Breaking Out

2018-11-16

Breaking Out of Tradition: Statistical Innovation in Psychological Science

S.N. MacEachern[1] & T. Van Zandt[2]

[1]Department of Statistics, The Ohio State University

[2]Department of Psychology, The Ohio State University

Friday, 16 November 2018

Supported by NSF awards SES-1424481 and DMS-1613110.

Thank you, Rich, for inviting me to speak at this symposium. For those of you who don't know him, Steve MacEachern, my co-author, is a my colleague and chair of the Department of Statistics. He and I have discussed the application of Bayesian methods to behavioral data for well over a decade, and we have lots of strong opinions on the topic.

Outline

Outline

Our talk is not very technical, and does not contain many surprising remarks. I'd like to spend a few minutes talking about the question that is the focus of the symposium, then I want to redirect our conversation to the real problem: the replication crisis. Then I'm going to discuss some statistical "innovations" (and you make disagree about the extent to which these things are innovative) that we believe serve as examples of good statistical practice.

Conclusions

Conclusions

- A black-box approach is guaranteed to produce poor-quality conclusions.
- Exploratory data analysis is necessary for good modeling and should be encouraged.
- Behavioral data are messy, and hierarchical models are the best way to explain individual differences.

For the most part I think I will be preaching to the choir. When I'm done, my conclusions, again not surprising, will be the following.

> *First, traditional statistical methods as applied in psychology have been based on black-box approaches. Black-box approaches, which emphasize asymptotic properties of tests and bright-line criteria for significance, will not yield good conclusions, will certainly not tell us what the truth is.*
>
> *Second, exploratory data analysis is absolutely critical for model development and experimental design. What is also critical is that such analyses are performed in a completely transparent fashion; preregistration is not a guarantee of transparency.*
>
> *Finally, behavioral data are messy. Individual differences are endemic, nonstationarity is pervasive. The best way to deal with data of this kind is to use hierarchical models.*

Fisher was a Geneticist

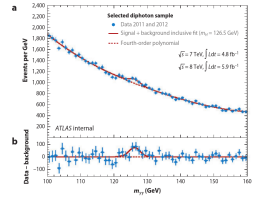


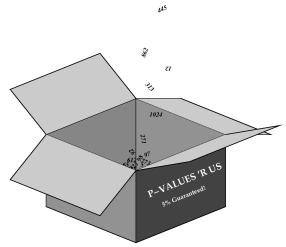
Figure 1
A subset of the 2011 and 2012 ATLAS event counts. (*a*) The observed event counts in the $\gamma\gamma$ decay channel in each invariant mass bin. (*b*) Residuals from the fitted background model. The excess counts with invariant mass near 125 GeV are apparent. The background models are doussd in Section 3, and the source model is based on the Standard Model with Higgs mass (m_H) of 126.5 GeV. The quantity $\sqrt{s}$ is the energy of the collider, and $\int Ldt$ is the volume of data at each energy used in the plot. Copyright CERN.

I'm going to start by complaining about the question on which our symposium is based, "Should Statistics Determine the Practice of Science, or Science Determine the Practice of Statistics?" The answer to the question is, "Neither." The question itself is wrong-headed, as it implicitly assumes that the practice of statistics is not scientific, and by implication that statisticians are not scientists. Yet nothing could be farther from the truth.

For centuries, statisticians have been making scientific contributions, including Fisher's contributions to genetics and evolutionary biology, and modern statisticians' contributions to global climate change, earthquake prediction, and tracking terrorist activity around the globe. The development of methods for data analysis and experimental design is a scientific enterprise, and the relationship between non-statistician scientists and statisticians has and will always be synergistic. The discovery of the Higgs boson, for example, required a close collaboration between particle physicists and statisticians, and both "kinds" of researchers were well-versed in the theory and models that drove their experiments and their data analysis. The statistical problems that the research team had to address were novel and complex, and a variety of strategies were employed to verify that the data in fact demonstrated the existence of the Higgs boson. These strategies included the addition of a random data shift that was not revealed until after the analyses were complete, a Poisson count model with nonstationary rate parameters, and both frequentist and Bayesian inference. To argue that one scientific discipline, like particle physics, can or should dictate terms to another, like statistics, is nonsense. So let's move on to the issue at hand: How should data be analyzed to ameliorate the replication crisis?

Breaking Out
The Problem
The Black-Box Approach

The Black-Box Approach



If the replication crisis exists, which is open for debate, what can be done to fix it? Today I would like to talk about the misconceptions about statistical inquiry that we believe have led to the crisis, and provide some concrete suggestions for improvement.

