Big Data and Security

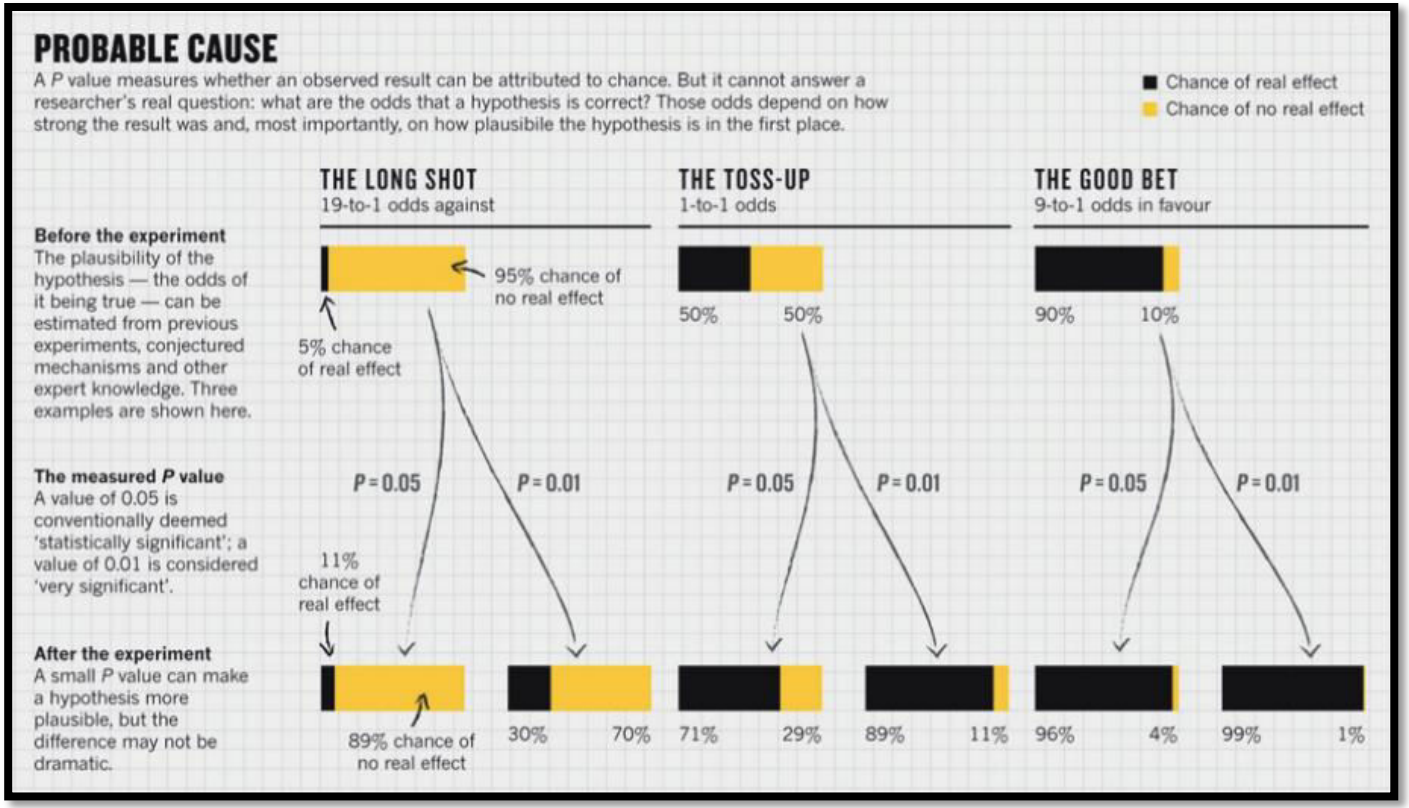
Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Applications of Bayesian Reasoning

An Example



Scientific method: Statistical errors. (n.d.). Retrieved from https://www.nature.com/news/scientific-method-statistical-errors-1.14700

Georgia Tech

A Problem with *p*-Values: Publication Bias

- It's typical in mainstream science to use p-values and the 5% confidence level to evaluate hypotheses
- A problem with this is publication bias
    - Studies with novel results are more likely to be published
    - But novel results are those which were unexpected (and thus are unlikely)
    - So these results are more likely to be found if they are in fact false. . .
- Note: this is not a problem with frequentist reasoning–the interpretation of a p-value is only valid if you do one hypothesis

Georgia Tech

Another Problem with *p*-Values: Multiple Hypotheses

- The p-value is designed to determine whether a particular hypothesis should be rejected
- But let's say you test 10 different, independent hypotheses, all of which in fact aren't true.
    - You will not reject the first one 95% of the time
    - You will not reject the second one 95% of the time
    - . . . You will not reject any of these hypotheses .9510 = .60 of the time
- So you will find one or more false hypotheses to be true 40% of the time So you have to be careful about this!!!

Georgia Tech

p-Values and Big Data

- So: there's a bias (in academics) towards novel results, and if you test to many hypotheses, you will find something to be true
- What implications does this have for big data analyses?
- You have to be extra careful because there always more X variables and hypotheses to test!

Georgia Tech

Lesson Summary

- The accuracy of frequentist hypothesis testing depends on how likely an assertion is to be true
- This problem interacts with publication bias to throw science off track
- As data sets get more complicated, prior beliefs are more important

Georgia Tech