Significance Testing in Theory and Practice

Daniel Greco
dlgreco@mit.edu

Introduction

- Significance testing is widely used, but can look thoroughly misguided:
  - "The results of a significance test, either of the Fisher or Neyman-Pearson variety, are often in flat contradiction to the conclusions which an impartial scientist or ordinary observer would draw" (Howson and Urbach 1993, p. 208)
  - "Although they are thoroughly fallacious, the methods of significance testing and classical estimation are still being advocated in hundreds of books, required texts in thousands of institutions of higher education, where hundreds of thousands of students are obliged to learn them." (p. 252)

Introduction

- In light of apparent problems with significance testing, we might seem to face an unpleasant dilemma:

  1. Find fault with criticisms of Significance testing, or
  2. Become skeptical about results in empirical disciplines that rely on significance testing

Introduction

- My aim in this talk is to show that can go between the horns of the dilemma.
- An analogy—the relationship between significance testing and Bayesian approaches to hypothesis testing is a bit like that between Newtonian mechanics and general relativity:
  - Significance testing is incorrect about the fundamentals, but still reliable in certain special cases which, conveniently, are the ones that usually come up in practice.

Introduction

- The Plan:
  1. Provide a brief summary of how significance testing works
  2. Identify two criticisms of significance testing—in particular, two types of cases in which significance testing would have us spuriously reject the null.
  3. For each type of case, identify special conditions under which it doesn't arise, and argue that those conditions are usually in place when significance testing is used.

How to Run a (Fisher) Significance Test:

Step 1. Formulate your null hypothesis

Step 2. Collect your data, and answer the following question:

"*Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?*"

Wikipedia Summarizing *Sage Dictionary of Statistics* (2004, p. 76)

Step 3. Depending on the answer computed in step 2, either reject the null hypothesis, or don't.

How to Run a (Fisher) Significance Test:

- Question: How should we interpret the instruction to "reject the null hypothesis"?
- I'll interpret it as placing a qualitative constraint on our degrees of belief:
  - "The rejection of a hypothesis…provides good reason, in the sense of rational degree of belief, for supposing the hypothesis to be false, but no numerical value can be placed upon this degree of belief." (Bulmer 1979, p. 165)

Problem 1—Probabilistic Modus Tolens

Probabilistic Modus Tolens (PMT):

1. If P, then *probably* ~Q
2. ~Q, therefore:
3. *Probably* ~P

In Significance Testing:

1. If the null hypothesis is true, then the value for the test statistic will *probably* not be at least as extreme as *x*.
2. The value for the test statistic is at least as extreme as *x*
3. So *probably*, the null hypothesis is false.

Problem 1—Probabilistic Modus Tolens

- As Sober (2008) explains, this form of inference is invalid. Examples:
  - Cases where all outcomes of some experimental process are equiprobable
  - Indeterministic theories faced with large bodies of data

Problem 2—Weakening the Evidence

- Example:
  - Someone's throwing a party. It might be Sam. Sam hates most plumbers, but likes Joe the plumber. Other potential hosts have the opposite preferences.
    - P(I meet Joe|Sam is hosting) > P(I meet Joe|Sam isn't hosting)
    - P(I meet a plumber|Sam is hosting) < P(I meet a plumber|Sam isn't hosting)
  - In this case, reasoning with weakened evidence would be disastrous.

Problem 2—Weakening the Evidence

- Analogous example in significance testing:
  - P(test statistic takes value *x*|The null hypothesis is true) ≥

