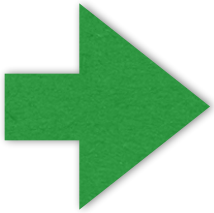
Q1) How important is the problem of adaptivity and its various guises as a cause of false discovery and false inference?

Response:

- With very few exceptions scientific practice requires an approach capable of dealing with adaptivity

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.



3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.

4. "Fishing": computing $T(y; \phi_j)$ for $j = 1, \ldots, J$: that is, performing J tests and then reporting the best result given the data, thus $T(y; \phi^{\text{best}}(y))$.

from Gelman and Loken 2013

Q2) Has the null hypothesis significance test approach meet the end of its lifespan?

Q3) What measure of statistical significance could replace them?

Response:

- Conceptually, NHT has been dead for quite a while.
- Bayesian statistics logic and methods should be the norm.

- In practice, NHT will see a painfully slow death:
  — NIH grant review
  — Statistically clueless editorial process
  — Inadequate statistical training

Q4) How do you deal with this in your own work?

Q5) Given the problems raised by adaptivity (e.g., Freedman's Paradox), how do we assess the statistical validity of post-selection inference?

Ideal (before today):

- Divide dataset into “training” and “testing sets”
- Data-mine the training set at will without peaking into the testing dataset.
- Identify set of plausible models from training data
- Carry out all inference in testing set

- Multiple comparison issues often not a concern in this framework

Practice vs ideal (before today):

- Sometimes collect second dataset only after using first dataset for full exploration
- Sometimes use single dataset but differentiate between prior and post-hoc models

Other “rules” of thumb:

- Extensive robustness checks to changes in data grouping and model details and preprocessing
- Look for additional tests of adaptively discovered models
- Focus on testing models with theoretical foundations (instead of ad-hoc data mining)

Key:

- Golden standard is replication cross samples, experimental designs, labs …

- Don't fully believe our findings and estimates until repeated replication in and outside lab
- Current journal/review practices severely taxes reporting of statistics that is sufficiently detailed and qualified

- Fortunately, key findings from lab/field have been systematically replicated
- But I will be shocked if we have never reported a false positive

But:

- I will be shocked if we have never reported a false positive
- We need better tools (with wide acceptance) badly

Q6) Is the widely advocated pre-registration a solution?

Response:

- Only in very limited cases
- Effective science needs to cope with adaptivity

Q7) Could the **Reusable Holdout Set** proposed here help?
Q8) What problems do you anticipate for its adoption?

Response 1:

- This is an extremely important development
- Best solution to the adaptivity problem that I have seen

- We have already started to use it in existing data

Thresholdout

Input:

Data S, holdout H, threshold $T > 0$, tolerance $\sigma > 0$

Given function q:

> Sample η, η' from $N(0,\sigma^2)$
>
> If $|avg_H[q] - avg_S[q]| > T + \eta$:
>
> output $avg_H[q] + \eta'$
>
> Otherwise:
>
> output $avg_S[q]$

Theorem. Thresholdout gives valid estimates for any sequence of adaptively chosen functions until n^2 overfitting* functions occurred.

Response 2: Critical hurdles for adoption:

1. Statistical illiteracy and math phobia in many areas of science
2. Publication friction with reviewers and editors

Response 3: What can authors do to help?

- Educate, educate, educate …
- Characterize more transparently the properties of commonly used statistical methods (e.g., logistic regression)
- Develop user friendly code and user guidelines