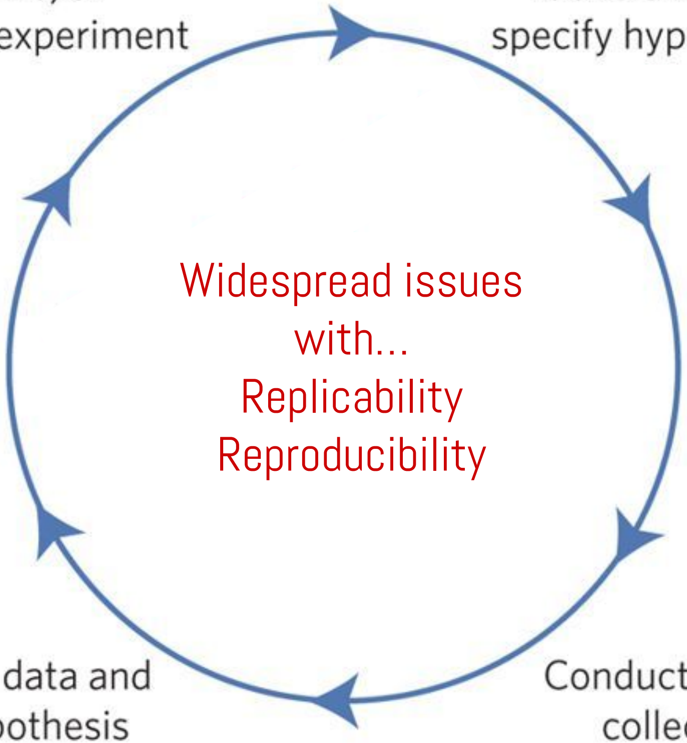
Day 09

Roundup

- Recap of main ideas & themes

What’s this course about?



Widespread issues with...
Replicability
Reproducibility

Generate and specify hypothesis

Failure to control for bias

Design study

Low statistical power
Not considering confounders

Conduct study and collect data

Poor quality control
Not recording all metadata

Analyse data and test hypothesis

No exploration
Model abuse
P-hacking

Interpret results

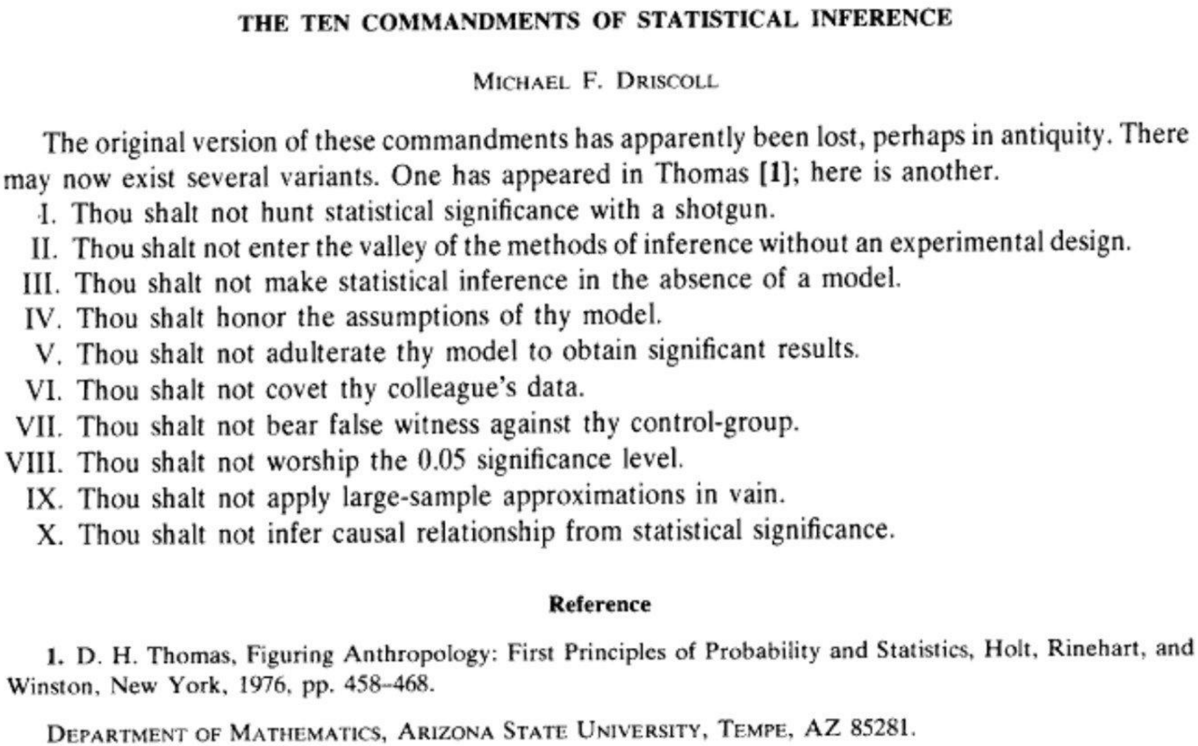
P-hacking
Ignoring multiple testing
Ignoring base rates

Publish and/or conduct next experiment

Visualization challenges
Publication bias
Lack of holistic analysis

https://www.nature.com/articles/s41562-016-0021

What’s this course about?