> *The central point of our argument is that the use of statistical black boxes does not lead to good conclusions, and in our opinion can be blamed for the current crisis. Data analysis must be informed by theory, and different experiments will lead to different data sets, each of which should be examined for contaminants, unexpected results, and errors.*
>
> *As statistics educators and quantitative researchers, we have all fallen into the trap of thinking that the methods that we have developed can be applied (perhaps blindly) to other situations: The hierarchical model that I develop for a recognition memory task can be used to explain data from a categorization task. The t-test can be applied in any situation where differences can be assumed to be normally (or symmetrically) distributed.*
>
> *Every statistical analysis is performed on the basis of a set of assumptions about the data that may or may not be true; the outcome of an analysis will depend on the assumptions that you carry into it. It is our opinion that the replication crisis has arisen from black-box techniques, often, unfortunately, applied by researchers with poor statistical insight, in the absence of a model, and without performing exploratory data analysis.*

Breaking Out
The Problem
A Recognition Memory Study: Lab A

A Recognition Memory Study: Lab A

Ten participants studied lists of words in two conditions (hard and easy). They were then tested with 20 words, 10 old and 10 new.

	Condition	
	Hard	Easy
Mean d'	2.05 ± 0.35	2.58 ± 0.26

$t(9) = 1.6658,\ p > 0.05$

Pitt & Van Zandt, in preparation

To make the problem concrete, consider the data in this table, which may have been collected in a recognition memory experiment. It was, as you can see, a small experiment, average performance under both easy and hard conditions was pretty good, but the test statistic did not reach significance at an $\alpha = 0.05$ level. Lab A researchers pitched the data into their file drawer, went home and cried.

Breaking Out
The Problem
A Recognition Memory Study: Lab B

A Recognition Memory Study: Lab B

Ten participants studied lists of words in two conditions (hard and easy). They were then tested with 20 words, 10 old and 10 new.

	Condition	
	Hard	Easy
Mean d'	2.05 ± 0.35	2.64 ± 0.25

$t(9) = 1.8855,\ p < 0.05$

Pitt & Van Zandt, in preparation

Lab B ran a very similar experiment, under the same constraints, and got similar results, except that performance in the Easy condition was slightly higher than what Lab A observed. Now the test statistic reached significance, and Lab B researchers popped open the champagne and started working on the first draft of their manuscript.

A closer look

	Hard, Labs A & B			Easy, Lab A			Easy, Lab B		
	Hits	FAs	d'	Hits	FAs	d'	Hits	FAs	d'
1	7	5	0.47	7	2	1.22	7	2	1.22
2	8	2	1.50	10	0	3.38	10	0	3.38
3	9	1	2.19	10	2	2.44	10	2	2.44
4	10	2	2.44	8	2	1.50	8	2	1.50
5	9	1	2.19	9	1	2.19	9	0	2.79
6	10	0	3.38	10	1	2.79	10	1	2.79
7	5	5	0.00	7	0	2.16	7	0	2.16
8	10	0	3.38	10	0	3.38	10	0	3.38
9	9	0	2.79	10	0	3.38	10	0	3.38
10	9	1	2.19	10	0	3.38	10	0	3.38
	Mean d'		2.05			2.58			2.64

Snodgrass & Corwin (1988) correction applied.

But consider the raw data. The numbers were poured into the *t*-test black box and the judgment of the presence or absence of an effect was made without looking at the numbers.

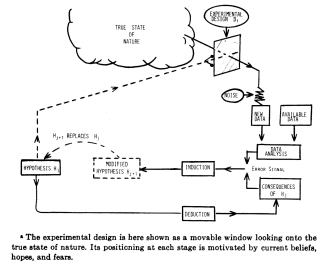
We can see that what happened in the two labs was very similar. The data are so similar that it's hard to see where they might be different.The only difference is in Participant 5, and consists of a single response: One response that was a false alarm in Lab A was a correct rejection in Lab B. One single response was sufficient to make the difference between the conditions "significant" by the use of a fixed threshold.

Problems

- No thought given to:
  - models that provide process/theoretical explanations;
  - what constitutes qualitative (vs. quantitative) differences over conditions;
  - whether effects are interesting.
  - the assumptions that are brought into the analysis.
- Any threshold/"bright line" approach to determining what is true or false is ludicrous.

I want to emphasize several things about this example. First, it is the application of a black-box approach with no thought of looking at the data that leads to this peculiar situation. There is no thoughtful model applied to the data, there is no attempt to describe what an interesting or qualitative difference between the conditions might be, there is no examination of the assumptions required for the analysis in question, and finally, the $p < 0.05$ "bright line" criterion is simply not how science should determine whether a particular model is true or false.

Box (1976)



* The experimental design is here shown as a movable window looking onto the true state of nature. Its positioning at each stage is motivated by current beliefs, hopes, and fears.

