Hypothesis Testing

SML201: Introduction to Data Science, Spring 2019

Michael Guerzhoy

P-Value

- Assuming the Null Hypothesis is true, the probability of observing a value that is as extreme or more extreme than what we observe
- Informally: if nothing is actually going on, how weird would it be to observe the data we do?

Hypothesis testing

- A low p-value indicates we have evidence against the null hypothesis
- Traditional rule:
  - *Reject* the null hypothesis if the p-value is smaller than 0.05
  - Note: if the p-value is not smaller than 0.05, we do NOT accept the null hypothesis as true. We merely don't have evidence against it

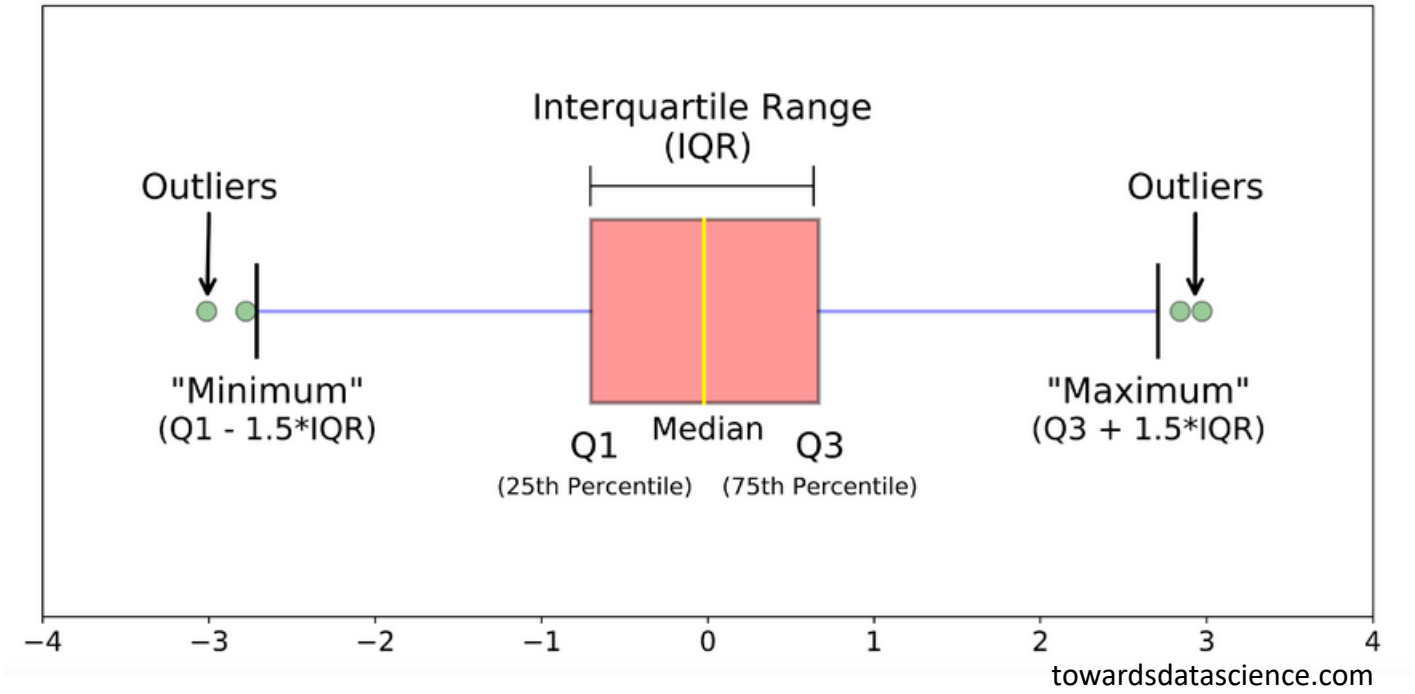
Hypothesis testing procedure

- Check that the *model assumptions* are satisfied by visualizing the data
- Compute the p-value
- *Reject* the null hypothesis if the p-value is smaller than a threshold

Is a distribution normal?