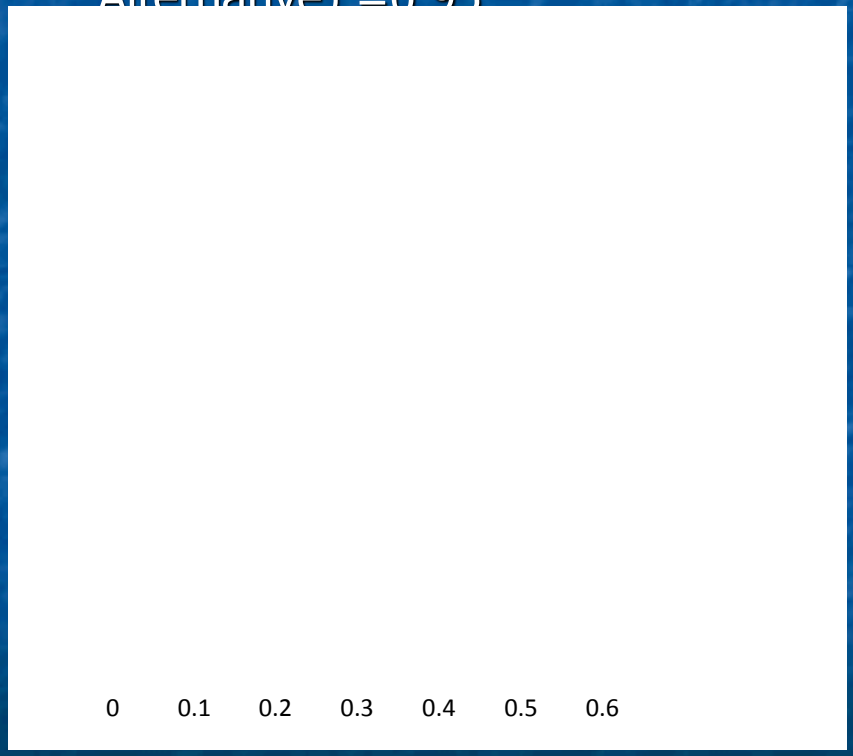
    P(test statistic takes value *x*|The null hypothesis is false)
  - P(test statistic takes some value at least as extreme as *x*|The null hypothesis is true) <

    P(test statistic takes some value at least as extreme as *x*|The null hyopthesis is false)

Solving the PMT Problem

- PMT is a good heuristic when there's an alternative hypothesis that:
  - Assigns the evidence a higher likelihood than the hypothesis being rejected
  - Has a sufficiently high prior probability

Posterior probability of null against prior probability of alternative when P(E|Null = 0.05), and P(E|Alternative) =0.95



Solving the PMT Problem

- Question: Why think there are such alternative hypotheses when significance tests are used?
- Answer: Norms of predesignation ensure that researchers have incentives to use significance tests just in case there are.
  - "It should be stressed, however, that the exact test to be used must be decided *before* the experiment has been done." (Bulmer 1979, p. 143)

Solving the PMT Problem

- How does predesignation help?
  - Without it, researchers who want to reject the null can perform an experiment, look for (perhaps arbitrary, gerrymandered) respects in which the data are unusual, and tailor their test to reject the null.
  - With it, researchers who want to reject the null must instead design a test that, *ex ante*, makes it likely that the null will be rejected.

