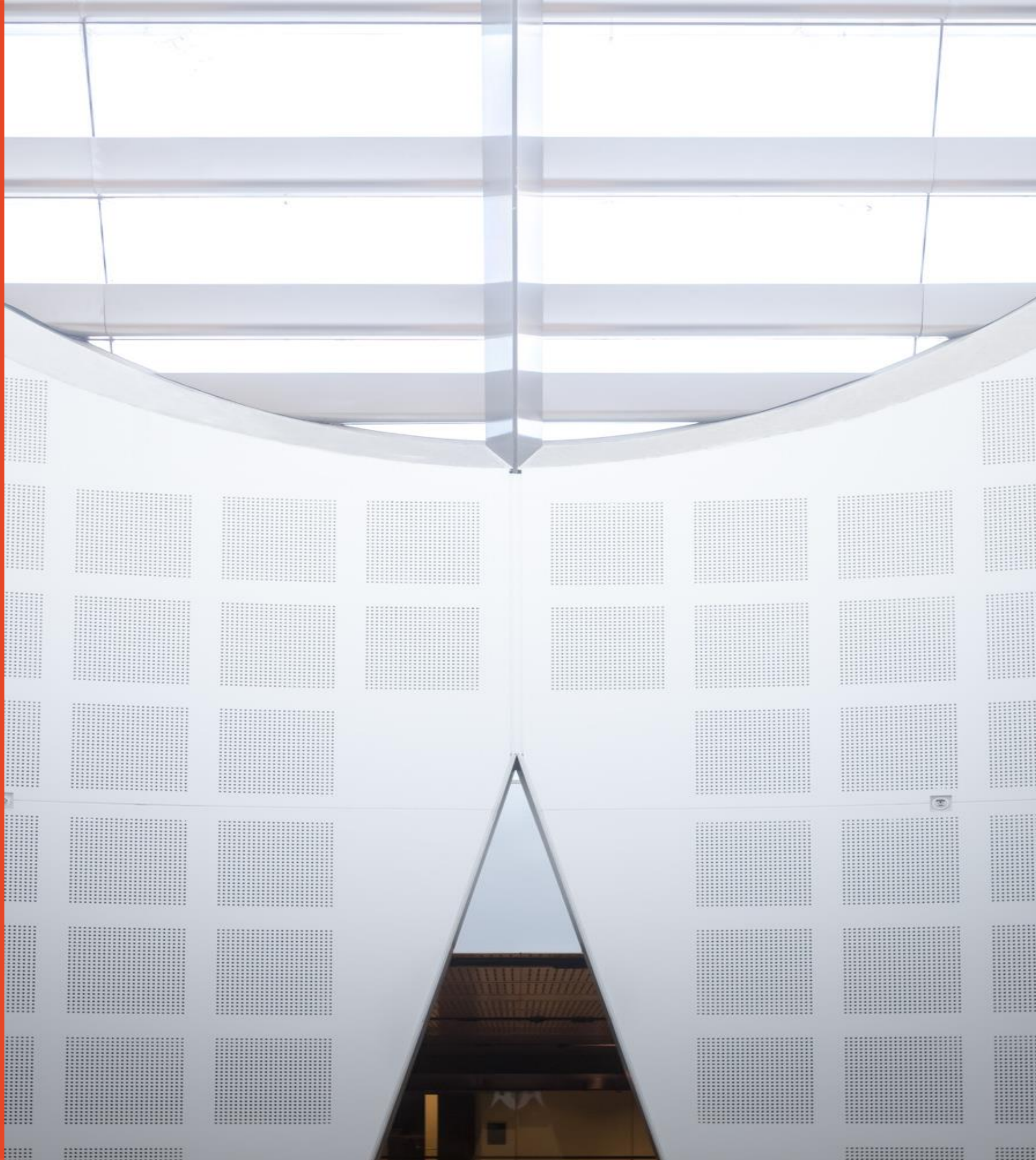
INFO5993/4990: Research Methods
Statistical Analysis & Research Evaluation

Clément Canonne

School of Computer Science

THE UNIVERSITY OF SYDNEY



Acknowledgement of Country

I would like to acknowledge the Traditional Owners of Australia and recognise their continuing connection to land, water and culture. I am currently on the land of the Gadigal people of the Eora Nation and pay my respects to their Elders, past, present and emerging.

I further acknowledge the Traditional Owners of the country on which you are on and pay respects to their Elders, past, present and future.

Goal(s) of this lecture

- Get familiar with (some) standard statistical estimators
- Learn (a bit of) how to scientifically assess the relevance of your results
- Point out some pitfalls to be wary of
- Introduce some keywords and notions you can look up later

Do we really have to?

