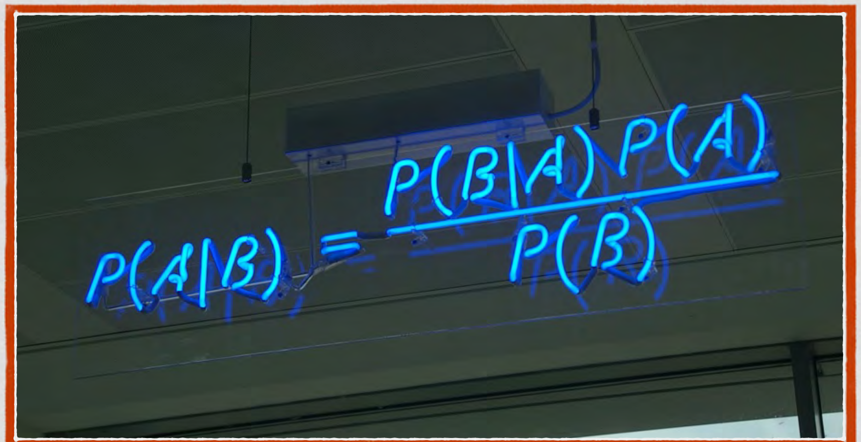


Bayesian Hypothesis Testing

Elizabeth Page-Gould, University of Toronto
Rotman Research Institute Conference 2014

1

Workshop Overview

http://page-gould.com/bayesian

- ☐ Bayesian Statistical Inference
- ☐ Brief Intro to R
- ☐ Bayesian Hypothesis Testing How-To
  - ☐ 2 Ways to Go Bayesian
  - ☐ Reporting and Visualizing Results

2

Probability

- ☐ Prior Probability
  - ☐ P(A)
- ☐ "Conditional Probability" = Posterior Probability
  - ☐ P(A|B)

3

Bayes Theorem

- ☐ Took the formula for conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- ☐ Permuted it in a most useful way:

$$P(A \cap B) = P(A \mid B)\,P(B)$$

$$P(B \cap A) = P(B \mid A)\,P(A)$$

- ☐ Then substituted some terms:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

4

Implications

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- If “A” and “B” are ideas and data …
  - Use probability to quantify logic
  - Quantify how much a single belief changes on the basis of evidence

5

Bayesian Hypothesis Testing

$$P(\, A \mid B\,) = \frac{P(\, B \mid A\,)\,P(\, A\,)}{P(\, B\,)}$$

A = Your Theory, B = The Data

$$P(\text{ Theory} \mid \text{Data }) = \frac{P(\text{ Data} \mid \text{Theory })\,P(\text{ Theory })}{P(\text{ Data })}$$

(Jeffreys, 1935)

6

But … why is everyone FREAKING OUT?

7

Science Is Dominated By One Statistical Approach

- Null Hypothesis Significance Testing (NHST)
- The Null Hypothesis
  - The default hypothesis that people who are skeptical of your hypothesis believe before you do your science
- It’s main value:
  - The null hypothesis is always falsifiable

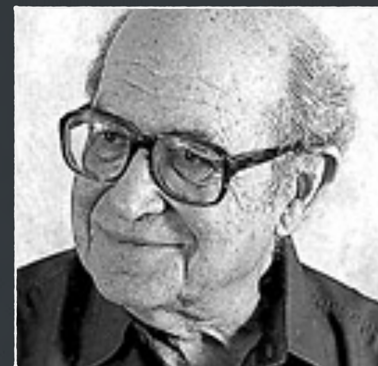
8

Null Hypothesis Significance Testing (NHST)

- Testing the probability of observing your data, given that the null hypothesis is true
- Bayesian expression of a “p-value:”

P(Data I Null Hypothesis)

Issues with NHST



- Conceptual
  - The question you want to ask vs. the question that is answered
- Pragmatic
  - Inferential errors change as a function of sample size and effect size

(Cohen, 1994)

Conceptual Problems With Null Hypothesis Significance Testing

- Fundamental
  - It doesn’t answer the question we need answered!
- Cultural
  - But people typically make the mistake of thinking it does