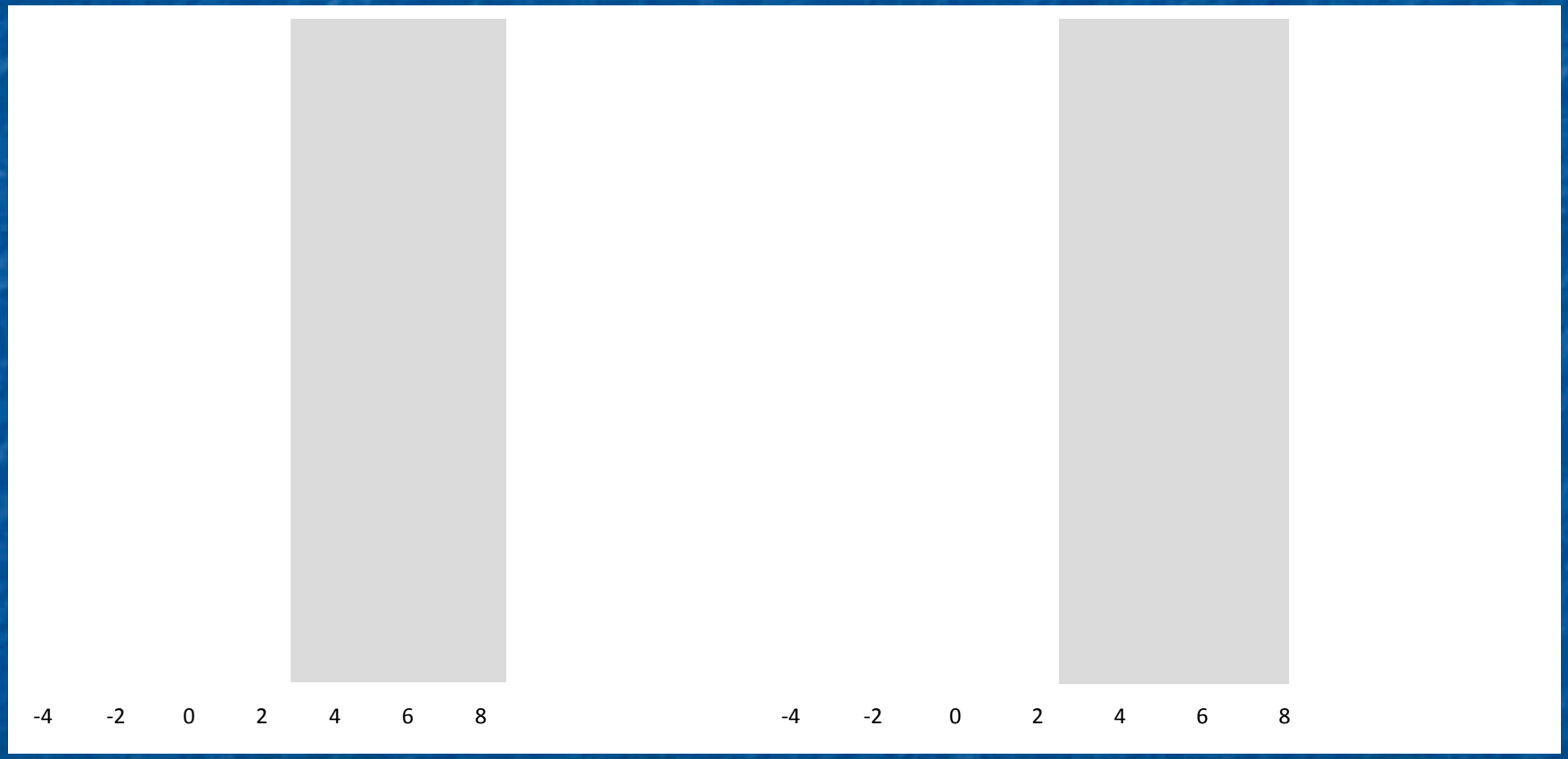
Solving the PMT Problem

- When is it *ex ante* likely that the null hypothesis will be rejected in a significance test?
- When there's an alternative hypothesis *H′* such that:
  - P(the test statistic should take a value that is extreme with respect to the null|*H′*) is high, and
  - *H′* is itself plausible, i.e., P(*H′*) is not too low.

Solving the PMT Problem

- But when there's such an alternative hypothesis, PMT is a reliable heuristic.
- Summing up:
  - Significance testing looks to commit PMT, which would suggest that there would be cases where even though P(test statistic takes a value at least as extreme as x|Null) $\leq$ 0.05, P(Null|test statistic takes a value at least as extreme as *x*) is relatively high.
  - These cases don't arise when there are alternative hypotheses with certain properties, and predesignation norms make it likely that there are such hypotheses when significance tests are used.

Weakening the Evidence—The Problem



Case where weakening is OK Case Where it's not

Natural and Unnatural Properties

An Argument Schema:

P1. Objects $o_1, o_2 \ldots o_n$ are all both F and G

P2. Object o_{n+1} is F

C. Object o_{n+1} is G

Properties for which the schema looks OK:

