Statistical Inference

"I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world." — Borges (1941)

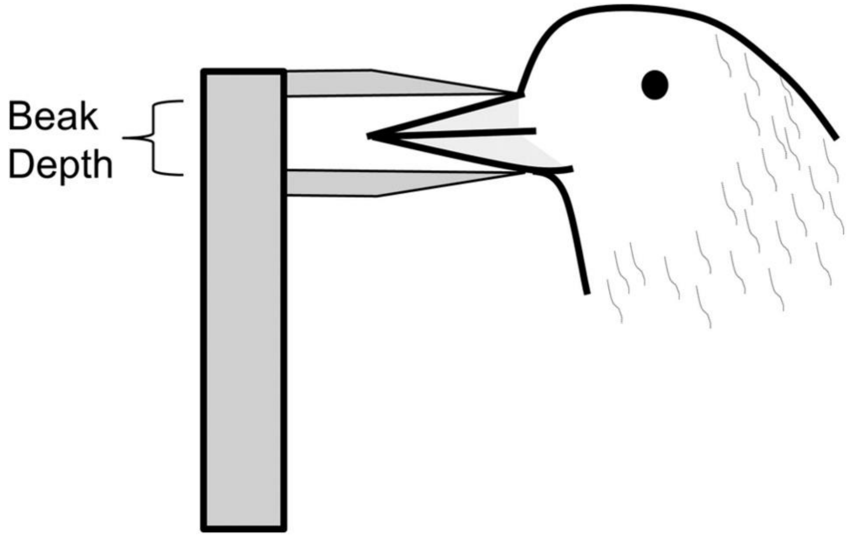
SML310: Research Projects in Data Science, Fall 2019

Michael Guerzhoy

Darwin's Finches

- Darwin's Conjecture (mid 19th century): the different kinds of finches on Galapagos Islands have different kinds of beaks because they adapted to their respective environments
- Data (from 1980s): the beak depths of more than 20 generations of finches
- In 1977, there was

a drought, and only large, tough seeds were available



(Beak data)

Review: The Null Hypothesis

- Null Hypothesis:
  - *Assume the beak depths are normally distributed, both pre-draught and post-draught*
  - The difference between the means of the distributions of beak depths pre-draught and post-draught is 0

- Why not a difference of e.g. 0.1 as the null hypothesis?

Null Hypothesis

- Ideally, the Null Hypothesis being true means that *nothing interesting is going on*. Ideally, rejecting the Null Hypothesis means that we learned something new. By default, we'd rather keep believing nothing interesting is going on and not believe something false
  - More or less the case here. It wouldn't be that surprising if the beak depth didn't have much to do with the toughness of the seeds

Review: p-values

- Test statistic: some function of the observed sample
  - **The sample**: all the measurements made both pre-draught and post-draught
  - **Example of a test statistic:** the difference between the means pre-draught and post-draught
  - **A more useful example of a test statistic:** the difference between the means pre-draught and post-draught, divided by the estimate of a standard deviation
- What is a p-value, in terms of a test statistic?

Review: p-values

- Test statistic: some function of the observed sample
  - **The sample**: all the measurements made both pre-draught and post-draught
  - **Example of a test statistic:** the difference between the means pre-draught and post-draught
  - **A more useful example of a test statistic:** the difference between the means pre-draught and post-draught, divided by the estimate of a standard deviation
- P-value: a measure of how extreme the test statistic is, *assuming the null-hypothesis is true*

Review: p-value of a test statistic

- Assuming the Null Hypothesis is true, the probability that the test statistic will be as extreme, or more extreme, than the observed value of the test statistic, when data is repeatedly sampled from the model
  - Null Hypothesis: $X_1, X_2, X_3, \dots, X_{89} \sim N(\mu_X, \sigma^2)$
    $Y_1, Y_2, Y_3, \dots, Y_{89} \sim N(\mu_Y, \sigma^2)$
    $\mu_X = \mu_X$
  - Test statistic: $\mathrm{T} = \frac{\bar{X} - \bar{Y}}{SE(\bar{X} - \bar{Y})}$ ($SE(\bar{X} - \bar{Y})$ is the estimate of the SD of $\bar{X} - \bar{Y}$)
  - We obtain the particular sample $x_1, \dots, x_{89}, y_1 \dots y_{89}$, and compute a particular t
  - (One-sided) p-value: $P(T > t)$, assuming the null hypothesis is true

