Meta-Topics in Behavioral Data and in Its Analysis

Wayne D. Gray
CRN 46232 COGS-6965-01

Posted 2016/10/28 at 14:41

This is a graduate level course.

1 DESCRIPTION

We will discuss enduring and new issues confronting researchers who collect and analyze behavioral data. This is not a course on probability theory, this is not a course that will teach statistical methods or techniques. We will focus on; (a) current controversies in the use and misuse of statistics by well-intended members of various research communities, (b) discussions of how statistical issues are presented to the public and interpreted by the media, (c) implications of the above for teaching statistics.

Updated 160828 to include the 2016-Aug-28 Sunday New York Times column titled: "Do you believe in God, or is that a software glitch?", the infamous "Dead Atlantic Salmon fMRI" poster, as well as the very recent PNAS paper referred to in the column.
# 2 TABLE OF CONTENTS

## 1 DESCRIPTION

## 2 TABLE OF CONTENTS

## 3 READINGS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>wk01: The Reproducibility Crisis – Introduction</td>
</tr>
<tr>
<td>3.1.1</td>
<td>p-values</td>
</tr>
<tr>
<td>3.1.2</td>
<td>wk02: p-hacking made easy</td>
</tr>
<tr>
<td>3.1.3</td>
<td>wk03: p-hacking – just below the threshold!?</td>
</tr>
<tr>
<td>3.2</td>
<td>wk04: Power to the Experimental Designs . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>3.3</td>
<td>wk05: p-values and Confidence Intervals</td>
</tr>
<tr>
<td>3.4</td>
<td>Perspectives</td>
</tr>
<tr>
<td>3.4.1</td>
<td>wk05: 2015 Saturday Night Brainstorming and Task Forces: (4th draft)</td>
</tr>
<tr>
<td>3.4.2</td>
<td>50 Years of Data Science</td>
</tr>
<tr>
<td>3.4.3</td>
<td>wk06: Exploratory and Confirmatory Research</td>
</tr>
<tr>
<td>3.4.4</td>
<td>wk06: The Influence of Statistics on Theory</td>
</tr>
<tr>
<td>3.4.5</td>
<td>To Get Ahead, Get a Theory</td>
</tr>
<tr>
<td>3.5</td>
<td>wk07: Injunctions and Policies and Blowback</td>
</tr>
<tr>
<td>3.5.1</td>
<td>The More Things Change, the More They Stay the Same . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>The American Psychological Association’s 1999 Statement</td>
</tr>
<tr>
<td>3.5.2</td>
<td>The American Statistics Society Statement on p-values</td>
</tr>
<tr>
<td>3.5.3</td>
<td>The Psychonomics Society’s Statements</td>
</tr>
<tr>
<td>3.5.4</td>
<td>The Association for Psychological Sciences Statements</td>
</tr>
<tr>
<td>3.5.5</td>
<td>wk09: Blowback</td>
</tr>
<tr>
<td>3.6</td>
<td>wk10: Reporting Practices and Proofing Stats</td>
</tr>
<tr>
<td>3.7</td>
<td>fMRI – Pretty Pictures are Good/Bad for Science</td>
</tr>
<tr>
<td>3.7.1</td>
<td>wk11: Draw me a picture – or two: Data Viz for Big and Small Data</td>
</tr>
<tr>
<td>3.7.2</td>
<td>wk11: Glowing brains – Is one pretty picture worth a thousand serious thoughts?</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Maybe more on this topic, maybe not . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>3.8</td>
<td>wk12: Data Sharing</td>
</tr>
<tr>
<td>3.9</td>
<td>wk13: Other stuff</td>
</tr>
<tr>
<td>3.10</td>
<td>NODS and More: Naturally Occurring Data Sets</td>
</tr>
<tr>
<td>3.10.1</td>
<td>wk13: Manifestos</td>
</tr>
<tr>
<td>3.10.2</td>
<td>wk13: Dangers of Big Data</td>
</tr>
<tr>
<td>3.10.3</td>
<td>wk14: NODs – pick one, any one!</td>
</tr>
</tbody>
</table>

## 4 REQUIREMENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>PreRequisites</td>
</tr>
<tr>
<td>4.2</td>
<td>About the Instructor</td>
</tr>
</tbody>
</table>

## 5 Honors Policy

## 6 Grading Policy

## 7 References

## 3 READINGS

### 3.1 wk01: The Reproducibility Crisis – Introduction

Abstract: John Oliver discusses how and why media outlets so often report untrue or incomplete information as science.


Abstract: “Recently, I lost a day at work when my Upshot colleague Austin Frakt emailed me first thing in the morning to tell me that headlines were appearing declaring that an “explosive new cellphone-cancer” study was making the rounds. As I have long been interested in this topic, I started to read the headlines and news.”


Abstract: More than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from Nature’s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.


Abstract: A recent study of the replicability of key psychological findings is a major contribution toward understanding the human side of the scientific process. Despite the careful and nuanced analysis reported, the simple narrative disseminated by the mass, social, and scientific media was that in only 36% of the studies were the original results replicated. In the current study, however, we showed that 77% of the replication effect sizes reported were within a 95% prediction interval calculated using the original effect size. Our analysis suggests two critical issues in understanding replication of psychological studies. First, researchers’ intuitive expectations for what a replication should show do not always match with statistical estimates of replication. Second, when the results of original studies are very imprecise, they create wide prediction intervals—and a broad range of replication effects that are consistent with the original estimates. This may lead to effects that replicate successfully, in that replication results are consistent with statistical expectations, but do not provide much information about the size (or existence) of the true effect. In this light, the results of the Reproducibility Project: Psychology can be viewed as statistically consistent with what one might expect when performing a large-scale replication experiment.