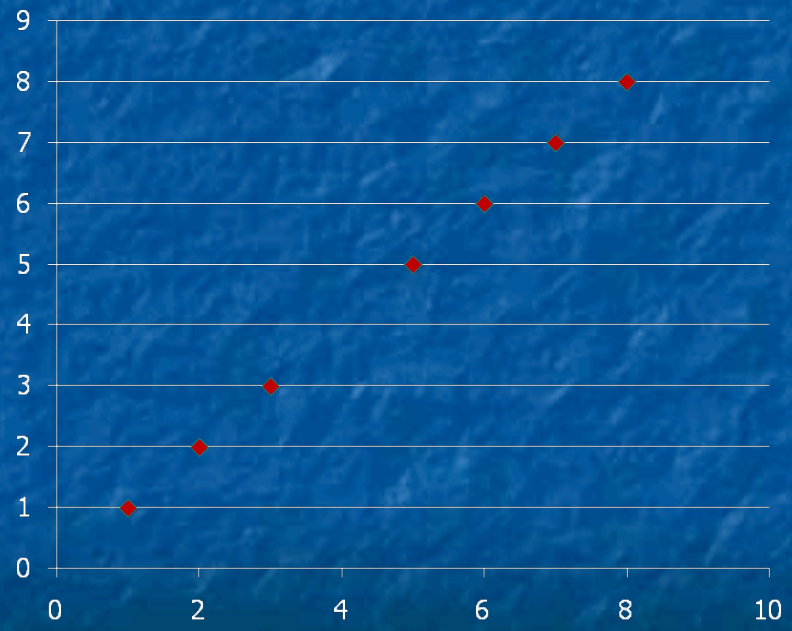
Being an emerald, being green, etc.

Properties for which it doesn't:

Being an emerose, being grue, etc.

Natural and Unnatural Quantities

- Some data from a hypothetical experiment:



- Should we fit a line to predict new data?
- Depends on what quantities the axes measure
  - If weight against food per day, yes

    Queight(x) = weight(x), unless queight(x) = 4 quounces, in which case weight(x) = 100 pounds
  - If queight against food per day, no

Natural and Unnatural Quantities

- The Moral: trying to fit simple curves to datasets is only attractive, even with *ceteris paribus* hedges, when the quantities in terms of which the data is expressed are natural quantities.

- More Generally:

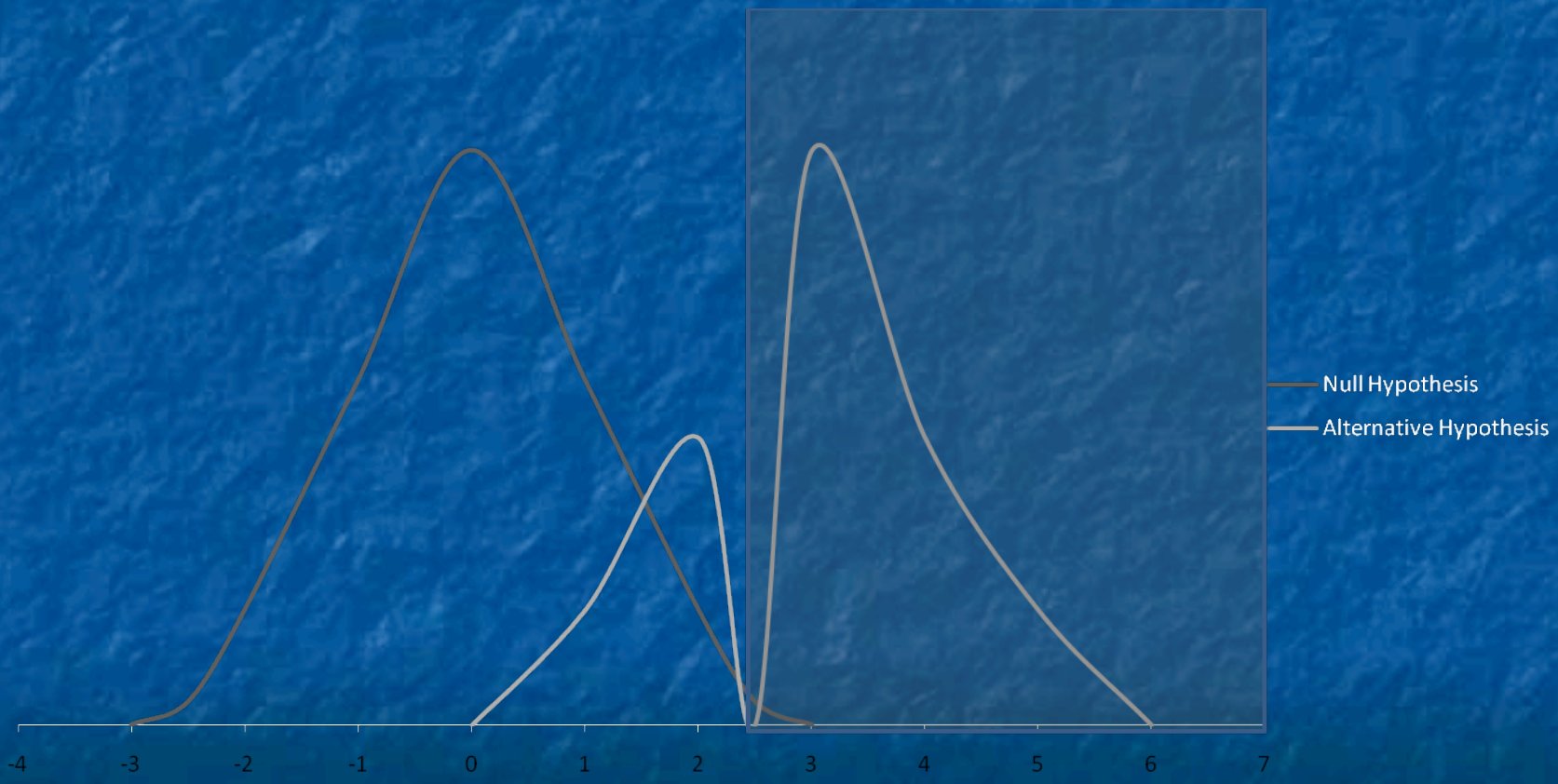
  Call the kinds of variables in terms of which we usually work our "standard" variables. It seems to be the case that, for whatever reason, our standard variables are usually smoothly distributed. If we go ahead and generalize from this observation (by enumerative induction), we arrive at the conclusion that most standard variable distributions are smooth. We may consequently take ourselves to have empirical grounds for adopting a revised and differently deployed "Principle of Insufficient Reason" of the following form:

  In the absence of any reason to think otherwise, assume that any standard variable is fairly smoothly distributed. (Strevens 1998, p. 241)

Natural and Unnatural Quantities

- Is Strevens' principle valid?
  - As a general, abstract claim, it's hard to evaluate, though (I think) it's plausible in the cases he applies it.
- Weaker claim:
  - In the context of significance testing, all plausible hypotheses will induce smooth probability distributions over natural choices for test statistics (e.g., avg. height in a random sample)

Weakening the Evidence Revisited



Weakening the Evidence Revisited

- Is ruling out unsmooth probability distributions over test statistics enough to rule out all spurious rejections of the null due to weakened evidence?
  - Not obviously, but I argue in the paper that it covers more cases than you might think

Summing Up

- I considered two types of cases (PMT, and weakening the evidence) in which significance testing would have us spuriously reject the null hypothesis.
- I identified special conditions under which these cases don't arise:
  - For PMT, the existence of a plausible alternative hypothesis that assigns the evidence a high likelihood
  - For weakening the evidence, the smoothness of the probability distribution over the test statisitic on the alternative hypothesis
- I argued (for PMT) or just claimed (for weakening) that these conditions are usually met

Summing Up

- What I didn't do
  - Argue that Bayesians should think that frequentist statistics is harmless in practice
    - Medical testing example
- What I did do
  - Argue that Bayesians should think that in typical cases when significance testers tell us to reject the null hypothesis, it has a low posterior probability

References

- Bulmer, M.G. (1979) *Principles of Statistics*. New York, Dover
- Duncan, and Howitt, Laurence. (2004) *The SAGE Dictionary of Statistics: A Practical Resource for Students in the Social Sciences*. London, SAGE
- Howson, Colin, and Urbach, Peter. (1993) *Scientific Reasoning: The Bayesian Approach*. Chicago. Open Court
- Sober, Elliott. (2008) *Evidence and Evolution*. Cambridge, UK. Cambridge University Press
- Strevens, Micahel. (1998) "Inferring Probabilities from Symmetries," *Noûs* 32, pp. 231-46