Do we really have to?

PLOS MEDICINE

PLoS Med. 2005 Aug; 2(8): e124.
Published online 2005 Aug 30. doi: 10.1371/journal.pmed.0020124

PMCID: PMC1182327
PMID: 16060722

Why Most Published Research Findings Are False

John P. A. Ioannidis

DOI: 10.1177/1745691612465253 • Corpus ID: 26361121

Editors' Introduction to the Special Section on Replicability in Psychological Science

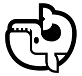
H. Pashler, E. Wagenmakers • Published 1 November 2012 • Psychology • Perspectives on Psychological Science

Is there currently a crisis of confidence in psychological science reflecting an unprecedented level of doubt among practitioners about the reliability of research findings in the field? It would certainly appear that there is. These doubts emerged and grew as a series of unhappy events unfolded in 2011: the Diederik Stapel fraud case (see Stroebe, Postmes, & Spears, 2012, this issue), the publication in a major social psychology journal of an article purporting to show evidence of extrasensory… Expand

https://en.wikipedia.org/wiki/Replication_crisis
https://en.wikipedia.org/wiki/Data_dredging

Parameter estimation

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale."

True parameter: θ ("population parameter" )

Estimator: some function T ("statistic") which takes data in, spits out a value to approximate θ

Estimate: a particular realization $\hat{\theta}$ of the estimator T(X) (data X is random, so $\hat{\theta}$ is as well)

Parameter estimation

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale."

Types of estimate:

- Point estimate: a single value $\hat{\theta}$ (best guess for θ)
- Interval estimate: a range of values (confidence interval), "θ is most likely in there"

Parameter estimation

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale."

Two points of view:*

– **Frequentist:** there is a true parameter θ, randomness comes from the observations

– **Bayesian:** θ itself is a random variable (reflects degree of certainty, evidence...)

Parameter estimation (Frequentist)

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale." 🐋

What could we want from an estimator?

- Unbiased
- Consistent
- Precise

Parameter estimation (Frequentist)

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale." 🐳

What could we want from an estimator?

- **Unbiased:** $\mathbb{E}[\hat{\theta}] = \theta$
- **Consistent:** $\hat{\theta} \to \theta$ as we get more data
- **Precise:** $\mathrm{std}(\hat{\theta})$ is small (not too much random fluctuation)

Parameter estimation (Frequentist)

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale." 🐳

Note: we cannot get all of it (e.g., bias/variance tradeoff): 😔

$$\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2 + \mathrm{std}\left(\hat{\theta}\right)^2$$

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

Parameter estimation (Frequentist)

You want to learn something about the world from observed/gathered data. E.g., "average weight of a blue whale." 🐳

Note: we cannot get all of it (e.g., bias/variance tradeoff): 😔

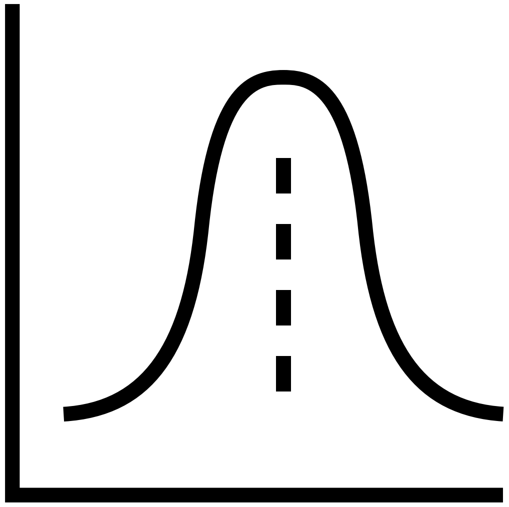
$$\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2 + \mathrm{std}\left(\hat{\theta}\right)^2$$

MSE = Bias^2 + Variance

Different estimators for the same quantity might be good for different things

Parameter estimation (Frequentist)

Example: assume the weight of a whale is Gaussian, with parameter θ: $\mathcal{N}(\theta, 1)$



θ

Parameter estimation (Bayesian)

Start with a prior, observe the data, update the prior to get a posterior

E.g., “my a priori knowledge is that whales could weigh anything between 500 and 20000 kg”

Prior: $\theta \sim \text{Uniform}(500, 20000)$

Parameter estimation (Bayesian)

Start with a prior, observe the data, update the prior to get a posterior

E.g., “my a priori knowledge is that whales could weigh anything between 500 and 200000 kg” 🐳

Prior: θ ~ Uniform(500, 200000)

Then I go out at sea, find $n = 14$ whales, measure their weight (?): ⚖️

133996. | 137230. | 143330. | 141119. | 132318. | 134366. | 139451. |
130268. | 146805. | 150000. | 143923. | 131957. | 148290. | 146080. |

What is my new state of knowledge/belief about whales?