- The density should be roughly bell-shaped
- There should be very few (if any) data points further away than 3 standard deviations from the median

Boxplots



25th percentile: the point such that 25% of the data is smaller than the point, and
75% of data is larger

Outliers: datapoints that are very far away from most other datapoints. The definition is context-dependent

Finches example

Type I/Type II errors

- Type I error: rejecting a true null hypothesis (analogous to “false positive”)
- Type II error: failing to reject a false null hypothesis (analogous to “false negative”)
- Trade-off between the probability of Type I and Type II errors
  - Can’t make a Type I error if we never reject a hypothesis

Probability of a Type I error

- If we reject a null hypothesis whenever we get a p-value of 0.05 or smaller, we'll reject 1 out of 20 true null hypotheses
- One out of 20 studies will report evidence against true null hypotheses
  - Assuming each study contained exactly one hypothesis
  - Assuming every study actually gets published

Multiple hypotheses

- In Project 2, we compute about 7000 t-statistics
- A t-statistic outside of approximately [-2, 2] would lead to a p-value of less than 5%
- Even if no gene actually has different expression levels in ALL and AML leukemia tumors, we would conclude that 7000*0.05 = 350 genes do, if we are not careful

Multiple Hypotheses

- Solution 1: just have one null hypothesis
  - Would make science *really* slow
- Solution 2: pre-register all your null hypotheses, and report all results
  - If you report the results of 5 hypotheses, we know that you have a more than 1/20 chance of rejecting a true hypothesis
  - Required by the NIH for serious studies
- Solution 3 (in conjunction with Solution 2): adjust your p-value thresholds to compensate
  - There are formulas to do this. Most result in needing very large sample sizes (or large differences in the data) to reject any hypothesis at all

Science-Wide Multiple Hypotheses

- If a whole community of scientists keeps testing the same hypothesis (or variations of the same hypothesis), *someone* will reject it
  - "The desk drawer effect": journals will generally publish interesting results (a null hypothesis was rejected) and not publish boring results
- This is the same as trying slightly different versions of the hypothesis again and again

What is the Science-Wide False Discovery Rate?

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Modeling the Framework for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

Most Research Findings

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a

- Estimates vary from 15% to over 50%

Other causes of false discoveries

- Fraud
- P-hacking: rather than pre-registering a single hypothesis, testing multiple different hypotheses until one is rejected, and publishing that
  - Also a kind of fraud
- Honest research that is nevertheless like p-hacking
  - “The Garden of Forking Paths”
- Bad experiments

Solutions

- Pre-registration of studies
- Publishing negative as well as positive results
- Setting p-values to be really low (in particle physics, the standard for discover is p = 0.0000003)
- Replication: a study is only accepted if it was replicated
- Not believing “just one study”
  - Standard practice in medicine
- Only testing hypotheses when there is sound scientific basis for believing that something might be going on
  - E.g., a theory about biology, physics etc.
  - Limits number of hypotheses

The Social Psychology Replication Crisis

- Many studies in social psychology used very small samples
  - Of college undergrads
- In recent years, many studies failed to replicate
- Several famous examples of fraud or near-fraud
  - Coaching in the Stanford Prison Experiment?
- Currently, there is a movement toward more rigorous procedures and larger sample sizes

Fake Neuroscience News

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

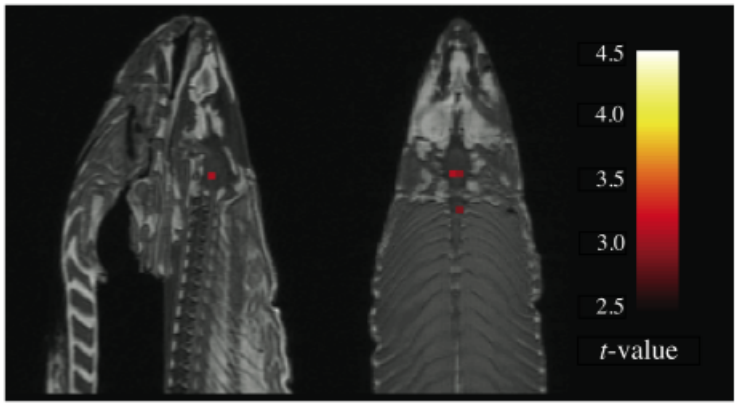
With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been

GLM RESULTS



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Type II errors

- Inevitable with small sample sizes
  - A small sample will not provide evidence against a null hypothesis a lot of the time
- Not really an error

“I have never in my life committed either a type I or a type II error” – Andrew Gelman

- All null hypotheses are false
- A type II error is not an error anyway

Type M errors and Type S errors

- Type M error: incorrectly estimating the *magnitude* of the effect
- Type S error: incorrectly estimating the *direction* of the effect
- Doesn’t really fit in well in the p-value framework

ASA Statement on P-values

- P-values can indicate how incompatible the data are with a specified statistical model.
- 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.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

P-values and Q-tips