3.1.1 p-values
- The Dance of the P Values: http://www.youtube.com/watch?v=ez4DgdurRPg
- Youtube video on the BIC

3.1.2 wk02: p-hacking made easy


Abstract: If science were a game, a dominant rule would probably be to collect results that are statistically significant. Several reviews of the psychological literature have shown that around
96% of papers involving the use of null hypothesis significance testing report significant outcomes for their main results but that the typical studies are insufficiently powerful for such a track record. We explain this paradox by showing that the use of several small underpowered samples often represents a more efficient research strategy (in terms of finding $p < .05$) than does the use of one larger (more powerful) sample. Publication bias and the most efficient strategy lead to inflated effects and high rates of false positives, especially when researchers also resorted to questionable research practices, such as adding participants after intermediate testing. We provide simulations that highlight the severity of such biases in meta-analyses. We consider 13 meta-analyses covering 281 primary studies in various fields of psychology and find indications of biases and/or an excess of significant results in seven. These results highlight the need for sufficiently powerful replications and changes in journal policies.


**Abstract**: My dear fellow scientists! "If you torture the data long enough, it will confess." This aphorism, attributed to Ronald Coase, sometimes has been used in a disrespectful manner, as if it was wrong to do creative data analysis. In fact, the art of creative data analysis has experienced despicable attacks over the last years. A small but annoyingly persistent group of second-stringers tries to denigrate our scientific achievements. They drag psychological science through the mire. These people propagate stupid method repetitions; and what was once one of the supreme disciplines of scientific investigation – a creative data analysis of a data set – has been crippled to conducting an empty-headed step-by-step pre-registered analysis plan. (Come on: If I lay out the full analysis plan in a pre-registration, even an undergrad student can do the final analysis, right? Is that really the high-level scientific work we were trained for so hard?). They broadcast in an annoying frequency that p-hacking leads to more significant results, and that researcher who use p-hacking have higher chances of getting things published. What are the consequences of these findings? The answer is clear. Everybody should be equipped with these powerful tools of research enhancement!


**3.1.3 wk03: p-hacking – just below the threshold!?**


**Abstract**: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively
rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.


**Abstract:** In null hypothesis significance testing (NHST), p values are judged relative to an arbitrary threshold for significance (.05). The present work examined whether that standard influences the distribution of p values reported in the psychology literature. We examined a large subset of papers from three highly regarded journals. Distributions of p were found to be similar across the different journals. Moreover, p values were much more common immediately below .05 than would be expected based on the number of p values occurring in other ranges. This prevalence of p values just below the arbitrary criterion for significance was observed in all three journals. We discuss potential sources of this pattern, including publication bias and researcher degrees of freedom.


**Abstract:** In this blog, I’ll explain how p-hacking will not lead to a peculiar prevalence of p-values just below .05 (e.g., in the 0.045-0.05 range) in the literature at large, but will instead lead to a difficult to identify increase in the Type 1 error rate across the 0.00-0.05 range. I re-analyze data by Masicampo & LaLande (2012), and try to provide a better model of the p-values they have observed through simulations in R (code below). I’d like to thank E.J. Masicampo and Daniel LaLande for sharing and allowing me to share their data, as well as their quick response to questions and a draft of this post. Thanks to Ryne Sherman for his p-hack code in R, to Nick Brown for asking me how M&L’s article related to my previous blog post and for comments and suggestions on this post, and Tal Yarkoni for feedback on an earlier draft.


**Abstract:** Some effects are statistically significant. Other effects do not reach the threshold of statistical significance and are sometimes described as “marginally significant” or as “approaching significance.” Although the concept of marginal significance is widely deployed in academic psychology, there has been very little systematic examination of psychologists’ attitudes toward these effects. Here, we report an observational study in which we investigated psychologists’ attitudes concerning marginal significance by examining their language in over 1,500 articles published in top-tier cognitive, developmental, and social psychology journals. We observed a large change over the course of four decades in psychologists’ tendency to describe a p value as marginally significant, and overall rates of use appear to differ across subfields. We discuss possible explanations
for these findings, as well as their implications for psychological research.

3.2 wk04: Power to the Experimental Designs ... NOT!!


**Abstract:** Many psychology studies are statistically underpowered. In part, this may be because many researchers rely on intuition, rules of thumb, and prior practice (along with practical considerations) to determine the number of subjects to test. In Study 1, we surveyed 291 published research psychologists and found large discrepancies between their reports of their preferred amount of power and the actual power of their studies (calculated from their reported typical cell size, typical effect size, and acceptable alpha). Furthermore, in Study 2, 89% of the 214 respondents overestimated the power of specific research designs with a small expected effect size, and 95% underestimated the sample size needed to obtain .80 power for detecting a small effect. Neither researchers’ experience nor their knowledge predicted the bias in their self-reported power intuitions. Because many respondents reported that they based their sample sizes on rules of thumb or common practice in the field, we recommend that researchers conduct and report formal power analyses for their studies.