Argument: Nhst Doesn’t Answer The Question We Really Want To Ask

- The question answered by NHST:
  - What is the probability of observing my data given that the null hypothesis is true?
  - Answer from NHST: The value of your p-value!
- The question we really want to know:
  - What is the probability that my hypothesis is true given the data I have observed?
  - Answer from NHST:

Pragmatic Problems With Null Hypothesis Significance Testing

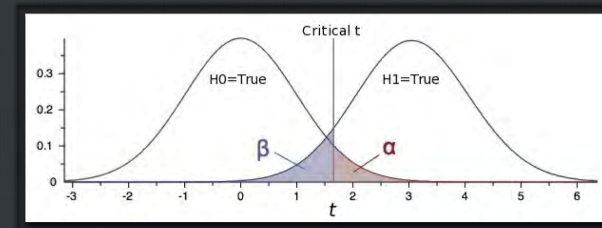
¡¡ERRORS!!

- ☐ Sample size and effect size have a tumultuous, scandalous relationship full of drama

13

Errors In Hypothesis Testing

- ☐ Type I Error — P(Type I Error) = α= 0.05
  - ☐ Rejecting the null hypothesis when the null hypothesis is true
- ☐ Type II Error — P(Type II Error) = β
  - ☐ Failing to reject the null hypothesis when the null hypothesis is false



14

The Smaller Your Sample …

- ☐ If your effect is real:
  - ☐ Only large effects will be significant
    - ☐ More Type II errors
- ☐ But the estimates of effect size are unreliable
  - ☐ “Large” effects may really not be as large and seemingly “small” effects may really not be small
    - ☐ So, more Type I and Type II errors

15

The Larger Your Sample …

- ☐ … everything is significant, even if it is meaningless
  - ➡ So, NHST in very large samples is meaningless; focus on your estimates of effect size
- ☐ Note: Although the significance test no longer matters, the effect size is a very good estimate in large samples

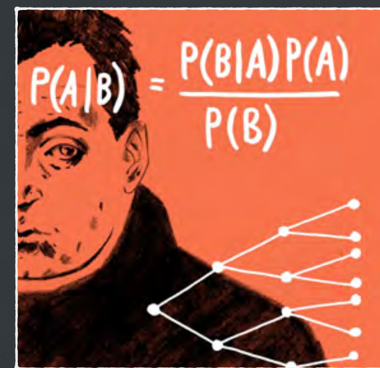
16

The Conundrum!

- Small N = unreliable estimates
  - …. and low NHST sensitivity
- Large N = reliable estimates
  - … but NHST is rendered meaningless

17

Going Bayesian



- Basic ideas
- Actually doing it:
  1. Bayesian Model Comparison
     - Bayes Factors
  2. Bayesian Data Analysis
     - MCMC Sampling

18

Bayesian Hypothesis Testing Terminology

How well your data fit your model

"Posterior" "Likelihood" "Prior"

P(Theory I Data)= P(Data I Theory) P(Theory) / P(Data)

The question you always wanted to test

Expected distribution of posterior

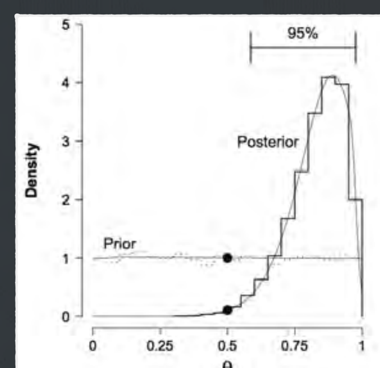
"Marginal Likelihood"

Evidence

19

Posterior Distributions

P(Theory I Data)



- It is a distribution of the values of your parameters, given your data … amazing!!
- What's special about the posterior?
  - At any given point, the posterior distribution represents the culminating influence of all the causal factors that brought you up to that point

20

P(Theory I Data)= $\frac{\text{P(Data I Theory) P(Theory)}}{\text{P(Data)}}$

Prior Distributions
P(Theory)

