

# MODERN APPROACHES TO DATA ANALYSIS

Introduction

# Basic Approach to a Scientific Understanding of Human Behavior

- Start with theory based on previous research, intuition, common sense, immediate needs etc.
- State the theory in terms of a *testable* model(s)
  - ▣ Predictor(s) and Outcome(s)
- Sufficiently define what is to be measured and find good measurement of the variables of interest
- Collect information regarding behavior of interest measured as operationally defined previously
- Examine the worth of the models
  - ▣ Compare model fit
- Make predictions regarding future behavior and reassess model/theory
- Replicate



# An (unfortunately) Common Approach

- ❑ Start with theory based on previous research, intuition, common sense, immediate needs etc.
- ❑ State the theory in terms of a lone testable model
- ❑ Largely ignore precision of measurement
  - ❑ Reliability not reported or reported only for initial development of the measure not the current study.
- ❑ Collect poor data as a result, insufficient as well (low N)
- ❑ Use dated methods to assess model fit in comparison to a null model no one believes in or is otherwise useless
- ❑ Ignore model fit if it doesn't fit well and focus entirely on explanation (mere description) and statistical significance rather than prediction (science) and practical effect
- ❑ Ignore validation and adequate exploration of what the data has to offer
- ❑ The results: are conflicting outcomes, few solid conclusions<sup>1</sup>, fad theories, few predictions made and that hold up, and in general slow progress.

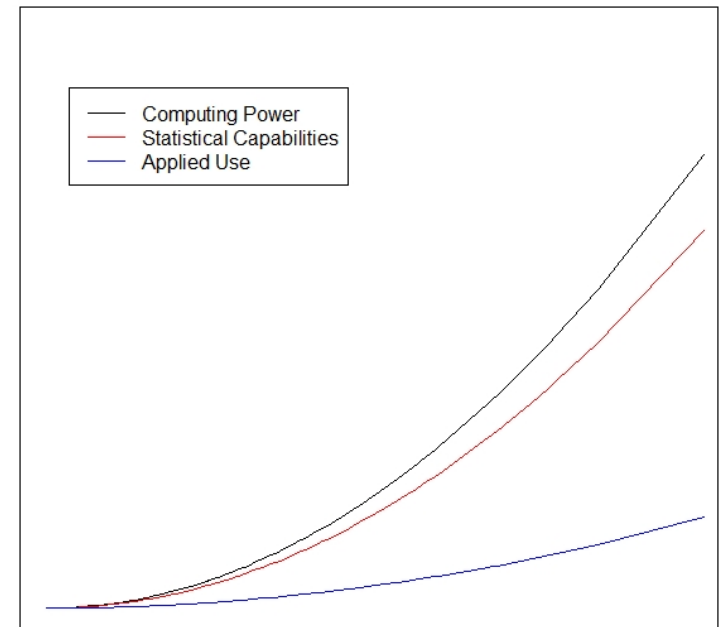
Example: Featured article on [APA journals website](#)- It's the Thought That Counts: The Role of Hostile Cognition in Shaping Aggressive Responses to Social Exclusion, *Journal of Personality and Social Psychology*, Jan 2009, a top 10 journal in impact factor (if review journals are not considered, top 20 even still). APA thought this warranted a press release.