Parameter estimation (Bayesian)

Start with a prior, observe the data, update the prior to get a posterior

Bayesian update (Bayes' rule).

Posterior: $\theta \sim \text{SomethingElse}(100000, 150000)$

Gives a probability distribution.

Now, what's a Confidence Interval?

Wikipedia:

In frequentist statistics, a **confidence interval** (CI) is a range of estimates for an unknown parameter, defined as an interval with a lower bound and an upper bound […].The interval is computed at a designated confidence level. The 95% confidence level is most common, but other levels (such as 90% or 99%) are sometimes used. The confidence level represents the long-run frequency of confidence intervals that contain the true value of the parameter. In other words, 95% of confidence intervals computed at the 95% confidence level contain the parameter, and likewise for other confidence levels.

Now, what's a Confidence Interval?

Wikipedia:

In frequentist statistics, a **confidence interval** (CI) is a range of estimates for an unknown parameter, defined as an interval with a lower bound and an upper bound […].The interval is computed at a designated confidence level. The 95% confidence level is most common, but other levels (such as 90% or 99%) are sometimes used. The confidence level represents the long-run frequency of confidence intervals that contain the true value of the parameter. In other words, 95% of confidence intervals computed at the 95% confidence level contain the parameter, and likewise for other confidence levels. 😕

Now, what's a Confidence Interval?

You want to estimate θ: you return two random variables L, U (depend on the data, so on θ) such that

$$\Pr[\, L < \theta < U \,] \geq 95\%$$

where the randomness is taken over the data (so over L, U): θ itself is not random.

Now, what's a Confidence Interval?

You want to estimate θ: you return two random variables L, U (depend on the data, so on θ) such that

$$\Pr[\ L < \theta < U\ ] \geq 95\%$$

where the randomness is taken over the data (so over L, U): θ itself is not random.

If you find it confusing: **it is**. Always be careful when interpreting things!

https://en.wikipedia.org/wiki/Confidence_interval#Misunderstandings

And now, for something completely different

Forget estimation: hypothesis testing

You have a hypothesis, want to test whether the data supports it.

Forget estimation: hypothesis testing

You have a hypothesis, want to test whether the data supports it.

Null hypothesis H_0 what people before you believe

Alternative hypothesis H_1 what you would like to show*

Statistical test: “*does the data give enough evidence against H_0 to reject it*”?

Forget estimation: hypothesis testing

You have a hypothesis, want to test whether the data supports it.

Null hypothesis H_0 what people before you believe

Alternative hypothesis H_1 what you would like to show*

Statistical test: *"does the data give enough evidence against H_0 to reject it"*?

p-value: "the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis (H_0) holds."

Forget estimation: hypothesis testing

You have a hypothesis, want to test whether the data supports it.

Null hypothesis H_0 what people before you believe

Alternative hypothesis H_1 what you would like to show*

Statistical test: *"does the data give enough evidence against H_0 to reject it"*?

p-value: "the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis (H_0) holds." 😕

Forget estimation: hypothesis testing

You have a hypothesis, want to test whether the data supports it.

Null hypothesis H_0 — what people before you believe

Alternative hypothesis H_1 — what you would like to show*

Statistical test: *"does the data give enough evidence against H_0 to reject it"*?

p-value: "the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis (H_0) holds." 😕

Again, confusing, and one of the most misinterpreted/"hacked" things in statistical analysis. Be careful!

Forget estimation: hypothesis testing

p-value: “the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis (H_0) holds.” 😕

“If Pr[what we saw | H_0] $< \alpha := 0.05$, we reject H_0“

Again, confusing, and one of the most misinterpreted/”hacked” things in statistical analysis. Be careful!

“Since the precise meaning of p-value is hard to grasp, misuse is widespread and has been a major topic in Metascience” (Wikipedia)

Some more: adaptive data analysis

(Be careful!)

Some more: adaptive data analysis

(Be careful!)

(This one-hour lecture is not enough!)

Some more: adaptive data analysis

(Be careful!)

(This one-hour lecture is not enough!)

(Ask a statistician)

Some more: adaptive data analysis

(Be careful!)

(This one-hour lecture is not enough!)

(Ask ~~a~~ two statisticians)