Lakens, D. (2015, JUL 30). On the challenges of drawing conclusions from p-values just below 0.05. *PEERJ*, 3(1-14). doi:{10.7717/peerj.1142}.

**Abstract:** In recent years, researchers have attempted to provide an indication of the prevalence of inflated Type 1 error rates by analyzing the distribution of p-values in the published literature. De Winter & Dodou (2015) analyzed the distribution (and its change over time) of a large number of p-values automatically extracted from abstracts in the scientific literature. They concluded there is a ‘surge of p-values between 0.041-0.049 in recent decades’ which ‘suggests (but does not prove) questionable research practices have increased over the past 25 years.’ I show the changes in the ratio of fractions of p-values between 0.041-0.049 over the years are better explained by assuming the average power has decreased over time. Furthermore, I propose that their observation that p-values just below 0.05 increase more strongly than p-values above 0.05 can be explained by an increase in publication bias (or the file drawer effect) over the years (cf. Fanelli, 2012; Paltasso, 2010), which has led to a relative decrease of ‘marginally significant’ p-values in abstracts in the literature (instead of an increase in p-values just below 0.05). I explain why researchers analyzing large numbers of p-values need to relate their assumptions to a model of p-value distributions that takes into account the average power of the performed studies, the ratio of true positives to false positives in the literature, the effects of publication bias, and the Type 1 error rate (and possible mechanisms through which it has inflated). Finally, I discuss why publication bias and underpowered studies might be a bigger problem for science than inflated Type 1 error rates, and explain the challenges when attempting to draw conclusions about inflated Type 1 error rates from a large heterogeneous set of p-values.


**Abstract**: Statistical rituals largely eliminate statistical thinking in the social sciences. Rituals are indispensable for identification with social groups, but they should be the subject rather than the procedure of science. What I call the “null ritual” consists of three steps: (1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis, (2) use the 5% significance level for rejecting the null and accepting your hypothesis, and (3) always perform this procedure. I report evidence of the resulting collective confusion and fears about sanctions on the part of students and teachers, researchers and editors, as well as textbook writers.

### 3.3 wk05: p-values and Confidence Intervals


**Abstract**: In the field of psychology, the practice of p value null-hypothesis testing is as widespread as ever. Despite this popularity, or perhaps because of it, most psychologists are not aware of the statistical peculiarities of the p value procedure. In particular, p values are based on data that were never observed, and these hypothetical data are themselves influenced by subjective intentions. Moreover, p values do not quantify statistical evidence. This article reviews these p value problems and illustrates each problem with concrete examples. The three problems are familiar to statisticians but may be new to psychologists. A practical solution to these p value problems is to adopt a model selection perspective and use the Bayesian information criterion (BIC) for statistical inference (Raftery, 1995). The BIC provides an approximation to a Bayesian hypothesis test, does not require the specification of priors, and can be easily calculated from SPSS output.


**Abstract**: Null hypothesis significance testing (NHST) is undoubtedly the most common inferential technique used to justify claims in the social sciences. However, even staunch defenders of NHST agree that its outcomes are often misinterpreted. Confidence intervals (CIs) have frequently been proposed as a more useful alternative to NHST, and their use is strongly encouraged in the APA Manual. Nevertheless, little is known about how researchers interpret CIs. In this study, 120 researchers and 442 students – all in the field of psychology – were asked to assess the truth value of six particular statements involving different interpretations of a CI. Although all six statements were false, both researchers and students endorsed on average more than three statements, indicating a gross misunderstanding of CIs. Self-declared experience with statistics was not related to researchers’ performance, and, even more surprisingly, researchers hardly outperformed the students, even though the students had not received any education on statistical inference whatsoever. Our findings suggest that many researchers do not know the correct interpretation of a confidence interval. The misunderstandings surrounding p values and CIs are particularly unfortunate because they constitute the main tools by which psychologists draw conclusions from data.

### 3.4 Perspectives

#### 3.4.1 wk05: 2015 Saturday Night Brainstorming and Task Forces: (4th draft)

Abstract: [from wdg] This is statistical satire. The link was sent by Gigerenzer in an email to the ABC mailing list about a month after the blog post came out. After spending the last four weeks wading through various critiques of the Frequentist approach to statistics, you will find it refreshing. (Also check out some of the many links embedded herein to other juicy and intellectually invigorating blogs by Deborah Mayo.

3.4.2 50 Years of Data Science


Abstract: More than 50 years ago, John Tukey called for a reformation of academic statistics. In 'The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or 'data analysis'. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name “Data Science” for his envisioned field. ... This paper reviews some ingredients of the current “Data Science moment”, including recent commentary about data science in the popular media, and about how/whether Data Science is really different from Statistics.