Interpreting p-values

- $P(T > t) < 0.05$
  - (Small p-values in general)
  - We say “There is evidence against the Null Hypothesis”
  - If the Null Hypothesis is true, observing the value of the test statistic t that we actually observe would be unlikely
- $P(T > t) \geq 0.05$
  - (Large p-values in general)
  - We say “There is no (or weak) evidence against the Null Hypothesis”
  - If the Null Hypothesis is true, we would not be surprised to observe the value of t that we observed

ASA Statement on P-values

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.

Q-tips

VARIETY OF USES
Q-tips
Cotton Swabs
More soft cotton at the tip than any other swab
100% PURE COTTON
170 COTTON SWABS
170 cotton swabs
Q-tips

Say $p < 0.05$

- Is it reasonable to say we are 95% sure the null hypothesis is false?
- When can we not conclude anything at all, even if $p < 0.05$?

Types of errors

- Type I error: rejecting the null hypothesis, even though it is not false
- Type II error: not rejecting the null hypothesis, even though it is false

“I’ve never in my professional life made a Type I error *or* a Type II error. But I’ve made lots of errors. How can this be?

- A Type 1 error occurs only if the null hypothesis is true (typically if a certain parameter, or difference in parameters, equals zero). In the applications I’ve worked on, in social science and public health, I’ve never come across a null hypothesis that could actually be true, or a parameter that could actually be zero.
- A Type 2 error occurs only if I *claim* that the null hypothesis is true, and I would certainly not do that, given my statement above!”