- “The present studies were designed to test the hypotheses that rejection or social exclusion instills a broad inclination to perceive hostility in the social environment and that this tendency in turn increases aggressive behavior”
- “Experiments 1a and 1b tested the hypothesis that social rejection creates a hostile cognitive bias.”
  - No mention of quality of measurements
    - Brief Mood Introspection Scale, which they use/cite and which the author has the original paper on their website along with other resources and tips on how to make it more reliable (advertised as a free-ware instrument)
    - It actually was shown to have good reliability with a similar sample (college students), so this would have strengthened the current paper to report
  - Analysis consisted of two t-tests (though conducted as ANOVA), no effect size/prediction error reported (it would have helped their case), no mention of assumptions met
  - Experiment 1a had 33, 1b 45 participants, despite all authors being from large state schools (it is not stated which of the 3 the participants came from)
- Experiment 2
  - N of 30
  - Reported the reliability for their own rating statements
  - Post hoc a priori contrasts: Odd mix in which the regular ANOVA was conducted then followed by ‘planned’ contrasts on all possible pairwise comparisons (you can call it a priori if you want, but that’s a post hoc procedure if I ever saw one)
    - Note that if you conduct a planned contrast, one of the primary reasons for doing so aside from theory is increased power relative to the omnibus F. You do it in lieu of the regular ANOVA, not in post hoc fashion.
  - Despite previous (and actually strong) results in Experiment 1 regarding BMIS, null hypotheses were a value of zero.
  - No assumptions reported nor effect sizes. Given the small Ns and no mention of whether assumptions have been met, you can guarantee anyone attempting to replicate this would not find the same results.
  - Continues on with overly simple models and tests a 3 variable mediation model, which is the quoted hypothesis above. It is unclear how ‘Social Exclusion’ is represented. In the previous Experiment 2 models/analyses (which along with Exp 1 are now rendered misspecified since this is the real model of interest), SE was a 3 group random assignment. This cannot be represented as a single variable in a regression ( $k \text{ groups} - 1 = 2$  coded predictors, not 1), unless recoded to only be two groups, excluded vs. not, or theoretical reasoning suggest at least an ordinal relation among treatment conditions.
- Experiment 3 & 4
  - Continued trend of small N, no effect sizes, no assumption testing, including another scale w/o reporting reliability, post hoc ‘planned’ comparisons, no use of prior results (in one case almost exact replication) to inform present study
- Conclusion is social rejection → aggression in college undergrads, and possibly in indirect fashion. Given the approach above, it would be hazardous at best to generalize any specific details from the study, as almost all problems/issues listed on the previous slide are seen here.

# Statistical Computing

- Statistical analysis *is* computation and as such has exploded in what's available to us now alongside the rise in computing power
- The same trend cannot be seen in the actual use of this technology in psychological research, though this appears to be changing<sup>1</sup>
- Any adequate course in methods in the 21<sup>st</sup> century has to have statistical computing as a focus<sup>2</sup>
  - This does not mean magically obtaining output from a program and interpreting in rulebook fashion
    - “If  $p < .05$  then I say ‘Important’.”

Discrepancy between use and capability



Years since 1950

# Consider

- Is this computer viable for your needs?
  - ▣ Apple IIe from 1983
- Is this car something you'd prefer to drive today?
  - ▣ AMC Gremlin 1970s
- Would you use this for your portable music needs?
- How many movies do you watch on this?
  - ▣ Betamax





# Technology

- Anything based on technological advances will change, and usually rapidly
- Modern statistical analysis obviously relies on computers<sup>1</sup>, and today's desktops can perform thousands of statistical calculations in seconds
- Modern techniques take advantage of today's computing power resulting in more accurate estimates and better prediction, and are better able to handle the complexities of the things we're interested in
- In short, they allow for a better model of reality



# Modern Approaches

- Modern problems require modern approaches to dealing with them
- We don't *think* like we used to, and our (available) methods reflect that obvious fact
- Modern approaches to data analysis entail:
  - Old techniques (if appropriate- 'if it ain't broke...')
  - New techniques (largely)
    - More flexible approaches e.g. bootstrap
    - More robust approaches of old techniques e.g. robust regression
    - Not forcing linear models on data to which it is not appropriate
  - Discipline blurring problem solving
  - Dealing with data problems rather than ignoring them (testing assumptions, estimating missing values etc.)
  - Clearly displaying uncertainty via interval estimates
  - A focus on the actual effect/fit as opposed to statistical significance
  - A focus on exploration but not without subsequent validation with new information
    - Modeling reality as opposed to forcing reality into our theory
  - Updated graphical exploration<sup>1</sup>
  - *Statistically* adjusting models based on previous research
  - Comparing multiple meaningful models instead of comparing one model to a completely ignorant one ('Nil' hypothesis testing)
  - Works that assuming an interaction with the consumer/interested party (e.g. via the web) as opposed to passive absorption (via printed book/journal)
  - Etc.