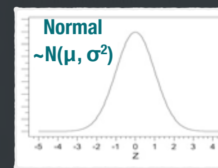
- ☐ An unconditional probability distribution representing a priori belief about a parameter
  - ☐ Commonly denoted by “P(θ)”
- ☐ Sometimes, P(Theory) is also expressed as a conditional statement:
  - ☐ P(Theory) = P(θ| M) = P(Measurements I Theoretical Constructs)

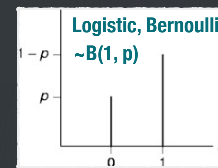
The measures you use to operationalize the constructs

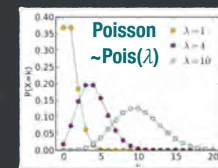
Your theoretical constructs

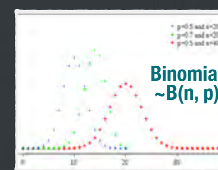
What are Prior Distributions?

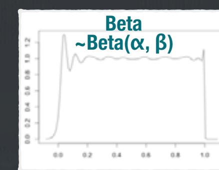
- ☐ The expected probability distribution of your outcome variable!











Likelihood and Marginal Likelihood

How well your data fit your model

“Likelihood”

P(Theory I Data)= $\frac{\text{P(Data I Theory) P(Theory)}}{\text{P(Data)}}$

“Marginal Likelihood”

Evidence

P(Theory I Data)= $\frac{\text{P(Data I Theory) P(Theory)}}{\text{P(Data)}}$

Likelihood
P(Data I Theory)

P( | y = x1 + x2)

- ☐ How well your data fit your hypothesized model
- ☐ Most important component for most forms of Bayesian Hypothesis Testing

P(Theory) = P(Measurements | Theoretical Constructs)

P(Theory | Data)= $\frac{\text{P(Data | Theory) P(Theory)}}{\text{P(Data)}}$

Marginal Likelihood
P(Data)

- Probability of your data, unconstrained by your theoretical model
- Remember, P(Theory) = P(Measurements | Theoretical Constructs)
  - Marginal Likelihood = P(Data | Theoretical Constructs) after "marginalizing" out P(Model Parameters)
- It is typically ignored because it's constant across model comparisons

Going Bayesian: Route #1

Bayesian Model Comparison
Aka "Bayesian Inference"

Brief Introduction to R

Prelude to Bayesian Analysis in R

What Is R?

- R is:
  - A computer program that can do statistics
  - Open-source
    - It's free!!
    - Widely used
    - Most cutting-edge
  - Syntax-based
  - Object-oriented

How To Use R

- Type commands into the “console”
- The output is printed below the command line prompt
- Basic statistical and graphing commands are pre-installed
- R is infinitely extendable through “packages” and user-defined functions
  - Search for packages at cran.r-project.org

29

What Is Object Orientation?

- A programming approach where concepts are represented as “objects”
- An object is a thing that has:
  - Attributes
    - Features of the object that describe it
  - Functions
    - Actions that can be done with the object

30

Object-Oriented Statistics

- Why do people love R so much?
  - Once you start thinking about statistics in an object-oriented way … it’s a whole new world
- Object orientation applied to statistics
  - Both data and statistical analyses are things that you want to know stuff about (i.e., attributes) and want to do stuff to (i.e., functions)
  - Example: What if I did a t-test and put it in an object?
    - Attribute: its degrees of freedom
    - Function: print a nice summary table of results

31

Best Practices With R

- Store all your results as objects so you can further manipulate them
  - The “assignment arrow” is used to create new objects
  - new.object <- function(old.object)
- Save your commands in a separate syntax file with a .R extension
- Read the help file every time you use a function!
  - help(function) or ?function, where “function” is a real function (e.g., mean)
- Read your error messages and search for them online

32

Health Before Wealth!