3.4.3 wk06: Exploratory and Confirmatory Research


Abstract: Adrianus Dingeman de Groot (1914-2006) was one of the most influential Dutch psychologists. He became famous for his work “Thought and Choice in Chess”, but his main contribution was methodological - De Groot co-founded the Department of Psychological Methods at the University of Amsterdam (together with R. F. van Naerssen), founded one of the leading testing and assessment companies (CITO), and wrote the monograph “Methodology” that centers on the empirical-scientific cycle: observation-induction-deduction-testing-evaluation. Here we translate one of De Groot's early articles, published in 1956 in the Dutch journal Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden. This article is more topical now than it was almost 60 years ago. De Groot stresses the difference between exploratory and confirmatory (“hypothesis testing”) research and argues that statistical inference is only sensible for the latter: “One ‘is allowed’ to apply statistical tests in exploratory research, just as long as one realizes that they do not have evidential impact”. De Groot may have also been one of the first psychologists to argue explicitly for preregistration of experiments and the associated plan of statistical analysis. The appendix provides annotations that connect De Groot's arguments to the current-day debate on transparency and reproducibility in psychological science. (c) 2014 Elsevier B.V. All rights reserved.

3.4.4 wk06: The Influence of Statistics on Theory

**Abstract:** Conducted a historical analysis of the period from 1925 to 1950 to investigate the incorporation of ANOVA techniques in psychological research. In addition to attempts to identify the earliest uses of ANOVA in psychology, the gradual incorporation of the technique was examined by counting its appearance in 6,457 articles appearing in major American psychological journals from 1935 through 1952. Expository articles and texts directed at psychologists were also identified, and a questionnaire survey of graduate psychology programs was conducted to establish how psychologists were introduced to ANOVA. Finally, the phylogeny of major contributors to American psychological statistics was established for the period. Data suggest a segmentation into 3 historical periods: (a) an initial, expository phase lasting until the onset of World War II; (b) a wartime interregnum during which use of ANOVA declined; and (c) a postwar resurgence, characterized by the institutionalization of ANOVA training. Although ANOVA certainly had a major impact on experimental psychology, the data do not permit the conclusion that the incorporation of ANOVA in psychology in itself constituted a revolutionary “paradigm shift.”


**Abstract:** The study of scientific discovery-where do new ideas come from?-has long been denigrated by philosophers as irrelevant to analyzing the growth of scientific knowledge. In particular, little is known about how cognitive theories are discovered, and neither the classical accounts of discovery as either probabilistic induction (e.g., Reichenbach, 1938) or lucky guesses (e.g., Popper, 1959), nor the stock anecdotes about sudden “eureka” moments deepen the insight into discovery. A heuristics approach is taken in this review, where heuristics are understood as strategies of discovery less general than a supposed unique logic of discovery but more general than lucky guesses. This article deals with how scientists’ tools shape theories of mind, in particular with how methods of statistical inference have turned into metaphors of mind. The tools-to- theories heuristic explains the emergence of a broad range of cognitive theories, from the cognitive revolution of the 1960s up to the present, and it can be used to detect both limitations and new lines of development in current cognitive theories that investigate the mind as an “intuitive statistician.”

### 3.4.5 To Get Ahead, Get a Theory


**Abstract:** Although Genevan research has provided a detailed analysis of cognitive structures, our knowledge of cognitive processes remains fragmentary. The focus is now not only on macro-development but also on changes occurring in children’s spontaneous action sequences in micro-formation. A series of experiments designed to study goal-oriented behavior is in progress. This paper describes the action sequences of 67 subjects between 4;6 and 9;5 years in a block balancing task. It is not a study of children’s understanding of a specific notion in physics, but an attempt to pave the way towards understanding the more general processes of cognitive behavior. The analysis focuses on the interplay between the child’s action sequences and his implicit theories which the observer infers from the sequences rather than from his verbal comments. Emphasis is placed on the role of counterexamples and on shifts in attention from goal to means. The construction and overgeneralization of ‘theories-in-action’ appear to be dynamic and general processes which are not stage-linked. The results also suggest certain functional rather than struc-
tural analogies between the acquisition of physical knowledge and the acquisition of language.

3.5 wk07: Injunctions and Policies and Blowback

3.5.1 The More Things Change, the More They Stay the Same . . . The American Psychological Association’s 1999 Statement


**Abstract**: The sections in italics are proposed guidelines that the TFSI recommends could be used for revising the APA publication manual or for developing other BSA supporting materials. Following each guideline are comments, explanations, or elaborations assembled by Leland Wilkinson for the task force and under its review. This report is concerned with the use of statistical methods only and is not meant as an assessment of research methods in general. Psychology is a broad science. Methods appropriate in one area may be inappropriate in another.

3.5.2 The American Statistics Society Statement on p-values

eprint: http://dx.doi.org/10.1080/00031305.2016.1154108.

**Abstract**: Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly. Underpinning many published scientific conclusions is the concept of “statistical significance,” typically assessed with an index called the p-value. While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced. In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed principles underlying the proper use and interpretation of the p-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

3.5.3 The Psychonomics Society’s Statements

Abstract: These guidelines focus on the analysis and reporting of quantitative data. Many of the issues described below pertain to vulnerabilities in null hypothesis significance testing (NHST), in which the central question is whether or not experimental measures differ from what would be expected due to chance. Below we emphasize some steps that researchers using NHST can take to avoid exacerbating those vulnerabilities. Many of the guidelines are long-standing norms about how to conduct experimental research in psychology. Nevertheless, researchers may benefit from being reminded of some of the ways that poor experimental procedure and analysis can compromise research conclusions. Authors are asked to consider the following issues for each manuscript submitted for publication in a Psychonomic Society journal. Some of these issues are specific to NHST, but many of them apply to other approaches as well.