# Modern Techniques: Examples

- Bootstrapped estimates of coefficients
- Robust regression
- Bayesian methods
- Multilevel models, Growth curve modeling etc.
- Neural networks
- Tree classification/regression
- Model averaging
- Boosting
- Updated factor analytic/principal component/ SEM techniques



# State of Affairs

- Gist: it's extremely difficult to keep up with what's available for those that look at methods everyday, and even more so for the applied researcher
- This was not the case many decades ago when knowledge growth was slower and lasted longer. But that era started to fade before most if not all of us in this course were born.
- We should not be doing something new just for the heck of it<sup>1</sup>
- We should however be keeping up with the developments of the last 50 years rather than ignoring them
- Times change, and we must deal with it.



# Outline<sup>1</sup>

- More focus on prediction and replication
- Bootstrapping
- Robust
- Nonlinear approaches
- Model comparison
- Bayesian
- Context matters: multilevel modeling

# Focus on prediction

- This is mostly just a 'new' thing for social sciences that have been almost exclusively on the explanation side of the explanation-prediction continuum
  - In contrast, physical sciences seem more on the prediction end
- As a result many articles spend a lot of time about things like variable importance when the overall model predicts horribly
  - Having a statistically significant predictor for a model that doesn't account for 10% of the variance in the DV doesn't do much to progress a discipline
- However, having 'good' models with no validation doesn't do a whole lot either
  - Had the previous model been validated, one would have seen that  $R^2$  drop even lower
    - It has to, initial regression output is overfitted
- Furthermore, the explanation (variable importance) is almost never tested statistically, and could easily and dramatically change with a new sample
  - Just because a standardized coefficient is larger than another, doesn't mean it will stay that way with a new sample
- Part of this is related to the NHST issue of overemphasis on statistical significance
  - You could talk about anything as long as it was statistically significant
- However, adding more emphasis on prediction coupled with better explanatory procedures will result in better science

