MSSE Data C200

Discussion #12

Name:

P-Hacking

The American Statistical Association (ASA) released a statement addressing misuse of the p-value with six principles:

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The main purpose of today's discussion is to elaborate on the fourth point.

Data Dredging, p-hacking, Scientific Studies

If one makes comparisons between enough variables, one is bound to find strong correlations in data where there is obviously no real, meaningful relationship. Look at some automatically generated Spurious Correlations.

Reports omit parts of their analyses. What usually happens is we only report the interesting findings without accounting for how we arrived at such a conclusion (e.g., behaving as if we loaded our data and tested a single hypothesis and obtained a statistically significant result). As such, often what we see is too good to be true (the 'file drawer effect"). There are many ways to p-hack:

- Analyze many measures (e.g. the 538 demo)
- Exclude data so findings look significant
- Collect data until findings look significant
- Transform data to make findings look more significant

Watch Last Week Tonight on Scientific Studies
`https://youtu.be/0Rnq1NpHdmw`.

538 Demo

Experiment with and Discuss Hack Your Way to Scientific Glory.
`https://projects.fivethirtyeight.com/p-hacking/`

Display and interact with the 538 demo.

Give students the opportunity to see that many choices could lead to either conclusion. Explain that even if the p-value was very small, it might not be solid evidence to conclude a causal statement such as ”The U.S. economy is affected by whether Republicans or Democrats are in office” due to confounding.

From the associated article
(`https://fivethirtyeight.com/features/science-isnt-broken/`:

> The data in our interactive tool can be narrowed and expanded (p-hacked) to make either hypothesis appear correct. That’s because answering even a simple scientific question — which party is correlated with economic success — requires lots of choices that can shape the results. This doesn’t mean that science is unreliable. It just means that it’s more challenging than we sometimes give it credit for.

1. What aspects of the analysis represent the researcher’s ‘degrees of freedom”?

2. What issue led us to hack artificial results? (Hint: read the description at the top of the demo)

   Possible options include

   - selection bias (Texas sharpshooter fallacy)
   - confirmation bias
   - post hoc ergo propter hoc
   - commercial bias
   - Simpson’s paradox

3. How could we protect against the hazards of confirmation bias in this sort of analysis?