3.5.4 The Association for Psychological Sciences Statements

Abstract: Psychological Science, the journal, and psychological science, the field, continue to struggle with the challenge of establishing interesting and important and replicable phenomena. As I often tell my students, “If scientific psychology was easy, everyone would do it.” We can take some comfort in knowing that other sciences, too, face similar challenges (e.g., Begley & Ellis, 2012). But our business is with psychology.

3.5.5 wk09: Blowback

Abstract: The ASA statement about the misuses of the p-value singles it out. It is just as well relevant to the use of most other statistical methods: context matters, no single statistical measure suffices, specific thresholds should be avoided and reporting should not be done selectively. The latter problem is discussed mainly in relation to omitted inferences. We argue that the selective reporting of inferences problem is serious enough a problem in our current industrialized science even when no omission takes place. Many R tools are available to address it, but they are mainly used in very large problems and are grossly underused in areas where lack of replicability hits hard.


Abstract: The journal of Basic and Applied Social Psychology banned the p-value in 2015, after Trafimow (2014) had explained in an editorial a year earlier that inferential statistics were no longer required. In the 2014 editorial, Trafimow notes how: “The null hypothesis significance procedure has been shown to be logically invalid and to provide little information about the actual likelihood of either the null or experimental hypothesis (see Trafimow, 2003; Trafimow & Rice,
The goal of this blog post is to explain why the arguments put forward in Trafimow & Rice (2009) are incorrect. Their simulations illustrate how meaningless questions provide meaningless answers, but they do not reveal a problem with p-values. Editors can do with their journal as they like – even ban p-values. But if the simulations upon which such a ban is based are meaningless, the ban itself becomes meaningless.

3.6 wk10: Reporting Practices and Proofing Stats


Abstract: Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so-and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the P values they produce) can lead to small P values even if the declared test hypothesis is correct, and can lead to large P values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of P values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.


Abstract: This study documents reporting errors in a sample of over 250,000 p-values reported in eight major psychology journals from 1985 until 2013, using the new R package “statcheck.” statcheck retrieved null-hypothesis significance testing (NHST) results from over half of the articles from this period. In line with earlier research, we found that half of all published psychology papers that use NHST contained at least one p-value that was inconsistent with its test statistic and degrees of freedom. One in eight papers contained a grossly inconsistent p-value that may have affected the statistical conclusion. In contrast to earlier findings, we found that the average
prevalence of inconsistent p-values has been stable over the years or has declined. The prevalence of gross inconsistencies was higher in p-values reported as significant than in p-values reported as nonsignificant. This could indicate a systematic bias in favor of significant results. Possible solutions for the high prevalence of reporting inconsistencies could be to encourage sharing data, to let co-authors check results in a so-called “co-pilot model,” and to use statecheck to flag possible inconsistencies in one’s own manuscript or during the review process.

### 3.7 fMRI – Pretty Pictures are Good/Bad for Science

#### 3.7.1 wk11: Draw me a picture – or two: Data Viz for Big and Small Data


**Abstract:** This article discusses the role of data visualization in the process of analyzing big data. We describe the historical origins of statistical graphics, from the birth of exploratory data analysis to the impacts of statistical graphics on practice today. We present examples of contemporary data visualizations in the process of exploring airline traffic, global standardized test scores, election monitoring, Wikipedia edits, the housing crisis as observed in San Francisco, and the mining of credit card databases. We provide a review of recent literature. Good data visualization yields better models and predictions and allows for the discovery of the unexpected.

#### 3.7.2 wk11: Glowing brains – Is one pretty picture worth a thousand serious thoughts?


**Abstract:** 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.


**Abstract:** We’ve all seen them, those colorful images that show how our brains “light up” when we’re in love, playing a video game, craving chocolate, etc. Created using functional magnetic resonance imaging, or fMRI, these pictures are the basis of tens of thousands of scientific papers, the backdrop to TED talks and supporting evidence in best-selling books that tell us how to maintain healthy relationships, make decisions, market products and lose weight. But a study published last month in the Proceedings of the National Academy of Sciences uncovered flaws in the software researchers rely on to analyze fMRI data. The glitch can cause false positives — suggesting brain activity where there is none — up to 70 percent of the time.

Abstract: The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data and a total of 3 million random task group analyses to compute empirical familywise error rates for the fMRI software packages SPM, FSL, and AFNI, as well as a nonparametric permutation method. For a nominal familywise error rate of 5%, the parametric statistical methods are shown to be conservative for voxelwise inference and invalid for clusterwise inference. Our results suggest that the principal cause of the invalid cluster inferences is spatial autocorrelation functions that do not follow the assumed Gaussian shape. By comparison, the nonparametric permutation test is found to produce nominal results for voxelwise as well as clusterwise inference. These findings speak to the need of validating the statistical methods being used in the field of neuroimaging.

3.7.3 Maybe more on this topic, maybe not . . .


Abstract: Brain images are believed to have a particularly persuasive influence on the public perception of research on cognition. Three experiments are reported showing that presenting brain images with articles summarizing cognitive neuroscience research resulted in higher ratings of scientific reasoning for arguments made in those articles, as compared to articles accompanied by bar graphs, a topographical map of brain activation, or no image. These data lend support to the notion that part of the fascination, and the credibility, of brain imaging research lies in the persuasive power of the actual brain images themselves. We argue that brain images are influential because they provide a physical basis for abstract cognitive processes, appealing to people’s affinity for reductionistic explanations of cognitive phenomena.