Results and Conclusion Focus

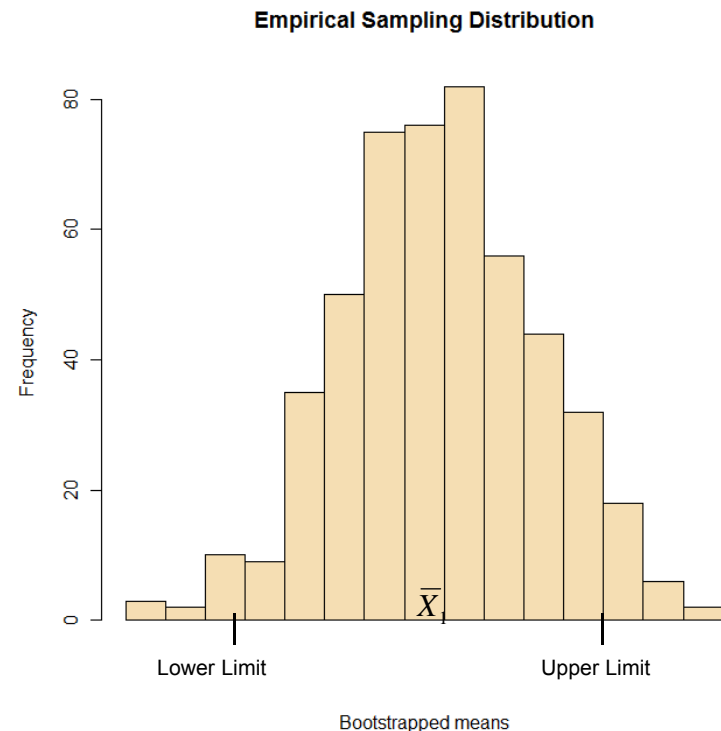
---

Explanation

Prediction

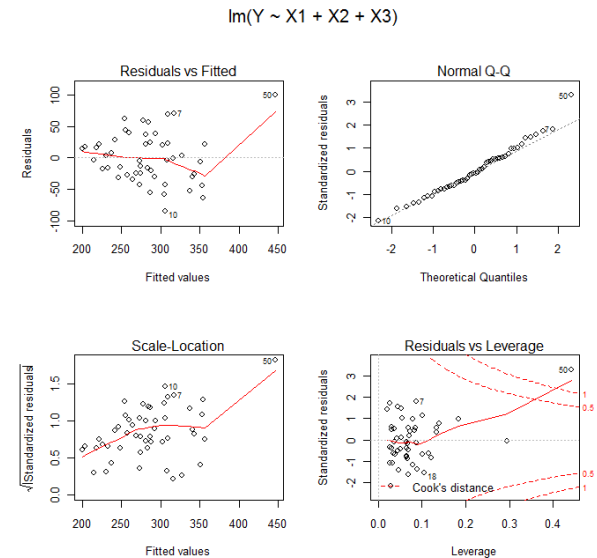
# Bootstrapping

- Very simple process of randomly sampling from the data set to create an empirical rather than theoretical sampling distribution for a particular statistic
- Purpose:
  - To give more accurate interval estimates when distributional assumptions do not hold (nonparametric procedure)
  - To give interval estimates to statistics where the standard error is otherwise undetermined
    - Often the case with modern statistics
- Perks:
  - Flexibility with problematic data
  - Often as good/accurate even when assumptions hold
  - Allow for statistical testing of *anything*
- Demo [Link](#)



# Robust

- Outliers and heteroscedasticity are common in research.
- They are detected:
  - Graphically: residuals vs. fitted values plot, influence plot etc.
  - Numerically: Cook's distance, Breusch-Pagan test for heteroscedasticity etc.
- In such cases, statistics may be biased, inefficient and are on the whole inaccurate and misleading
- Robust procedures are typically as easy to use and will be more accurate if there are problems, and should be checked regularly as a comparison 'just in case'