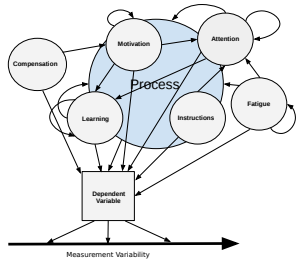
In 1976, George Box described the advancement of learning as "motivated iteration" between theory and practice. This rather janky figure appeared in his 1976 paper, "Science and Statistics," and depicts his view of the advancement of knowledge. Box shows the critical role played by data analysis in the construction of hypotheses or models, the exploratory phase of discovery, followed by the equally critical confirmatory phase, in which model predictions are contrasted with the results of new data analysis. It is important to notice two things: first, the data analysis occurs before the model is constructed or modified, and second, the current iteration of the model determines the location of the "window on truth," the experimental design, that determines the data we will obtain next.

Thoughtful Statistics

- Build models that are based on theory;
- Make use of exploratory data analysis;
- Draw conservative conclusions.
- Acknowledge that data are messy;

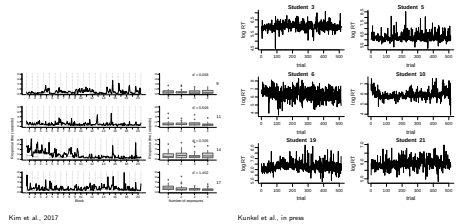
We can see from Box's figure his conception of thoughtful statistics. Generalizing from what we do not see in this figure, poor statistical practice makes heavy use of black boxes with no theory and bright-line criteria. Good practice involves theoretically motivated analysis: the development of models and the application of statistical techniques that are appropriate to those models.

> *Good model development in turn requires exploratory data analysis. T-tests can be run on csv files of numbers without ever considering the features hidden in those numbers. I had someone in my office not three days ago asking for help with an analysis in which the p-values were nowhere less than 0.05. When I suggested that he plot his data and look at it before running blind analyses he was surprised that I would suggest such a thing, and then he was embarassed that he hadn't thought of it himself. How can a reasonable model of the data be constructed if we don't know what the data are telling us, and if we don't know what a good model is, how do we know how to interpret our results?*
>
> *With a reasonable model in hand, conclusions drawn from theoretically motivated statistical analyses should be conservative. The p-value is only one piece of the puzzle, which is influenced by the assumptions on which a particular procedure is based and the robustness of the procedure to violations of those assumptions.*
>
> *Finally, and independently from Box's opinions on the matter, those of us who deal with behavioral measures must acknowledge that behavioral data are messy. The observations we obtain from participants in an experiment in which we ask those participants to repeat the performance of a task many many times are not independent and identical. The observations are not even exchangeable, because the way they perform the task is changing over time. This lack of exchangeability has profound implications for the conclusions we can draw from any black-box analysis or model-fitting exercise.*



It is this messiness that has preoccupied my research efforts over the past several years. We are trying to study a process, but that process operates in an impossibly complex reality. The data we observe are mixtures of measurements from that process, as well as measurements that are distorted or corrupted by other mental or physiological processes, and contaminants.

I know of one participant who confessed to me that she simply pressed keys as quickly as she could during the performance of her task because she had to pee so badly she couldn't concentrate. I could have performed an ANOVA or fit the diffusion model to her data in a black-box approach without ever looking at the time series of her responses and never have been the wiser. The approach we take to analyzing behavioral data has to acknowledge incidents like these, and our models have to be complex enough to accommodate (if not explain) the unexpected. Such model construction requires intimate familiarity with the data; this in turn again requires that we spend a little time doing exploratory analysis.



I present here some RT data reported in two recent papers published by our research group. The left panel shows four RT series from four participants performing a recognition memory task, the right shows 6 series from six participants performing a modified go/no-go task.
You can see a number of problematic features of the data right away without performing any statistical analyses at all.

> *First, there are large individual differences. Some participants show steadily decreasing RTs, while others do not. The variability of the RTs changes over time and blocks in different idiosyncratic ways. Analyses of these series show local dependence from trial to trial.*
>
> *The exploratory analysis we performed on these data sets told us that no one process was sufficient to explain the data. Each individual's data was modeled as a mixture of fast, slow, and process-associated observations.*
>
> *My point is that data are messy regardless of how spotless your design or well-fitting your model is. By fitting more complex, hierarchical mixture models, we were able to characterize individual performance in ways that were previously not possible. Data that would have been discarded as contaminated were preserved, and we were able to extract information from them. We have the computing power now to incorporate realistic contaminant processes within a hierarchical design that can account for individual differences. Bayesian hierarchical modeling is the easiest (though not the only) way to do this, but there is never going to be a black box that permits you to take our model and apply it to any other data set without first looking at the data to see if it makes sense to do so.*

Transparency!