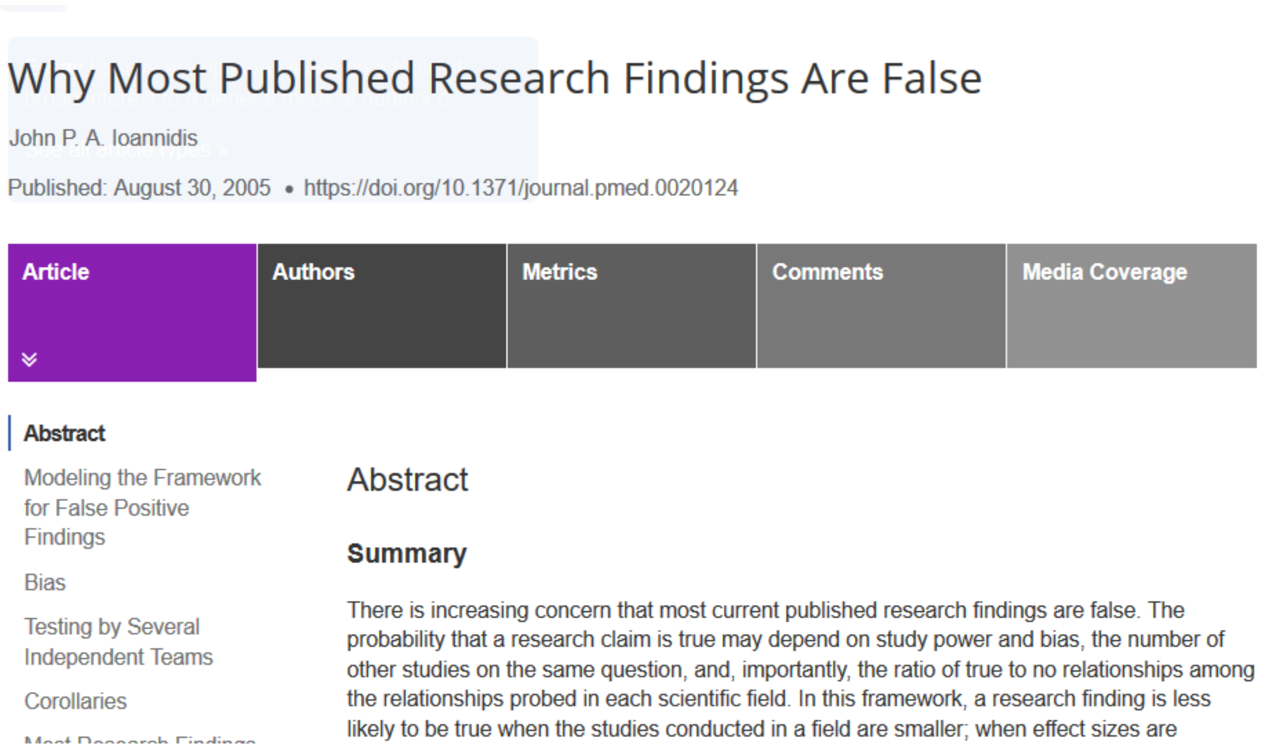
Andrew Gelman’s Error Typology

- Type S error: claiming that the effect is positive when it is actually negative
- Type M error: claiming a large-magnitude effect when the effect is actually small

Science-Wide False Discovery (Type I error) Rate

- What would you expect in an ideal world?