# Robust

- Unfortunately much software has not kept up with the times<sup>1</sup>, but good software can do it pretty easily allowing for incorporation into any analysis
- Example R code and output of a robust regression. The function uses a sophisticated procedure to downweight cases the more extreme they are in terms of the model.
- However understanding the concept, downweighting extreme cases, is the all that's really necessary to begin using it.
  - The default settings are the most generally applicable but one may tweak as they become more familiar with the technique
- Everything is interpreted the same as in regular OLS regression

```
library(robustbase)
modelrob = lmrob(Outcome~Predictor, data=mydata)
summary(modelrob)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	74.26659	3.01780	24.61	< 2e-16	***
Predictor	0.07828	0.01747	4.48	3.21e-05	***

# Robust

- Which to use?
- Look at the coefficients, prediction accuracy and model fit. Does the robust improve the model significantly?
- Coefficients: are they dramatically different? are the standard errors noticeably reduced with the robust?
- Prediction accuracy: compare residual standard error
- Model fit: compare BIC (lower = better)



# Beyond the linear

- Sometimes a linear model doesn't fit
- If using a non-arbitrary scale (e.g. income), transformation and subsequent linear fit may be viable, but often in psych research we are dealing with arbitrary scales meaning transformations would perhaps be largely uninterpretable (log BDI score?)<sup>1</sup>
- Scatterplots can clearly denote curvilinear relationships, and sometimes simple polynomial models may suffice
  - ▣  $\hat{Y} = b_0 + b_1X_1 + b_2X^2$
- Other methods such as loess regression can provide a smoothed fit to the data, and while prediction accuracy would be greatly improved it can sometimes it can be somewhat difficult to interpret in the sense of variable importance (which we are often interested in psych research)
  - ▣ Think of it as a bunch of little linear regressions for ranges of the predictor
- Including interaction terms, quantile regression and multilevel modeling are all still linear, but may be helpful in similar situations

# Model Comparison

- In both a general and specific sense you are familiar with this already
  - One way to think about of regression is that is a comparison of your model with  $k$  predictors vs. an intercept only model, i.e. a model in which only the DV mean is predicted with random error
  - Sequential regression compares a model with  $k$  predictors to one with  $k +$  additional predictors
  - You've also seen it in the Neyman-Pearson approach in which one chose between a Null and Alternative model
  - Theoretical debates are model comparisons which may or may not have a scientific basis
  - Anytime you argued with someone about which of you is right about something

$$Y = b_0 + e$$

# Model Comparison

- The problem with our usual approach in psychological research is that our comparisons are to ideas no one believes
  - ▣ Did you really think a model in which the predictors had no relationship with the DV is a viable one?
- Best would be to come up with *viable* models based on prior research/theory, common sense etc. and put them to the test in their ability to predict the construct of interest
- However, one could also start with many predictors, and compare possible models that are subsets of the 'all predictors in' model
  - ▣ Compare different combinations/numbers of predictors
  - ▣ Validate

# Bayesian

- So what's different here?
- Use an inferential procedure to determine the probability of a hypothesis, rather than a probability of some data assuming a hypothesis is true (classical inference)
- Uses prior information
- Takes a model comparison approach

# The Bayesian Approach

- The key to the Bayesian approach is model comparison (e.g. null and alternative hypotheses) and your estimate of prior evidence

$$P(H_i | D) = \frac{P(D | H_i)P(H_i)}{P(D | H_i)P(H_i) + \dots + P(D | H_k)P(H_k)}$$



# Bayesian

- Key concepts
- Prior probability
  - This can come from a variety of sources but prior research is the likely one, but it can also take on uniform values across hypotheses/models (i.e. uninformed)
- Likelihood<sup>1</sup>
  - This is related to ye ol' p-value, the probability of the data assuming a hypothesis our  $p(D | H)$
  - It is actually the probability density associated with the event (this is the y column of Appendix z regarding the standard normal distribution in the Howell text)
- Posterior probability
  - The fun stuff
  - The probability of the hypothesis given the data  $p(H | D)$
  - A function of the product of the prior and the likelihood



# Proportions: M & M's

- Let's say we want to make a guess as to the proportion of brown M&M's in a bag<sup>1</sup>
- They are fairly common but don't make a majority in and of themselves
- Guesses? Somewhere between .1 and .5
- Posit three null hypotheses
  - $H_{01}: \pi = .2$
  - $H_{02}: \pi = .3$
  - $H_{03}: \pi = .4$
- What now?
- Take a sampling distribution approach as we typically would and test each hypothesis

# Proportions

- Using normal approximation
- Bag 1:  $n = 61$  brown = 21 proportion = .344
  - $H_0: \pi = .2$
  - $H_a: \pi \neq .2$
  - p-value = .001    95% CI: (.23,.46)