A Statistical Ode to Loreta Bonomo Gould

33

Example Data

- General Social Survey
- N = 35,484 American Adults
- Variables of interest
  - Overall happiness ≈ Self-rated happiness ("HAPPY")
  - Health ≈ Self-rated health ("HEALTH")
  - Wealth = Household income ("REALINC")
- All predictors have been properly coded from raw dataset

34

Bayesian Model Comparison

- What matters more for your everyday happiness, health or wealth?
  - Hypothesis 1: Healthy people are happier
    - Model 1: Happiness = Health
  - Hypothesis 2: Rich people are happier
    - Model 2: Happiness = Wealth

35

Let's Compare Them

$$P(\text{Model 1} \mid \text{Data}) = \frac{\underbrace{P(\text{Data} \mid \text{Model 1})}_{\text{Happiness = Health}}\ \underbrace{P(\text{Model 1})}_{\text{Happiness = Health}}}{\cancel{P(\text{Data})}}$$

$$P(\text{Model 2} \mid \text{Data}) = \frac{\underbrace{P(\text{Data} \mid \text{Model 2})}_{\text{Happiness = Income}}\ \underbrace{P(\text{Model 2})}_{\text{Happiness = Income}}}{\cancel{P(\text{Data})}}$$

$$\boxed{\frac{P(\text{Model 1} \mid \text{Data})}{P(\text{Model 2} \mid \text{Data})}} = \frac{P(\text{Data} \mid \text{Model 1})\ P(\text{Model 1})}{P(\text{Data} \mid \text{Model 2})\ P(\text{Model 2})}$$

Bayes Factor

36

Bayes Factors

- A ratio of the posterior probabilities of two models (e.g., Model 1, Model 2)
  - Typically denoted with variable, "K" or "BF"
  - Historically hard to compute …
    - … good thing we live now!
- Bayes Factor evaluating the likelihood of the model with the smaller BIC relative to the model with the larger BIC:
  - Bayes Factor = I BIC2 - BIC1 I

(Raftery, 1995)

37

Bayesian Information Criterion (Bic)

- Log estimate of the likelihood that the observed data came from your model, with a penalty for models with lots of predictors
  - P(DatalModel)P(Model)
  - For the same set of data, the model with lower the BIC is always preferred
- BIC is the log likelihood, so BIC is in log units
  - Subtracting log variables is equivalent to dividing non-log variables
  - Bayes Factor = $\frac{\text{P(DatalModel 1)P(Model 1)}}{\text{P(DatalModel 2)P(Model 2)}}$ ≈ IBIC2 - BIC1I
- Huge advantage:
  - Can be used for non-nested model comparison
    - But only for models with the same dependent variables!

(Schwarz, 1978)

38

Bayesian Model Comparison In R

- `model.1 <- glm( happiness~health )`
- `model.2 <- glm( happiness~income )`
- `abs( BIC(model.2)-BIC(model.1) )`

39

Bayesian Inference

Bayes Factor	Interpretation
< 1	No functional difference between models
1 - 3	"Not worth more than a bare mention"
3 - 10	Positive evidence in favour of model with smaller BIC
10 - 30	Strong evidence in favour of model with smaller BIC
30 - 100	Very strong evidence in favour of model with smaller BIC
> 100	Decisive evidence in favour of model with smaller BIC

(Jeffreys, 1961, Appendix B)

40

Is Happiness Better Predicted By Health Or Wealth?

- BIC(Happiness = Health): 66448
- BIC(Happiness = Income): 67807
- Bayes Factor = 67807 - 66448 = 1359

41

Reporting Your Analysis

- State the original analyses you ran when calculating BIC
  - "We tested the hypothesis that health is a more important factor for predicting happiness than wealth with Bayesian Inference (Raftery, 1995). General happiness was regressed on each predictor in two general linear models. People who reported better daily health also reported greater happiness, $b = 0.197$, $SE = 0.004$, $t(35483) = 50.67$, $p < 0.001$. Income was also positively related to happiness, $b = 0.004e\text{-}3$, $SE = 0.001e\text{-}4$, $t(35483) = 33.73$, $p < 0.001$."

42

Reporting Your Analysis

- Report BIC of each model and their Bayes Factor (difference)
  - "Bayesian Information Criterion (BIC) values of each model were compared. The model predicting happiness from health had a smaller BIC, $BIC_{Health} = 66448$, than the model predicting stress from income, $BIC_{Wealth} = 67807$, suggesting that the health model is more than 1359 times more likely than the wealth model. Thus, we found decisive evidence that health is more closely related to a person's happiness than wealth."