Science-Wide False Discovery (Type I error) Rate



Why Most Published Research Findings Are False

John P. A. Ioannidis

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

Abstract

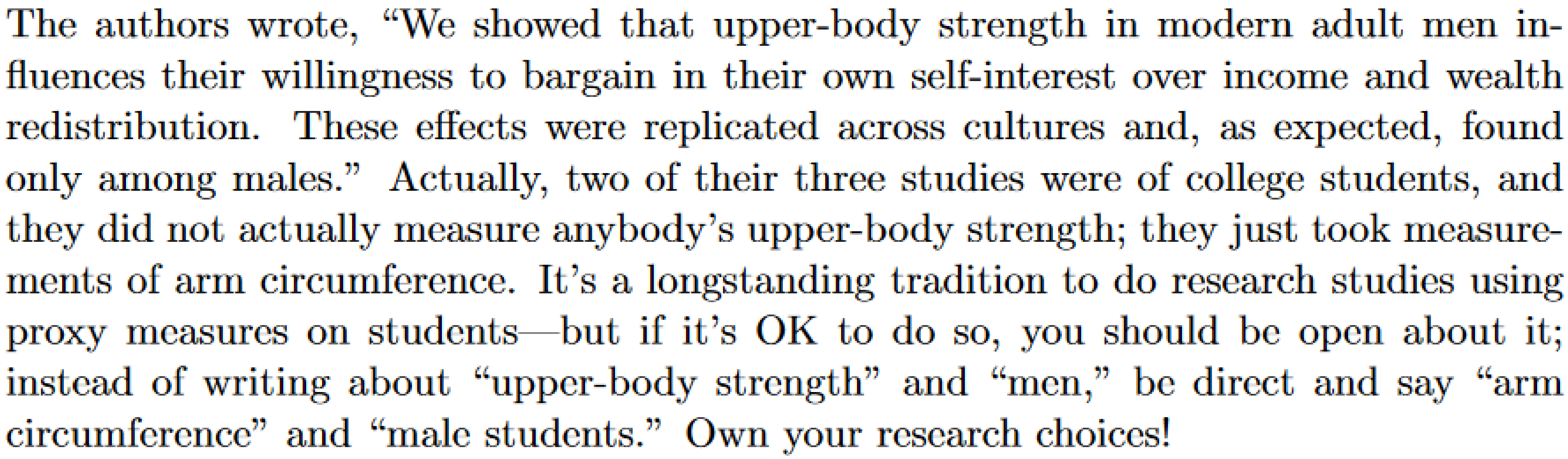
Summary

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

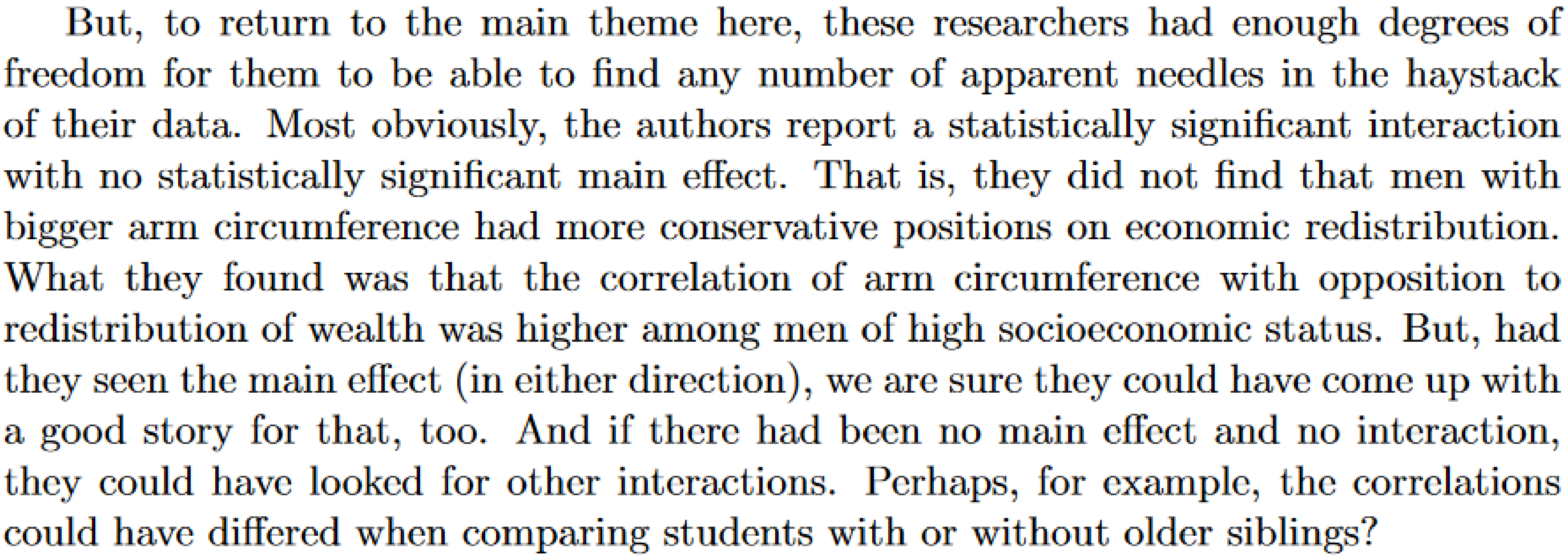
- More modest estimates are usually around 15% false findings

False Discoveries

- **Publication bias/file drawer effect:** a study showing there is an effect will be published, a study not rejecting a null hypothesis will not be
- **The garden of forking paths/p-hacking**: researchers will try different hypotheses, or shape hypotheses in such a way as to obtain a significant result (often unintentionally)
- **Bad models leading to bad p-values**



The authors wrote, “We showed that upper-body strength in modern adult men influences their willingness to bargain in their own self-interest over income and wealth redistribution. These effects were replicated across cultures and, as expected, found only among males.” Actually, two of their three studies were of college students, and they did not actually measure anybody’s upper-body strength; they just took measurements of arm circumference. It’s a longstanding tradition to do research studies using proxy measures on students—but if it’s OK to do so, you should be open about it; instead of writing about “upper-body strength” and “men,” be direct and say “arm circumference” and “male students.” Own your research choices!