Abstract: Abstract A series of highly-cited experiments published in 2008 demonstrated a biasing effect of neuroimages on lay perceptions of scientific research. More recent work, however, has questioned this bias, particularly within legal contexts in which neuroscientific evidence is proffered by one of the parties. The present research moves away from the legal framework and describes five experiments that re-examine this effect. Experiments 1 through 4 present conceptual and direct replications of some of the original 2008 experiments, and find no evidence of a neuroimage bias. A fifth experiment is reported that confirms that, when laypeople are allowed multiple points of reference (e.g., when directly comparing neuroimaging to other graphical depictions of neurological data), a neuroimage bias can be observed. Together these results suggest that, under the right conditions, a neuroimage might be able to bias judgments of scientific information, but the scope of this effect may be limited to certain contexts.


Abstract: The persuasive power of brain images has captivated scholars in many disciplines. Like others, we too were intrigued by the finding that a brain image makes accompanying information
more credible (McCabe & Castel in Cognition 107:343-352, 2008). But when our attempts to build on this effect failed, we instead ran a series of systematic replications of the original study—comprising 10 experiments and nearly 2,000 subjects. When we combined the original data with ours in a meta-analysis, we arrived at a more precise estimate of the effect, determining that a brain image exerted little to no influence. The persistent meme of the influential brain image should be viewed with a critical eye.

3.8 wk12: Data Sharing


Abstract: Data sharing and access are venerable problems embedded in a rapidly changing milieu. Pressure points include the increasingly data-driven nature of science, the volume, complexity, and distributed nature of data, new concerns regarding privacy and confidentiality, and rising attention to reproducibility of research. In the context of research data, this review surveys extant technologies, articulates a number of identified and emerging issues, and outlines one path for the future. Recognizing that data availability is a public good, research data archives can provide economic and scientific value to both data generators and data consumers in a way that engenders trust. The overall framework is statistical—the use of data for inference.


3.9 wk13: Other stuff


Abstract: Poor research design and data analysis encourage false-positive findings. Such poor methods persist despite perennial calls for improvement, suggesting that they result from something more than just misunderstanding. The persistence of poor methods results partly from incentives that favour them, leading to the natural selection of bad science. This dynamic requires no conscious strategizing—no deliberate cheating nor loafing—by scientists, only that publication is a principal factor for career advancement. Some normative methods of analysis have almost certainly been selected to further publication instead of discovery. In order to improve the culture of science, a shift must be made away from correcting misunderstandings and towards rewarding understanding. We support this argument with empirical evidence and computational modelling. We first present a 60-year meta-analysis of statistical power in the behavioural sciences and show that power has not improved despite repeated demonstrations of the necessity of increasing power. To demonstrate the logical consequences of structural incentives, we then present a dynamic model of scientific communities in which competing laboratories investigate novel or previously published hypotheses using culturally transmitted research methods. As in the real world, successful labs produce more ‘progeny,’ such that their methods are more often copied and their students are more likely to start labs of their own. Selection for high output leads to poorer methods and increasingly high false discovery rates. We additionally show that replication slows but
does not stop the process of methodological deterioration. Improving the quality of research requires change at the institutional level.

3.10  NODS and More: Naturally Occurring Data Sets

3.10.1  wk13: Manifestos


**Abstract:** The cognitive revolution offered an alternative to merely analyzing human behavior, using the notion of computation to rigorously express hypotheses about the mind. Computation also gives us new tools for testing these hypotheses - large behavioral databases generated by human interactions with computers and with one another. This kind of data is typically analyzed by computer scientists, who focus on predicting people’s behavior based on their history. A new cognitive revolution is needed, demonstrating the value of minds as intervening variables in these analyses and using the results to evaluate models of human cognition. (C) 2014 Elsevier B.V. All rights reserved.


**Abstract:** The very expertise with which psychologists wield their tools for achieving laboratory control may have had the unwelcome effect of blinding psychologists to the possibilities of discovering principles of behavior without conducting experiments. When creatively interrogated, a diverse range of large, real-world data sets provides powerful diagnostic tools for revealing principles of human judgment, perception, categorization, decision making, language use, inference, problem solving, and representation. Examples of these data sets include patterns of web-site links, dictionaries, logs of group interactions, collections of images and image tags, text corpora, history of financial transactions, trends in twitter tag usage and propagation, patents, consumer product sales, performance in high-stakes sporting events, dialect maps, and scientific citations. The goal of this issue is to present some exemplary case studies of mining naturally existing data sets to reveal important principles and phenomena in cognitive science, and to discuss some of the underlying issues involved with conducting traditional experiments, analyses of naturally occurring data, computational modeling, and the synthesis of all three methods.