- Preregistration of studies is not enough: planned/preregistered analyses *should* change after the researcher examines the data.
- All raw data and code should be included as supplements at publication.

I have emphasized exploratory analysis throughout my remarks today. You are all aware that there is great pressure today to preregister your studies and state, in advance, what analyses you will perform and what models you will test. While preregistration of experimental procedures may make some sense, we express here our concerns that, while well-intentioned, preregistration of analyses or models will not serve any useful purpose.

As Box described, advancement of learning proceeds in an iterative fashion as we move from exploratory to confirmatory analysis. Because exploratory analysis is such a fundamental component of model construction, the analyses that you perform SHOULD change after you collect and examine your data. The model SHOULD change as a function of the regular features of the data that you observe. Instead of preregistration of analyses, we believe that all raw data, code, and an explanation of exploratory analyses should be published as online supplements to empirical papers.

Things We Should See More Of

- Cross validation
- Model averaging
- Robust methods/Bayesian sensitivity analysis
- Nonparametric Bayesian methods

With regard to statistical innovations, I would like to mention but a few. One is cross-validation, the evaluation of a model on the basis of its ability to fit a set of hold-out data, probably the most powerful method for evaluating goodness-of-fit. The second is the use of model averaging. Model building is a subjective process, and different researchers address problems in different ways. Model averaging is a theoretically motivated approach that provides a coherent way to include uncertainty about different potential models in an analysis, and prevents overconfidence in one model to the exclusion of others.
Bayesian sensitivity analysis is something we don't see enough of. When there is uncertainty about the priors or even the likelihood, we should be evaluating our conclusions based on model fits that use a range of priors (and likelihoods). If the analysis is robust, the conclusions will not depend much on which prior or likelihood we use.

Finally, nonparametric Bayesian modeling, which makes use of Dirichlet processes, bypasses the question of evaluating several models by

using a single model that increases in complexity as more data are observed.

Conclusions

- A black-box approach will never be able to provide the insight we are looking for.
- Exploratory data analysis is necessary for good modeling and should be encouraged.
- Behavioral data are messy, and hierarchical models are the best way to explain individual differences.

So now I've said what I wanted to say, and here again are the points I want to emphasize. First, traditional statistical methods as applied in psychology have been based on black-box approaches. T-tests, ANOVAs, χ^2 tests, Kruskal-Wallace when we want to be risque. The assumptions that are carried into these analyses are rarely tested or questioned, and we rely on the proven asymptotic properties of statistical tests to save us from the bother of looking closely at the data. This black-box approach, which emphasizes asymptotic properties of tests and bright-line criteria for significance, will not yield good conclusions, will certainly not tell us what the truth is. Second, there is a tendency today to view exploratory data analysis with suspicion, and this is a mistake. Exploratory data analysis is absolutely critical for model development and experimental design. What is also critical is that such analyses are performed in a completely transparent fashion; preregistration is not a guarantee of transparency, and I fear that it will instead hinder transparency. People will see immediately that the planned and registered analyses are inappropriate after looking at the data and will inevitably experience cognitive dissonance as a result.

Finally, behavioral data are messy. Individual differences are endemic, nonstationarity is pervasive. The best way to deal with data of this kind is to use hierarchical models of sufficient complexity that "bad data" can be explained along with the good. The models don't need to be Bayesian, but the Bayesian approach is the easiest and the most intuitive. It seems to me that this modeling approach is becoming more of the norm rather than the exception. Let's do all we can to make sure that our community comes to appreciate these kinds of models and that they are equipped with the skills to deploy them.

Thank you.

Thank you for your attention.

Tightening Criteria ($p < .005$)

Benjamin et al. (2017)

I have also mentioned that good statistical analyses draw conservative conclusions. Some of you may have read a recent paper, which I signed, that advocated that the criterion for significance be reduced from 0.05 to 0.005 for new discoveries. I have also stated that bright-line criteria for determining truth are ludicrous, whether those criteria are based on p-values or Bayes factors.
How do I reconcile these two sentiments? As I explained recently in a different talk, my audience was not you, was not experimental psychologists more generally, but was instead the science reporters at the New York Times. My hope is that our declaration, bolstered by the demonstration that p-values less than 0.05 provide only weak evidence for the alternative hypothesis, would be noticed by journalists. If journalists take note, then the next time someone tries to claim that subliminal exposure of the American flag makes people more likely to vote for republican candidates the claim will be scrutinized more carefully before it is written up for consumption, later to stand as a permanent reminder to the public that psychology is not yet a "real" science.

As I then explained, reducing the p-value for significance to 0.005 is not a scalpel in the hands of a skilled surgeon operating to remove a tumor from our body of science. Instead, it is a towel that we can use, hopefully only temporarily, to staunch the bleeding.