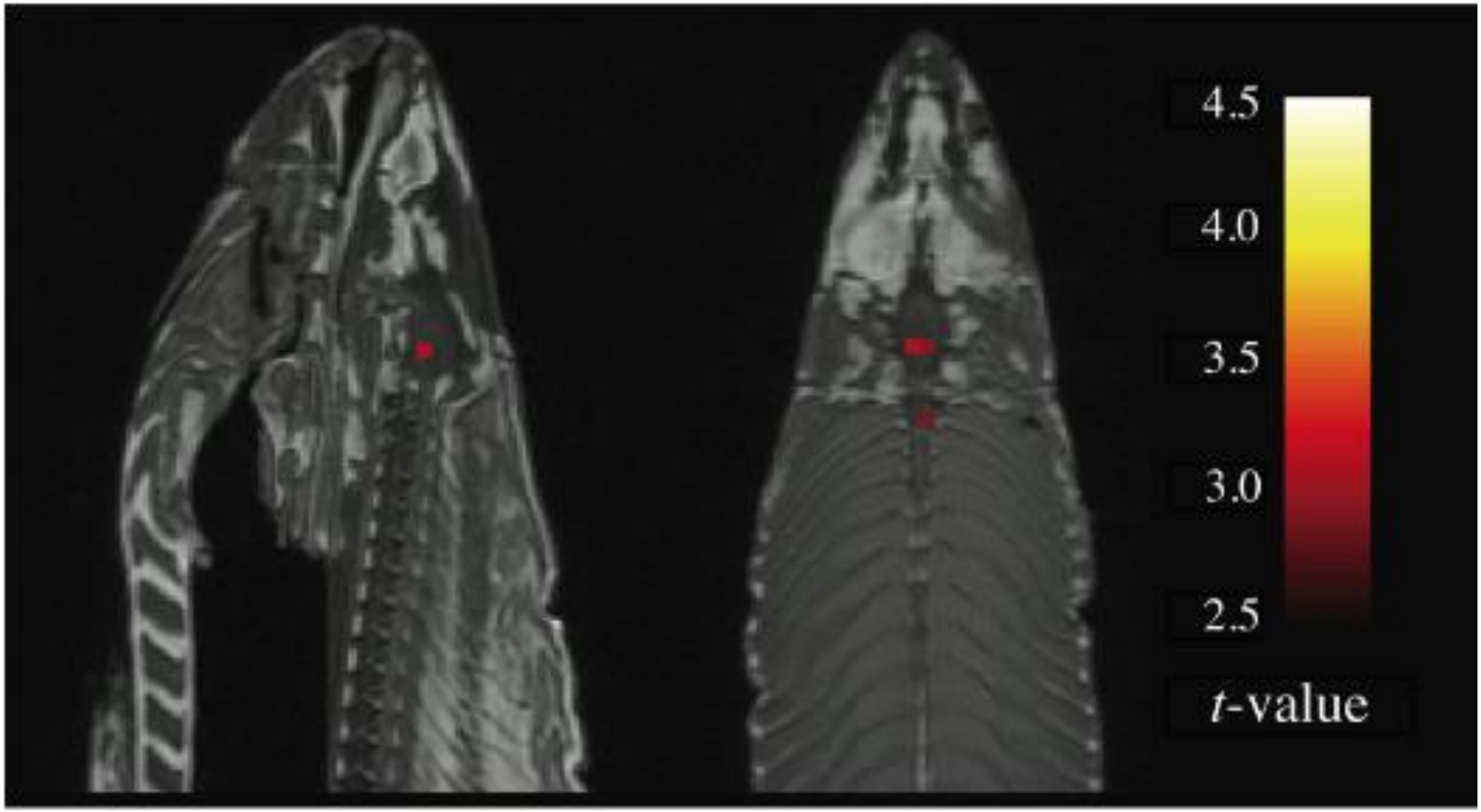
But, to return to the main theme here, these researchers had enough degrees of freedom for them to be able to find any number of apparent needles in the haystack of their data. Most obviously, the authors report a statistically significant interaction with no statistically significant main effect. That is, they did not find that men with bigger arm circumference had more conservative positions on economic redistribution. What they found was that the correlation of arm circumference with opposition to redistribution of wealth was higher among men of high socioeconomic status. But, had they seen the main effect (in either direction), we are sure they could have come up with a good story for that, too. And if there had been no main effect and no interaction, they could have looked for other interactions. Perhaps, for example, the correlations could have differed when comparing students with or without older siblings?