- Bag 2:  $n = 59$  brown = 15 proportion = .254
  - $H_0: \pi = .3$
  - $H_a: \pi \neq .3$
  - p-value = .481    95% CI: (.14,.37)
- Bag 3:  $n = 60$  brown = 21 proportion = .350
  - $H_0: \pi = .4$
  - $H_a: \pi \neq .4$
  - p-value = .435    95% CI: (.23,.47)
- At this point we might reject the hypothesis of  $\pi = .2$  but are still not sure about the other two



# Proportions: The Bayesian Way

## Bag 1: n = 61 x = 21

Hypothesis	Prior probability	P(D H <sub>i</sub> )	Prior X P(D H <sub>i</sub> )	Posterior Probability
H <sub>1</sub> : π = .2	.333	.0033394	.001130	.02
H <sub>2</sub> : π = .3	.333	.081093	.027044	.52
H <sub>3</sub> : π = .4	.333	.071586	.023838	.46
Total	1.0		.051972	1.0

$$P(D | H_1) = \binom{n}{x} (\pi_0)^x (1 - \pi_0)^{n-x} = \left( \frac{61!}{21!(61-21)!} \right) (.2)^{21} (.8)^{40} = .0033394$$

$$P(D | H_2) = \binom{n}{x} (\pi_0)^x (1 - \pi_0)^{n-x} = \left( \frac{61!}{21!} \right) (.3)^{21} (.7)^{40} = .081093$$

$$P(D | H_3) = \binom{n}{x} (\pi_0)^x (1 - \pi_0)^{n-x} = \left( \frac{61!}{21!} \right) (.4)^x (.6)^{n-x} = .071586$$

$$P(H_i | D) = \frac{P(D | H_i)P(H_i)}{P(D | H_i)P(H_i) + \dots + P(D | H_k)P(H_k)}$$

For H<sub>1</sub> : .001130/.051972 = .02

The rest based on updated priors (previous posterior)

### Bag 2: $n = 59$ $x = 15$

Hypothesis	Prior probability	$P(D H_i)$	Prior X $P(D H_i)$	Posterior Probability
$H_1: \pi = .2$	.021742	.071176	.001548	.03
$H_2: \pi = .3$	.520357	.087510	.045536	.90
$H_3: \pi = .4$	.458670	.007421	.003404	.07
Total	1.0		.050488	1.0

### Bag 3: $n = 60$ $x = 21$

Hypothesis	Prior probability	$P(D H_i)$	Prior X $P(D H_i)$	Posterior Probability
$H_1: \pi = .2$	.030661	.002782	.000085	.00
$H_2: \pi = .3$	.901917	.075965	.068514	.93
$H_3: \pi = .4$	.067422	.078236	.005275	.07
Total	1.0		.073874	1.0

Based on 3 samples of  $N \approx 60$ , which hypothesis do we go with?



# The Bayesian Approach: Means

- Say you had results from high school records that suggested an average IQ of about 108 for your friends
- You test them now ( $N = 9$ ) and obtain a value of 110, which hypothesis is more likely that it is random deviation from an IQ of 100 or 108?
- $z = (110 - 100)/5 = 2.00$ ;  $p(D | H_0) = .05^*$
- $z = (110 - 108)/5 = 0.40$ ;  $p(D | H_1) = .70^*$

# The Bayesian Approach

- So if we assume either the Null (IQ = 100) or Alternative (IQ = 108) are as likely i.e. our priors are .50 for each...

$$p(H_o | D) = \frac{p(D | H_o) * p(H_o)}{p(D | H_o) * p(H_o) + p(D | H_1) * p(H_1)}$$

$$p(H_o | D) = \frac{.05 * .50}{.05 * .50 + .70 * .50} = .067$$

$$P(H_1 | D) = \frac{.7 * .5}{.05 * .50 + .70 * .50} = .933$$

# The Bayesian Approach

- Now think, well they *did* score 108 before, and probably will be closer to that than 100, maybe I'll weight the alternative hypothesis as more likely (.75)

$$p(H_o | D) = \frac{.05 * .25}{.05 * .25 + .70 * .75} = .023$$

$$P(H_1 | D) = \frac{.7 * .75}{.05 * .25 + .70 * .75} = .977$$

- The result is that the end probability of the null hypothesis given the data is even less likely

# Bayesian Summary

- Gist is that by taking a Bayesian approach we can get answers regarding hypotheses, credible intervals that appeal to our intuition,
- It involves knowing enough about your research situation to posit multiple viable hypotheses or choosing from among a set of available ones given some predictors
- While one can go in ‘uninformed’, one of its strengths is the ability to use prior research to make worthwhile guesses regarding priors
  - ▣ Much if not most of Bayesian analysis is ‘objective’ however
  - ▣ Though it may seem subjective, it is no more so than other research decisions, e.g. claiming  $p < .05$  is ‘significant.’
    - ▣ And possibly less since it’s weighted by evidence rather than a heuristic that came out of agricultural research of the 1920s



# Modern Approach Summary

- Modern approaches:
- Take advantage of modern day computing power
  - ▣ Use sampling and simulation<sup>1</sup> techniques easily accomplished with that computing power
- Can test multiple theories
- Are more accurate with complex data
- Can take context into account and have it as a focus
- Allow for better prediction
- Allows more researcher choice

# Modern Approach Summary

- Still overwhelmed? Confused? Maybe even frightened?!?
- Don't go it alone, we're here to help
  
- Behavioral and Statistical Therapists of America can help you
  - Overcome Generalized NHST Confusion (GNC)
  - Acquire a vast and goofy vocabulary involving words like bootstrapping, minimum volume ellipsoid estimator, robustification and the like
  - Obtain a sense of superiority over those who do not engage in modern data analysis
  - Create graphs that will arouse the interest of people at conferences who would otherwise not talk to you
  - Stop making a big deal about the fact that men and women are different because nothing else was statistically significant
  
- After just 6 weeks you may have
  - Cleaner teeth
  - A dissertation that kicks butt
  - Something you can feel proud of
  - A feeling that you actually learned something and helped progress your chosen field of research
  - A higher probability of obtaining \$1 million dollars in the subsequent 6 weeks\*
    - \*Priors may be off in the calculation of said probability
  
- Try it today!!