43

Going Bayesian: Route #2

Bayesian Data Analysis with MCMC Sampling

44

Bayesian General Linear Modelling

- ☐ Rethinking regression as a Bayesian Model
- ☐ Example: Happiness as a function of health versus wealth

45

Classical Perspective

- ☐ General Linear Model:
  - ☐ $y_i = b_0 + b_1 * x_i + e_i$

Variance in Happiness That Your Model Doesn't Explain

Average Happiness

$$\boxed{happiness_i} = \boxed{b_0} + \boxed{b_1} * \boxed{health_i} + \boxed{e_i}$$

What you are trying to predict, $\hat{y}_i$

Data you collected to hone your prediction, y_i

Degree to which our prediction for happiness changes with health

Information That You Think Will Improve Your Happiness Prediction

46

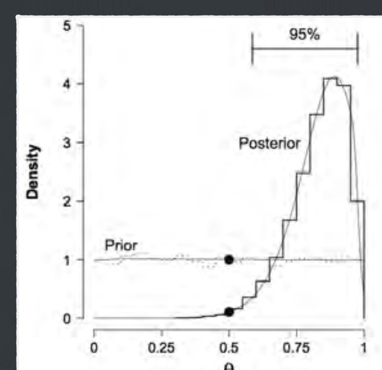
Bayesian Perspective

$$P(\text{ Theory I Data }) = \frac{P(\text{ Data I Theory })\, P(\text{ Theory })}{P(\text{ Data })}$$

- ☐ Does perceived health predict happiness?
  - ☐ $happiness_i \sim N(mean_i, spread)$
  - ☐ $mean_i$ <- intercept + slope * $health_i$
  - ☐ $happiness_i \sim N(\text{(intercept + slope * } health_i), spread)$

47

How Do You Find the Prior and Posterior Distributions?

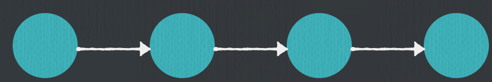


- ☐ Prior
  - ☐ Declare the probability distribution of your dependent variable, with certain starting values
- ☐ Posterior
  - ☐ Sample from the posterior distribution using Markov Chain Monte Carlo (MCMC) Sampling

48

MCMC Chains

- ☐ MCMC chains are samples from the Posterior Distribution of the theory given the data



- ☐ A computer uses a Monte Carlo sampling technique to build stochastic Markov Chains, abbreviated MCMC

49

MCMC Chains

- ☐ MCMC samples are dependent on each other
  - ☐ The first n samples are generated as “burn in” samples and they serve as the priors of the remaining MCMC samples
- ☐ Choose how many chains to run at once

50

Bayesian Linear Modelling In R

- ☐ Steps to Bayesian Hypothesis Testing
  1. Specify your model like you normally would
  2. Use the MCMCregress function to conduct the MCMC sampling
     - ☐ Create 3 chains
  3. Bind these chains together
  4. Evaluate the convergence
  5. Evaluate the output

51

Specify Your Model

- ☐ Health Model:
  - ☐ `happiness~health`
- ☐ Wealth Model:
  - ☐ `happiness~income`

52

Make Some Decisions

- MCMC samples: 5000
- Burn-in: 500
- Chains: 3
- Randomly-generated seeds

Create Each Chain With A Random Seed

```
model.1.chain.1 <-
MCMCregress( happiness ~ health ,
data=example.data, burnin=500,
mcmc=5000, seed=abs(rnorm(1)*1000),
b0=c(0,0), B0=c(1e-6, 1.6e-5) )
```

- …

<u>Important Note:</u> Each chain must have a different seed!