THE TEN COMMANDMENTS OF STATISTICAL INFERENCE

MICHAEL F. DRISCOLL

The original version of these commandments has apparently been lost, perhaps in antiquity. There may now exist several variants. One has appeared in Thomas [1]; here is another.

I. Thou shalt not hunt statistical significance with a shotgun.
II. Thou shalt not enter the valley of the methods of inference without an experimental design.
III. Thou shalt not make statistical inference in the absence of a model.
IV. Thou shalt honor the assumptions of thy model.
V. Thou shalt not adulterate thy model to obtain significant results.
VI. Thou shalt not covet thy colleague’s data.
VII. Thou shalt not bear false witness against thy control-group.
VIII. Thou shalt not worship the 0.05 significance level.
IX. Thou shalt not apply large-sample approximations in vain.
X. Thou shalt not infer causal relationship from statistical significance.

Reference

1. D. H. Thomas, Figuring Anthropology: First Principles of Probability and Statistics, Holt, Rinehart, and Winston, New York, 1976, pp. 458–468.

DEPARTMENT OF MATHEMATICS, ARIZONA STATE UNIVERSITY, TEMPE, AZ 85281.

The American Mathematical Monthly
Volume 84, Number 8, 1977 (p. 628)

Hypothesis testing, Multiple testing, P-hacking

- Collect data to disprove the hypothesis in addition to just support it. Check both expected and unexpected results.
- Check your assumptions and the assumptions of the statistical procedures.
- Remember what a p-value is and is not. ($p < 0.05 \neq 5\%$ chance the result is false)
- Be wary of selecting or discarding variables based on statistical significance.
- Avoid p-hacking or hypothesizing after the results are known.
- Control for multiple hypothesis testing, esp. excess false discoveries.
- Look beyond the p-value: effect size, other lines of evidence, prior knowledge, data quality, real world costs-and-benefits, and other explanations for the same results.

Statistical power & Sample size

- Calculate power; Be skeptical of findings from underpowered studies.
- If you can, generate pilot data to understand the variability of different factors in your system and to design a full experiment with sufficient power.
  - Published/existing publicly-available data can be really helpful here.
- If sample size is impractical, rethink your hypothesis and experimental design, and, in general, be aware of the limitations of your study.
- Not significant ≠ Zero or Nonexistent. There might not be enough power.

Pseudoreplication & Confounding factors

- Be aware of and capture biological and technical variation.
- Even when they are “different” samples, they might not be truly independent of each other.
- Record all variables and metadata (source, type/format, date/time, technician/machine) and use them to both explore data and to include in statistical analysis to detect potential confounders.
  - This is great for data management and reporting later.
- Published/existing publicly-available data can provide valuable replication.

Double-dipping & Regression to the mean

- Don't use the same data for deciding on the analysis procedure and doing the analysis itself.
  - Think about a pilot experiment. Bring in prior knowledge. Blind data-preprocessing and hypothesis generation.
- Be aware of how samples/individuals are being selected to be part of your study. The special criterion might not hold in future observations.
- Plan and decide/fix on stopping rules ahead of time. Report the rule when reporting the results.

Descriptive statistics & Visualization challenges

- Linear correlation is not appropriate for most cases. Correlation ≠ Causation.
- It is very easy to find spurious correlations/associations when testing many variables.
- Plot (different facets of) your data and overlay additional information/metadata. Visual inference is as powerful as statistical inference. Do not underestimate the power of exploratory visual data analysis.
- Even plots can be deceiving.
    - Bar plots are terrible for continuous data with small sample size. Show the actual data using dot plots and add box/violin plots for data with medium-to-large sample sizes. No pie charts or 3D either.

Planning & Registration

- Define your question specifically.
- Before collecting and analyzing data, plan your analysis and register it somewhere. After seeing the data, if you have to changed course, note this in your paper and provide an explanation.
- Don't do exploratory analysis and report just the interesting pattern. Use blinded analyses to avoid storytelling & rationalization after the fact.

Reproducibility

- Automate your analysis/visualization (avoid manual interventions & manipulations) using well-documented code.
- Keep track of all intermediate steps and results. Use R/Jupyter notebooks.
- Archive the exact versions of all external datasets (source/download-date) and code (programs/packages) used.
- Version control your entire project.
- Share raw data, tidy data, and the detailed analysis procedure including the code & recipe to perform the analysis step-by-step to reproduce all the results with a permanent identifier.
- Find out research-sponsor requirements.

Some general thoughts

- Conscious ignorance: from unknown unknown → known unknown
  - Dunning-Kruger effect: knowing that something is unknown is as hard as knowing that thing!
  - The importance of feeling stupid: threshold of learning something new!

- Intelligent persistence
  - I don't understand this → What about this don't I understand?
  - Gaps in my knowledge → Gaps in collective knowledge

Some general thoughts

- Thank you for all the discussions and active engagement!

- Keep in touch and let me know all the cool things you go on to do :)

You have already taken the right steps!

Now, go forth and apply what you've learned to your data!