3.10.2  wk13: Dangers of Big Data


**Abstract:** The recent proliferation of digital databases of cultural and linguistic data, together with new statistical techniques becoming available has lead to a rise in so-called nomothetic studies [1-8]. These seek relationships between demographic variables and cultural traits from large, cross-cultural datasets. The insights from these studies are important for understanding how cultural traits evolve. While these studies are fascinating and are good at generating testable hypotheses, they may underestimate the probability of finding spurious correlations between cultural traits. Here we show that this kind of approach can find links between such unlikely cultural traits as traffic accidents, levels of extra-martial sex, political collectivism and linguistic diversity. This suggests that spurious correlations, due to historical descent, geographic diffusion or increased
Noise-to-signal ratios in large datasets, are much more likely than some studies admit. We suggest some criteria for the evaluation of nomothetic studies and some practical solutions to the problems. Since some of these studies are receiving media attention without a widespread understanding of the complexities of the issue, there is a risk that poorly controlled studies could affect policy. We hope to contribute towards a general skepticism for correlational studies by demonstrating the ease of finding apparently rigorous correlations between cultural traits. Despite this, we see well-controlled nomothetic studies as useful tools for the development of theories.

3.10.3 wk14: NODs – pick one, any one!


Abstract: We analyze naturally occurring datasets from student use of educational technologies to explore a long-standing question of the scope of transfer of learning. We contrast a faculty theory of broad transfer with a component theory of more constrained transfer. To test these theories, we develop statistical models of them. These models use latent variables to represent mental functions that are changed while learning to cause a reduction in error rates for new tasks. Strong versions of these models provide a common explanation for the variance in task difficulty and transfer. Weak versions decouple difficulty and transfer explanations by describing task difficulty with parameters for each unique task. We evaluate these models in terms of both their prediction accuracy on held-out data and their power in explaining task difficulty and learning transfer. In comparisons across eight datasets, we find that the component models provide both better predictions and better explanations than the faculty models. Weak model variations tend to improve generalization across students, but hurt generalization across items and make a sacrifice to explanatory power. More generally, the approach could be used to identify malleable components of cognitive functions, such as spatial reasoning or executive functions.


Abstract: The growing availability of large datasets in a variety of domains presents an opportunity for researchers to use field data to better understand psychological concepts. I discuss, from an empirical economics point of view, steps for how to study cognition in large datasets. I use two recent papers that explore psychology in the used-car market as motivating examples. These examples help illustrate the potential importance of big data as a way to explore human psychology and cognition.


Abstract: When making a decision, humans consider two types of information: information they have acquired through their prior experience of the world, and further information they gather to support the decision in question. Here, we present evidence that data from search engines such as Google can help us model both sources of information. We show that statistics from search engines on the frequency of content on the Internet can help us estimate the statistical structure of prior experience; and, specifically, we outline how such statistics can inform psychological theories concerning the valuation of human lives, or choices involving delayed outcomes. Turning to information gathering, we show that search query data might help measure human information gathering, and it may predict subsequent decisions. Such data enable us to compare information...
gathered across nations, where analyses suggest, for example, a greater focus on the future in countries with a higher per capita GDP. We conclude that search engine data constitute a valuable new resource for cognitive scientists, offering a fascinating new tool for understanding the human decision-making process.


**Abstract:** This research addressed theoretical approaches in political science arguing that the American electorate is either poorly informed or dependent on party label cues, by assessing performance on political judgment tasks when party label information is missing. The research materials were created from the results of a national opinion survey held during a national election. The experiments themselves were run on nationally representative samples of adults, identified from another national electoral survey. Participants saw profiles of simulated individuals, including information about demographics and issue positions, but omitting party labels. In Experiment 1, participants successfully judged the likelihood of party membership based on the profiles. In Experiment 2, participants successfully voted based on their party interests. The results were mediated by participants’ political knowledge. Conclusions are drawn with respect to theories from political science and issues in cognitive science regarding categorization and reasoning.


**Abstract:** Most cognitive psychology experiments evaluate models of human cognition using a relatively small, well-controlled set of stimuli. This approach stands in contrast to current work in neuroscience, perception, and computer vision, which have begun to focus on using large databases of natural images. We argue that natural images provide a powerful tool for characterizing the statistical environment in which people operate, for better evaluating psychological theories, and for bringing the insights of cognitive science closer to real applications. We discuss how some of the challenges of using natural images as stimuli in experiments can be addressed through increased sample sizes, using representations from computer vision, and developing new experimental methods. Finally, we illustrate these points by summarizing recent work using large image databases to explore questions about human cognition in four different domains: modeling subjective randomness, defining a quantitative measure of representativeness, identifying prior knowledge used in word learning, and determining the structure of natural categories.


**Abstract:** Psychologists have used experimental methods to study language for more than a century. However, only with the recent availability of large-scale linguistic databases has a more complete picture begun to emerge of how language is actually used, and what information is available as input to language acquisition. Analyses of such “big data” have resulted in reappraisals of key assumptions about the nature of language. As an example, we focus on corpus-based research that has shed new light on the arbitrariness of the sign: the longstanding assumption that the relationship between the sound of a word and its meaning is arbitrary. The results reveal a systematic relationship between the sound of a word and its meaning, which is stronger for early acquired words. Moreover, the analyses further uncover a systematic relationship between words and their lexical categories — nouns and verbs sound differently from each other — affecting how we learn
new words and use them in sentences. Together, these results point to a division of labor between arbitrariness and systematicity in sound-meaning mappings. We conclude by arguing in favor of including “big data” analyses into the language scientist’s methodological toolbox.


