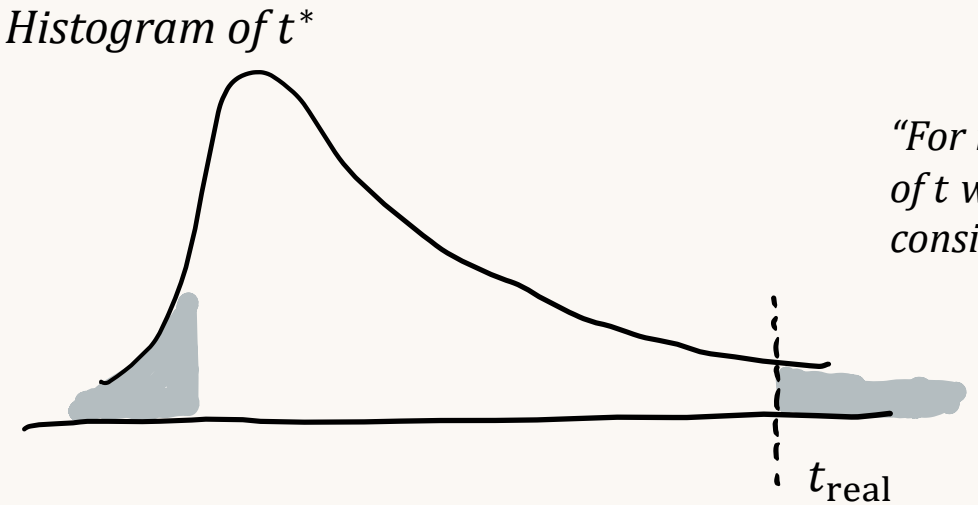
- H_0: the data from each of my two groups is $N(\mu, \sigma^2)$ for some μ, σ
- H_0: with multiple groups, the data from group g is $N(\mu, \sigma_g^2)$ for some $\mu, \{\sigma_g\}$
- H_0: the data from my single group of test subjects is $N(\mu, \sigma^2)$ for some $\mu \geq \text{thresh}$ and some σ

The mechanics of hypothesis testing

1. Choose a null hypothesis, H_0
2. Choose a test statistic t, i.e. a function $t : \text{dataset} \mapsto \mathbb{R}$
3. Assuming H_0 to be true, how might t have turned out in a parallel universe? Denote the parallel-universe value of the statistic by t^*, a random variable.

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

the value of t that we actually saw in this universe



"For my H_0, both large and small values of t would have made me doubt it, so I consider both tails to be extreme."

Choosing the right test (H_0 and t)

Choose a detailed and explicit H_0 so that it specifies precisely the distribution of t^*

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

The sign test

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 $\mathrm{Bin}(n, {}^1/_2)$.

An unpaired permutation test

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

Null hypothesis: the drug has no effect

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

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.

A paired permutation test

CorpusID	Algorithm	Outcome
1	alg1	93.2
	alg2	91.8
2	alg1	55.1
	alg2	58.3
3	alg1	33.5
	alg2	38.8
⋮	⋮	⋮

Null hypothesis: for a given CorpusID, the algorithm makes no difference to the distribution of Outcome

To find the distribution of t under H_0, we simply simulate many random swaps of Algorithm within CorpusID

If H_0 were true, we'd get the same distribution of Outcome if the Algorithm entries were randomly swapped.

A t-test (unpaired samples, pooled variance)

numerical outcome measure

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

two test conditions

each subject experiences one test condition

Null hypothesis: the scores are independent $N(\mu, \sigma^2)$ for some μ and σ, regardless of the condition

Test statistic: t is a transformed version of the average difference between the two groups, transformed so that it is insensitive to μ and to σ.

The cunning transformation means that we can write down the distribution of t^* using simple maths.

If your audience will let you get away with a full-blown probability model, great!

Null hypothesis: in the following model, the $\gamma_{\text{condition}}$ coefficients are all equal:

$$\text{HitRate} \sim \gamma_{\text{condition}} + \alpha_{\text{age}} + \beta_{\text{gender}} + A_{\text{subject}} + N(0, \sigma^2) \quad \text{where} \quad A_{\text{subject}} \sim N(0, \rho^2)$$

covariates

SubjectID	age	gender	trial	condition	game num	HitRate
1	23	female	1	touchpad	1	0.939
					2	0.950
			2	armtap	1	0.914
					2	1.000
			3	button	1	1.000
					2	0.965
2	22	male	1	armtap	1	0.988
					2	0.931
			2	touchpad	1	0.975
					2	0.947

panel data

repeated measure

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 find 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

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

A battery of significance tests

Four metrics

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

#tests = 112

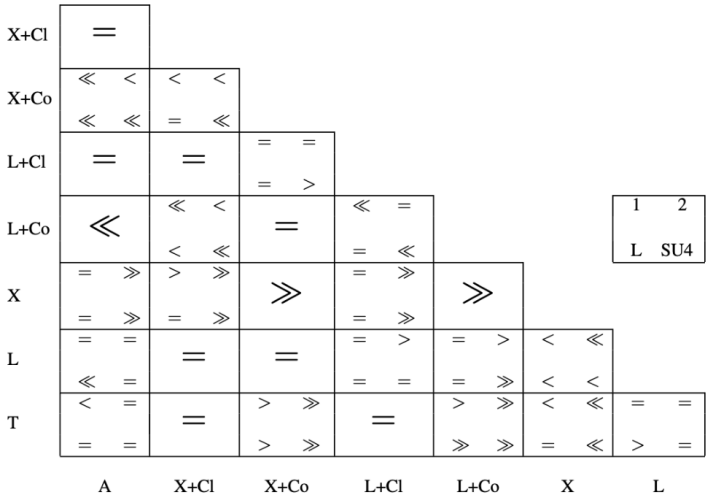


Table 2: ROUGE F-scores and statistical significance of the differences. The four positions in the significance table correspond to ROUGE-1, 2, L and SU4, respectively. "≫" means row statistically outperforms column at $p < 0.01$ significance level; ">" at $p < 0.05$ significance level, and "=" means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

The hypothesis "All models are equally good" has overall $p = 112 \times \min_i p_i$

Seeing the full battery of test results may help with exploratory model-building.

Attendance question

How do you strike fear into the heart of a simple-minded experimentalist?

What's a correct interpretation of the *p*-value?

"The probability that H_0 is true is p."

"Since $p <$ MAGIC_CONST we can reject H_0 in favour of the alternative."

"Since $p <$ MAGIC_CONST we can reject H_0."

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

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

FURTHER QUESTIONS

- Have I learnt a correlation, or a cause?
  (dependent / independent / control variables)
- Why does hypothesis testing go wrong with big data?
- ANOVA: how to test with multiple conditions
- Between-subjects versus within-subjects, and order effects
- Models for the Likert response measure