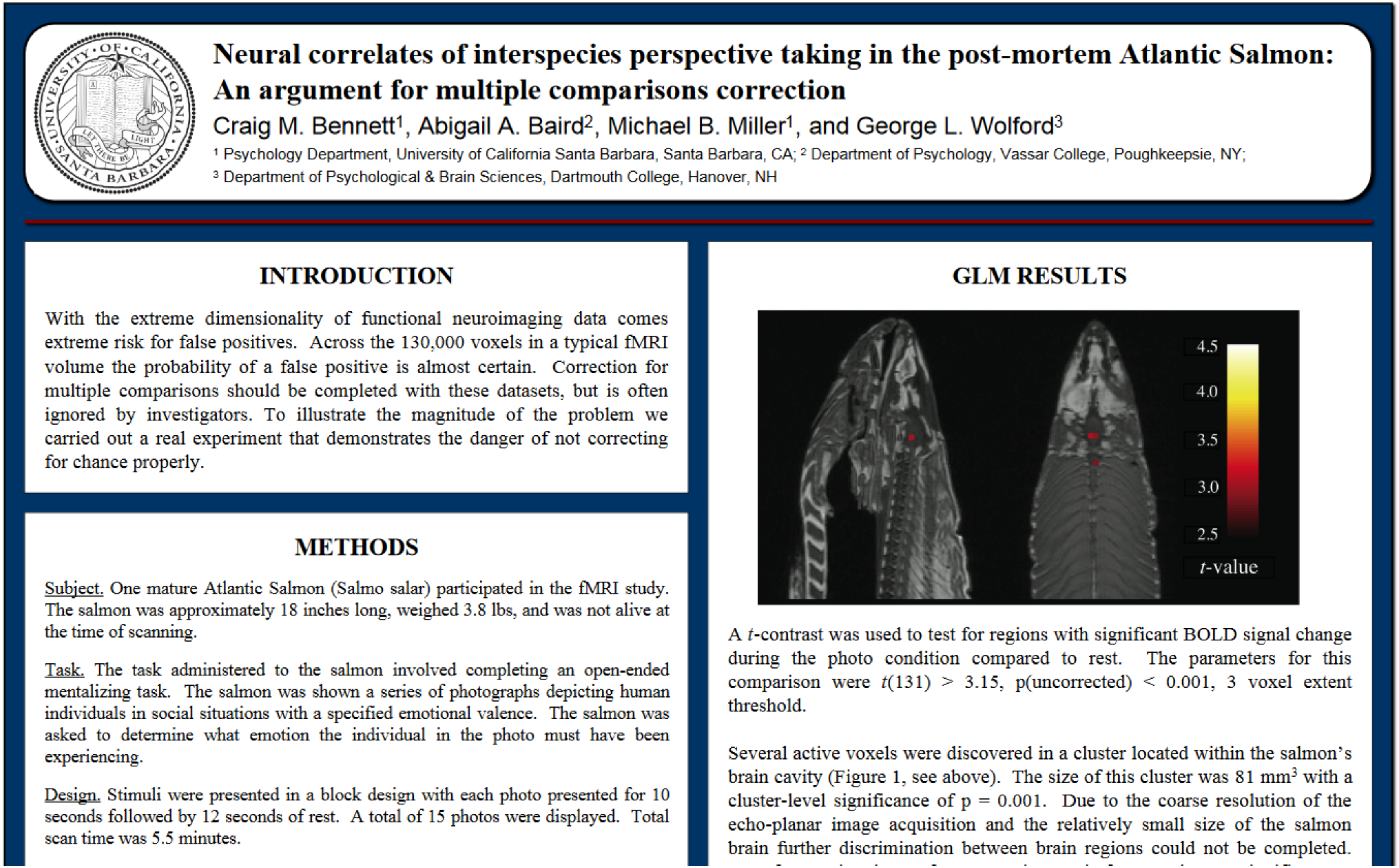
Remedies

- Never believe just one study on its own
  - Doctors never do
  - Attempts at replication and meta-analyses (looking at multiple studies) can address some of the issues
- Pre-registration
  - Specify what hypothesis is being tested *before* collecting the data
    - Expensive and difficult
    - The practice in biomedical fields
- Have solid theory behind models and null-hypotheses

Remedies

- In high-energy physics, p=0.003 means evidence of a particle. It’s only a discovery if p=0.0000003
    - How does one get really high p-values in general?


4.5
4.0
3.5
3.0
2.5
t-value



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 experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

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.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed.

Bayesian Inference (for finches)

- Frequentist inference/hypothesis testing framework: if the actual difference is 0, would we be very unlikely to obtain the value we actually obtain?
- Bayesian inference: obtain a posterior distribution for the difference *d*
  - Report the distribution
  - Requires a *prior* distribution
    - We'll discuss when those are possible to try to estimate using data
    - Bad prior distribution => don't need to bother with Bayesian inference

CI for the difference between populations of finches

- What does a 95% CI mean?

95% CI for the difference between populations of finches

- *If* the true difference is $\hat{d}$*, then* if we repeat the experiment 100 times, the estimated difference will fall within the CI 95% of the time
  - $\hat{d}$ is the estimated difference, not the true difference
  - Experiment: sample from the same populations (1976 and 1978 finches)
- Not true that the difference is within the CI with probability 95%