**Abstract:** Options are often presented incidentally in a sequence, but does serial position impact choice after delay, and if so, how? We address this question in a consequential real-world choice domain. Using 25 years of citation data, and a unique identification strategy, we examine the relationship between article order (i.e., position in a journal issue) and citation count. Results indicate that mere serial position affects the prominence that research achieves: Earlier-listed articles receive more citations. Furthermore, our identification strategy allows us to cast doubt on alternative explanations (i.e., editorial placement) and instead indicate that the effect is driven by psychological processes of attention and memory. These findings deepen the understanding of how presentation order impacts choice, suggest that subtle presentation factors can bias an important scientific metric, and shed light on how psychological processes shape collective outcomes.


**Abstract:** How many words—and which ones—are sufficient to define all other words? When dictionaries are analyzed as directed graphs with links from defining words to defined words, they reveal a latent structure. Recursively removing all words that are reachable by definition but that do not define any further words reduces the dictionary to a Kernel of about 10% of its size. This is still not the smallest number of words that can define all the rest. About 75% of the Kernel turns out to be its Core, a “Strongly Connected Subset” of words with a definitional path to and from any pair of its words and no word’s definition depending on a word outside the set. But the Core cannot define all the rest of the dictionary. The 25% of the Kernel surrounding the Core consists of small strongly connected subsets of words: the Satellites. The size of the smallest set of words that can define all the rest—the graph’s “minimum feedback vertex set” or MinSet—is about 1% of the dictionary, about 15% of the Kernel, and part-Core/part-Satellite. But every dictionary has a huge number of MinSets. The Core words are learned earlier, more frequent, and less concrete than the Satellites, which are in turn learned earlier, more frequent, but more concrete than the rest of the Dictionary. In principle, only one MinSet’s words would need to be grounded through the sensorimotor capacity to recognize and categorize their referents. In a dual-code sensorimotor/symbolic model of the mental lexicon, the symbolic code could do all the rest through recombinatory definition.

### 4 REQUIREMENTS

- **Group discussion.**
  
  - The most important contribution each person can make is to our discussions of the readings. I believe a seminar course in which everyone actively participates can be the most productive and educational forum in grad school (often for the instructor as well). Bringing together the various backgrounds and training of everyone in the room generally leads to a much richer perspective than would otherwise be possible. There is a lot of individual variability in tendency to speak up in this type of envi-
rionment, but it is critical to an academic career to be comfortable doing so. You cannot succeed in this field without a willingness (and desire) to share your ideas in the face of criticism, and this is perhaps the best context to practice. If you are someone who has no qualms about dominating a debate, this is also a good place to practice restraint and listening.

- **Written reactions.**

  - Each person should bring to each class a brief written reaction to the readings to be discussed. You can email your reactions to me before class, or you can bring a printed copy, but either way they must be complete before class begins. The reactions serve two purposes: as a nominal motivation to ensure everyone reads and carefully thinks about the articles, and as a starting point for the group discussion. Reactions should not be summaries. A few sentences at the beginning to summarize each article are generally useful, both for me to make sure everyone recognizes the critical points and for you to check your own understanding, but the primary content should be your own ideas in response to what you read. These ideas can be anything from connections to other research (from this class or elsewhere); to possible extensions, improvements, or follow-up work; to criticisms of the authors logic or methods.

- **Preview presentations.**

  - Each person will sign up to present a preview summary of next week’s readings at the end of this week’s class. We will rotate this responsibility among ourselves and spread it around so that for weeks when multiple readings are assigned, the readings will be divided among multiple people. In general, each presentation should be 8-12 minutes long and structured as though you were presenting your own work at a conference. A useful strategy is to copy key figures and tables out of each article and supplement with (scant) text stating the major points. Focus on summarizing the research, as the authors present it (including motivation, background, methods, results, and conclusions), and save your own reactions for the following week. You will also be in charge of getting the discussion going and answering any clarifying questions people may have after reading the articles themselves. Written reactions are not required for papers you present.

4.1 **PreRequisites**

Permission of the instructor. This is a graduate research seminar in the Cognitive Science Department. However, all interested undergraduates and interested graduate students from other departments are encouraged to contact the instructor to discuss their participation in the seminar. Responsibilities and assignments for undergraduates will be discussed and agreed on, in writing, by the student and the instructor.

4.2 **About the Instructor**

Professor Gray has been a member of the Cognitive Science Department at RPI since the Fall of 2002. For details on his research interests and activities see his
5 Honors Policy

- My expectation is that all of the work you do for me in this class will be the work of one individual. Exceptions to this rule will be broadcast to the class by email.

- As you will all find out, I explicitly encourage you to engage in public (using email and other media to broadcast a message to the entire) or private (one-to-one) discourse regarding the readings and topics raised in this class. Study groups are encouraged.

- If any of you have any questions regarding current situations or future situations, remember that I am your first contact on this. Please come and see me.

6 Grading Policy

- Examinations – none

- Group Discussion
  - 45% for active participation in all discussions on all weeks in which the seminar is held. Exceptions due to professional travel or other activities need to be discussed with the instructor ahead of time.

- Written Reactions
  - 20% Due before class, preferably the night before but definitely before class starts.

- Presentations: Leading the discussions and summarizing the readings
  - 40%

- Yes. I expect 105% out of you!

7 References


r=0


This document contains 56 references.