Introduction to Data Science

The Statistical Significance Filter

Collin Cademartori

November 13, 2020

Plan for This Lecture

- Review of p-values
- Definition of statistical power
- The statistical significance filter
- A possible solution
- Simulating the significance filter in R

P-values Review

- In science we are always comparing models to observation
- Suppose we find a model is consistent with some data
  - Can we say that our model is likely to be true?
  - Always possible there is another equally good model
  - If we discover such a model, can we believe first model?
- We always want to rule out alternative explanations
- P-values attempt to do this for certain “null” explanations

P-values Review

- Often have a single, simple effect in mind
- I.e. this drug will have some positive effect on some metric
- If we are right, can design a simple model of observations
- We observe an effect in patients by measuring this metric
- Measurements will differ due to:
  - Interaction b/w drug and other health factors
  - Measureent error from imperfect devices
  - Non-identical administration of the drug
- But on average, we expect the drug to help
- So we model measurements as $N(\mu, \sigma)$ with $\mu > 0$
  - Normal distribution captures variation b/w patients
  - σ is standard deviation - how much variation
  - $\mu > 0$ corresponds to positive effect of drug
- Want to rule out $\mu = 0$ explanation of data
- Otherwise data are consistent with drug doing nothing

P-values Review

- How do we rule out $\mu = 0$ as an explanation?
- What the data are highly improbable if $\mu = 0$?
- This would be good evidence against $\mu = 0$!
- The P-value is exactly a quantification of this idea
  - Suppose we make measurements x_i on patients $i = 1, \ldots, n$
  - Average measurement is denoted $\overline{x}$
  - If $\mu = 0$, $\overline{x}$ will be 0 on average
  - Therefore large $\overline{x}$ are unlikely if $\mu = 0$
- The P-value is **the probability that $\overline{x}$ is at least as large as the observed value in our experiment if in fact $\mu = 0$**

Sampling Distributions

- Some of the proceeding was imprecise
- What does "if $\mu = 0$, $\overline{x}$ will be 0 on average" mean?
- "Large" has to be defined relative to mesurement error
  - Imprecise measurement $\implies$ $\overline{x}$ could be large while $\mu = 0$
  - Measurement error depends on variation in x_i and size of n
  - More variation in measurements $\implies$ less precise average
  - Larger sample size n $\implies$ more precise average
  - Standard deviation of $\overline{x}$ is about $\sigma/\sqrt{n}$ (**standard error**)
- Also, what does "on average" mean?
  - We only observe one value of $\overline{x}$ from our data
  - But data give a sense of variation in future data
  - "On average" defined relative to replications of experiment
  - Distribution of a measure over replications is a **sampling distribution**

A Bit of Vocabulary

- An explanation like $\mu = 0$ is called a **null hypothesis**
    - This is the hypothesis we are trying to rule out
- A **significance level** α is a cut-off for the P-value
    - If $P < \alpha$ then we reject the null hypothesis
    - If $P \geq \alpha$ we say that we cannot reject
    - In this case we think the null is consistent with the data
    - A hard cut-off can be unnatural but is common in practice
- Often times the hypothesis like $\mu > 0$ is called the **alternative hypothesis**
- The process of calculating a P-value and comparing to a significannce level is called a **hypothesis test**

The Power of a Hypothesis Test

- The **power** of a hypothesis test quantifies its ability to rule out the null hypothesis when it is false (e.g. when $\mu > 0$)
- It is the probability of observing a P-value less than or equal to the significance level assuming the alternative hypothesis is true
- But this probability depends on the true μ value!
  - If μ is positive but close to zero it will be hard to distinguish this situation from $\mu = 0$ and power will be low
  - If μ is very far from zero (relative to the standard error), then it is easy to rule out $\mu = 0$
- The distance of μ from zero is called the **effect size**
- If we want to estimate power, we need an effect size estimate

Power Continued

- Why should we care about estimating power?
- Power depends on effect size relative to standard error
- But standard error usually decreases at rate $1/\sqrt{n}$
- We can't control the size of the effect...
- But we usually can control the sample size
- With an effect size estimate, we can control power by changing the sample size n
- If power is 20% we have only a 1/5 chance of detecting a nonzero effect *assuming that effect is there to begin with*
    - Studies are often expensive
    - Good power estimates help us set appropriate sample sizes
    - Better chance our study won't be a waste

Estimating Effect Size

- Key ingredient to power estimate is effect size estimate
- How do we estimate effect size?
  - Can retrospectively estimate as size of our measured effect
  - Can estimate from published studies of similar effects
  - Can just set to minimal size that would be of interest
- These are all usually bad ideas in practice
- Why? Published results routinely over-estimate effect sizes
- This is known as the **statistical significance filter**

Statistical Significance Filter

- Suppose there is a true effect of size 2
- Suppose we perform a study of this effect with a standard error of 1
- We also want the P-value of our estimate to be smaller than 5%
- This means that we want an estimate larger than 1.64
- Our estimate of the effect will be a random draw from a $N(2, 1)$ distribution
- Now suppose we only publish if we get a statistically significant result
- What does this do to the published estimates as a group?
- They will look like those draws from a $N(2, 1)$ distribution larger than 1.64
- We can simulate this!

Statistical Significance Filter

- If only statistically significant results are published, these estimates will systematically over-estimate the true effect
- In previous example with $\alpha = 0.05$, estimates were exaggerated by 30%
- If we take instead $\alpha = 0.01$, estimates exaggerated by 50%!
- What does this mean for power estimates?
    - We review literature to get effect size estimate
    - This estimate is an exaggeration of true effect size
    - The larger the true effect, the smaller sample size needed
    - If we set our sample size for 80% power, our actual power will be systematically lower than this
- We can derive a formula for power and simulate this problem in R

So What?

- What can we conclude about statistical significance and power?
  - If only statistically significant results are published, published estimates will over-estimate the true effect size
  - If these effect sizes are then used to compute power estimates, power can significantly over-estimated
  - This in turn can lead researchers to select sample sizes that are too small
  - They then proceed with confidence that they can detect an effect they expect
  - Statistically significant findings are then easily over-interpreted
  - We would then expect that many such studies would fail to replicate
  - And in an increasing number of fields this is exactly what we find