Create A MCMC List From The Chains

```
mcmc.model.1 <- mcmc.list( model.
1.chain.1, model.1.chain.2, model.
1.chain.3)
```


Evaluate Convergence

- You need to make sure that your MCMC chains all converged on the same solution before evaluating that solution
- Commonly-reported diagnostic criteria:
  - Gelman-Rubin Convergence Statistics
  - Autocorrelation
- Some people look at convergence for all non-burn in samples, others look at only the last half

Gelman-Rubin Convergence Statistics

- Measure of between-chain variance relative to within-chain variance
  - Typically denoted with $\hat{R}$
- Ideally, you will report that the average convergence and upper bound of 95% CI are both equal to 1
- If not:
  - Try running more chains (e.g., 100,000)

Obtaining Gelman-Rubin Statistic In R

- `gelman.diag( mcmc.model.1 )`

Autocorrelation

- Correlation of the chain with itself, lagged by k iterations
  - Reflects the "clumpiness" of the MCMC sampling
  - Usually lagged at k = -1, -5, -10, and -50
- Ideally, autocorrelation ≈ 0
- If the chains are autocorrelated:
  - Increase the number of chains
  - "Thin" the chains by only saving chains every kth interval

Obtaining Autocorrelation In R

- `autocorr.diag( mcmc.model.1 )`

Reporting Your Analysis

- Begin by describing the analysis
  - “We tested the hypothesis that perceived health predicted happiness with a Bayesian General Linear Model (Smith, 1973). The likelihood of this hypothesis was estimated through Monte Carlo Marchov Chain (MCMC) sampling using the MCMCpack (Martin, Quin, & Park, 2011) and the coda (Plummer, Best, Cowles, & Vines, 2006) packages for R 3.0.2 (R Core Team, 2013). Three MCMC chains were estimated for 5,000 iterations, discarding the first 500 iterations as burn-in samples. All priors were chosen based on recommendations of weakly-informative priors for the relevant distributions (Gelman, 2008) and initialization values were randomly generated. Happiness was assumed to be normally distributed. The mean of happiness was modelled as a function of the intercept (representing the population mean) with an additive effect for daily health symptoms. Both the intercept and the slope for health were assumed to be normally-distributed around a mean of zero.”

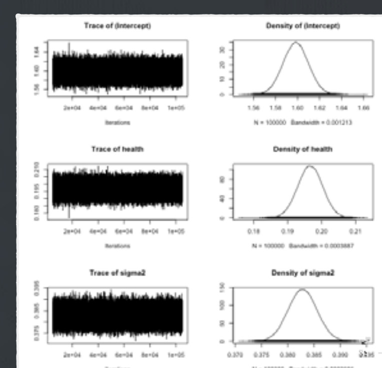
Reporting Your Analysis

- Next, report the convergence information:
  - “The Gelman-Rubin convergence criteria suggested that the chains stabilized on a reliable solution for both the intercept, $\hat{R} = 1$, 95% CI [, 1], and the slope for health, $\hat{R} = 1$, 95% CI [, 1]. The chains also showed low evidence of autocorrelation for either the intercept, $Lag_1 = -0.006$, $Lag_5 = 0.009$, $Lag_{10} = 0.002$, $Lag_{50} = 0.018$, or the slope for symptoms, $Lag_1 = -0.007$, $Lag_5 = 0.010$, $Lag_{10} = -0.0002$, $Lag_{50} = 0.016$. Together, these diagnostic criteria suggest that the linear model converged on a solution that should be able to make reliable predictions.”

Evaluate The Output

- `summary( mcmc.model.1 )`
- `plot( mcmc.model.1 )`

Can You Predict Happiness From Health?



- The “Highest Posterior Density” or HPD interval for the slope does not include 0

	2.5%	25%	50%	75%	97.5%
(Intercept)	1.5757	1.5907	1.5987	1.6065	1.6210
health	0.1894	0.1941	0.1965	0.1990	0.2039
sigma2	0.3775	0.3810	0.3828	0.3848	0.3881

- The processes converged
- The posterior distribution was reliably above 0

Reporting Your Analysis

- Finally, report all the interesting stuff you found
  - "The most common posterior values for the intercept were below the midpoint of the scale, MED = 1.60, 95% HPD [1.16, 2.04]. The slope for perceived health did not include zero and was positive, MED = 0.196, 95% HPD [0.058, 0.338], suggesting that knowing someone's perceived health improves the prediction of their happiness."

