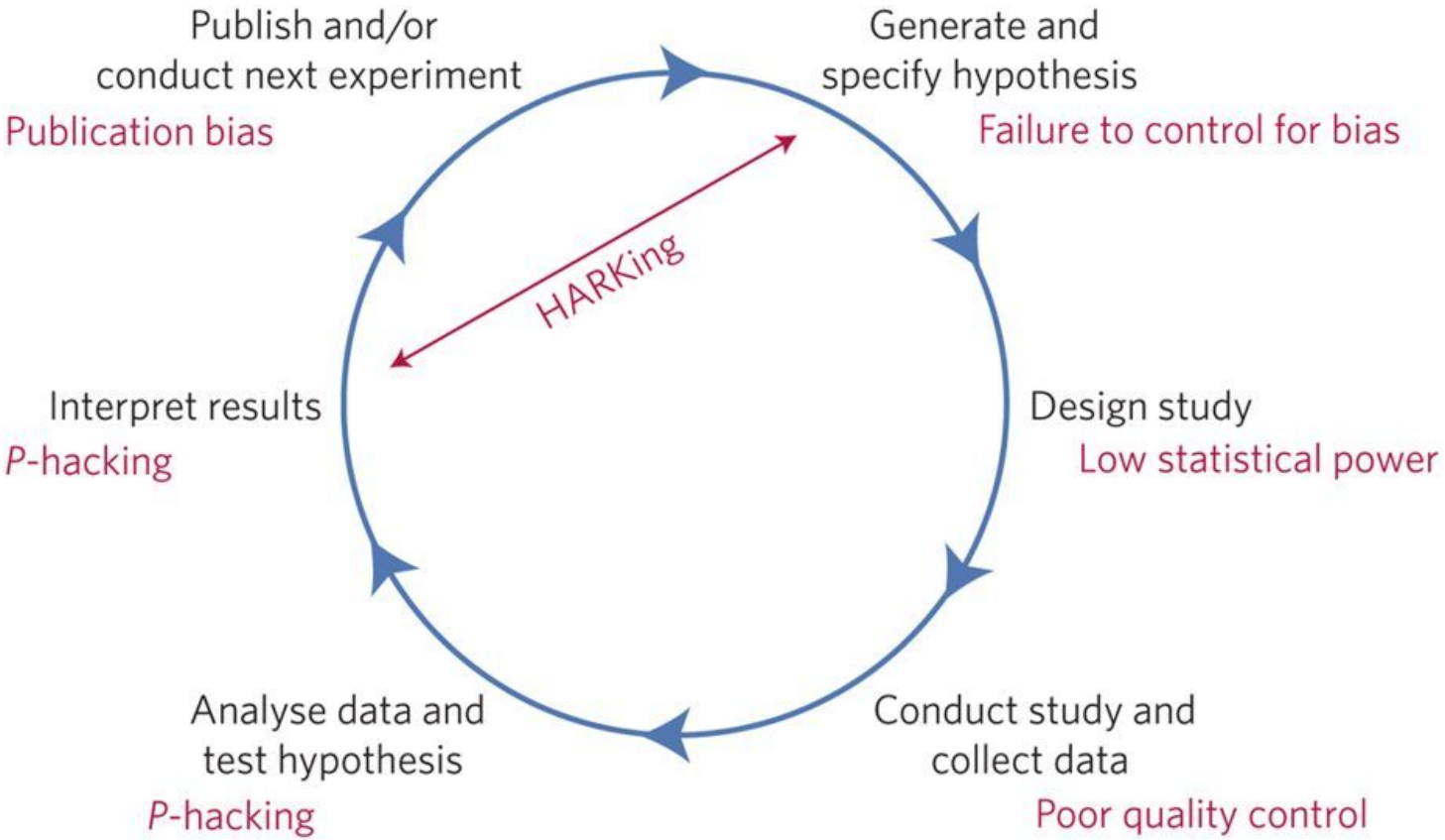
Topic 5: Review, Recap, Roundup

Lectures 10

- Review
- Recap
- Roundup

Idealized version of the scientific method & Various perils



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

Hypothesis testing, Multiple testing, P-hacking

- Collect data to disprove the hypothesis in addition to just support it. Check both expected and unexpected results.
- Remember what a p-value is and is not. ($p < 0.05 \neq 5\%$ chance the result is false)
- Control for multiple hypothesis testing, esp. excess false discoveries.
- Be wary of selecting or discarding variables based on statistical significance.
- 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 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 and use them to both explore data and to include in statistical analysis to detect potential confounders.

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.
- 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 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.
- But 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 violin plots for data with medium-to-large sample sizes. No pie charts or 3D either.

Planning, Registration, & Reproducibility

- Define your question specifically.
- Before collecting and analyzing data, plan your analysis and register it somewhere. After seeing the data, if you have to change 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.
- Automate your analysis (avoid manual interventions and manipulations) and keep track of all intermediate steps and results.
- 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.

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

Final exam

- Select a primary research article published by you/your group in the past 5-7 years that:
  - Has a specific question (exploratory or hypothesis-driven).
  - Contains experimental data (newly generated or previously published).
  - Involves a few different results obtained by a few different statistical analyses.
  - Contains a “Methods” section.
- Read this paper ***before*** the exam. You can also discuss the research question, methodology, and/or analyses with folks involved.