65

Best Practices and Conclusions

66

Bayesian Values

- Stochastic processes
  - "Yesterday's posterior is today's prior."
- Competing Models
  - Strong Inference (Platt, 1964, Science)
- Attitudes toward MCMC chains

67

Current Issues

- People argue about priors
  - Priors differ in how informative they are
  - Priors differ in how proper they are
- Creates two camps:
  - "Subjective Bayesians" v. "Objective Bayesians"

68

The "Informativeness" of Priors

- People vary in how strongly they state their prior beliefs
- If you state your belief strongly …
  - E.g., the true correlation is ~N with Mean = +0.3 and SD = 0.06
  - Pitfall: Your beliefs have greater influence over the shape of the posterior distribution
- If you state your belief weakly …
  - E.g., true correlation is equally likely at any real value between -1 and 1
  - Pitfall: You run the risk of overestimating the relative densities of the posterior distribution to the prior distribution

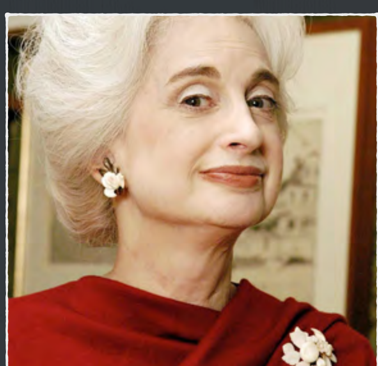
69

Different Classes Of Priors, Based On Informativeness

- Informative Priors ("Subjective Bayesians")
  - Prior distributions that are specific about the values of model parameters (e.g., true correlation ≈ N(μ = -0.5)
- Non-informative Priors ("Objective Bayesians")
  - Usually, uniform distributions that includes all values of a parameter (e.g., -1 ≤ true correlation ≤ +1, with every value having equal probability)
- Weakly-Informative Priors ("WIP"; Most Bayesians)
  - Specifying the distribution (e.g., Normal), with starting values known to bias estimates the least
  - See Gelman, Jakulin, Pittau, & Su (2008) for some default WIP

70

The "Propriety" of Priors



- Improper Priors
  - A probability distribution that integrates to infinity
  - e.g., Unbounded, continuous uniform distribution, U(-∞, +∞), seen with uninformative priors
  - Try to avoid that
    - … or don't (c.f., Jeffreys, 1961)
    - Better to go with "weakly informative priors" (Gelman et al., 2008)
- Proper Priors
  - A probability distribution whose integral is finite

71

Bayesian Hypothesis Testing: Pros

- Quantify the amount of support for one hypothesis relative to another
- Parsimony is rewarded
- Evidence can be gathered in favour of a hypothesis
- Sample size does not affect estimates as much as it does the stability of your posterior distribution

72

Bayesian Hypothesis Testing: Cons

- There can be ambiguity around the choice of priors
  - Bad priors are quickly remedied through MCMC
  - Why we always “burn in” the first 1000 chains
- Culturally, relatively uncommon
  - But present in mainstream discourse

73

Bottom Line

- Statistics is both Classical and Bayesian
- You are only as intellectually flexible as your statistical toolkit is broad

74

Where To Go From Here

- Run the examples yourself, following the steps in the distributed slides
- Bayesian Inference
  - Immediately begin calculating BIC and using it to compare competing hypotheses
- Bayesian Data Analysis
  - Think of how you would analyze your data as regression, then apply it to the second example
- Check out the readings on the last slide

75

¡¡Thank You!!

- Workshop Materials:
  - http://page-gould.com/bayesian/
- Questions? Comments?
  - elizabeth.page-gould@utsc.utoronto.ca
- R Bayesian Task View:
  - http://cran.r-project.org/web/views/Bayesian.html
- Recommended papers:
  - Nice General Intro and Savage-Dickey Density Ratio: Wagenmakers, Lodewyckx, Kuriyal, & Grasman (2010)
  - Bayes Factors and Model Comparison: Raftery (1995)
- Recommended book:
  - Gelman, Carlin, Stern, Rubin, & Dunson (2013), Bayesian Data Analysis, 3rd Edition.

76