A photograph of a chalkboard with the Bayesian formula written in blue chalk. The formula is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The chalkboard is framed by a light-colored border with dark corner mounts. The background shows a classroom setting with a projector screen and some equipment.
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

BAYESIAN HYPOTHESIS TESTING

ELIZABETH PAGE-GOULD

WORKSHOP OVERVIEW

<http://page-gould.com/bayesian>

- Bayesian Statistical Inference
- Bayesian Hypothesis Testing How-To
- Reporting Results

PROBABILITY

- Prior Probability
 - $P(A)$
- “*Conditional Probability*” = Posterior Probability
 - $P(A | B)$

BAYES THEOREM

- Took the formula for conditional probability:

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

- Permuted it in a most useful way:

$$P(A \cap B) = P(A | B)P(B)$$

$$P(B \cap A) = P(B | A)P(A)$$

- Then substituted some terms:

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



IMPLICATIONS

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

- Use probability to quantify logic
- Allows you to ...
 - Quantify how much a single belief changes on the basis of evidence
 - Compare the likelihood of competing possibilities

BAYESIAN HYPOTHESIS TESTING

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

A = Your Theory, B = The Data

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

**BUT ... WHY IS EVERYONE
FREAKING OUT?**



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

NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

- Testing the probability of observing your data, given that the null hypothesis is true

- “ p -value”:

$$P(Data \mid Null Hypothesis)$$



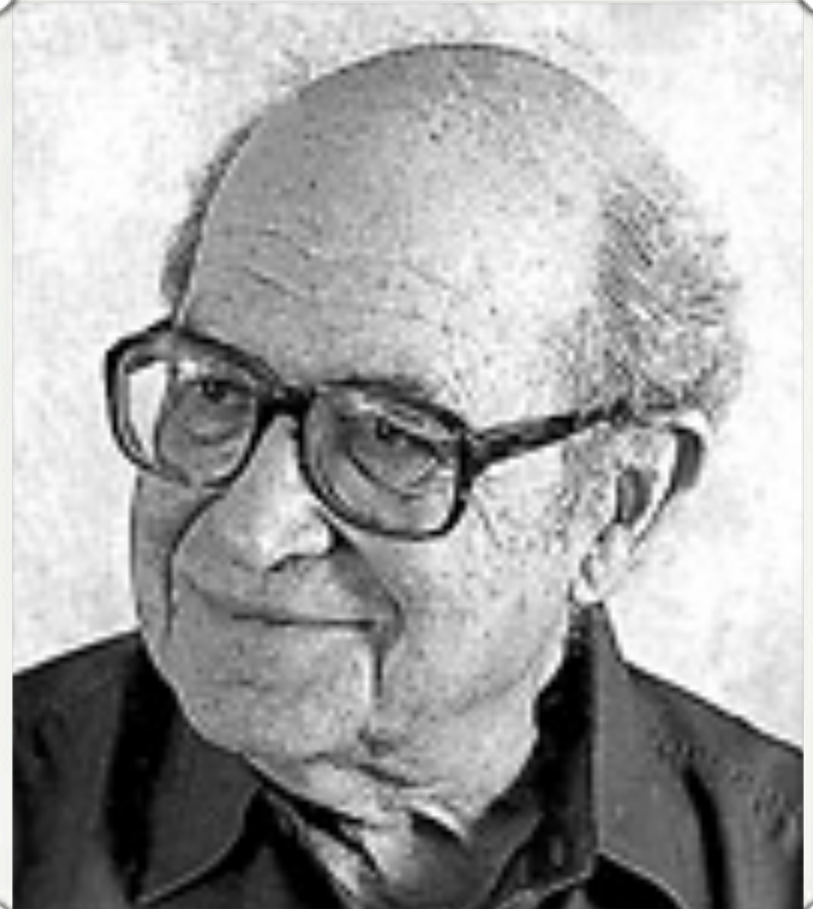
ISSUES WITH NHST

- **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*: <crickets>

PRAGMATIC PROBLEMS WITH NULL HYPOTHESIS SIGNIFICANCE TESTING

!!ERRORS!!

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

ERRORS IN HYPOTHESIS TESTING

- Type I Error

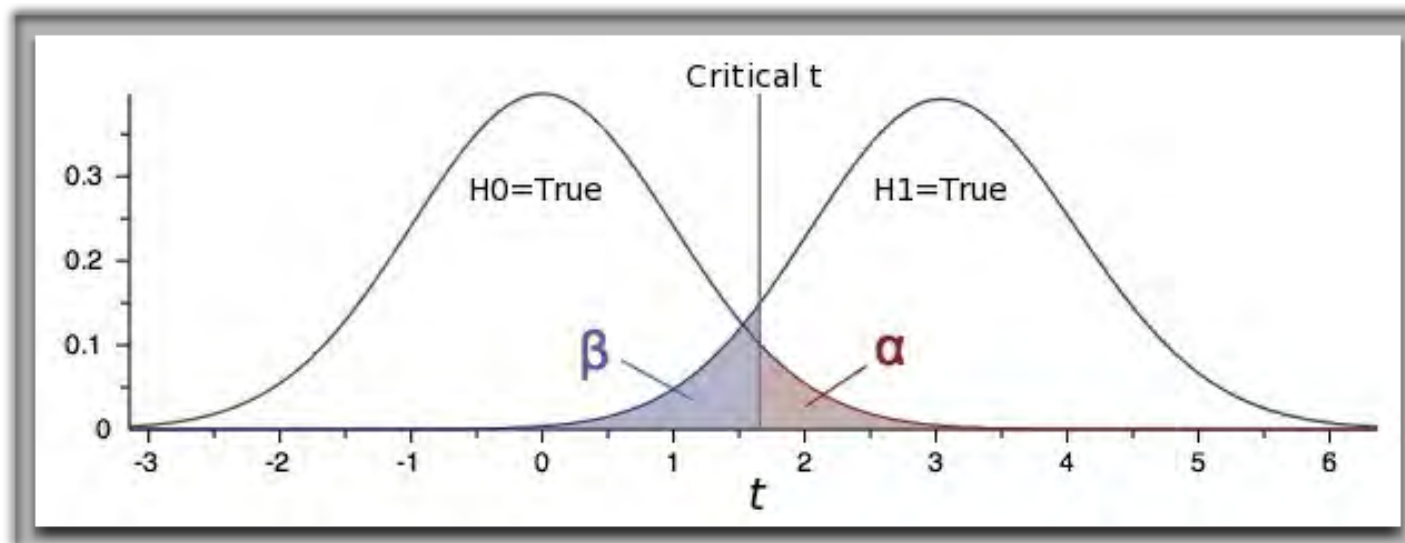
$$\mathbf{P(\text{Type I Error})} = \alpha = 0.05$$

- *Rejecting the null hypothesis when the null hypothesis is true*

- Type II Error

$$\mathbf{P(\text{Type II Error})} = \beta$$

- *Failing to reject the null hypothesis when the null hypothesis is false*



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

THE LARGER YOUR SAMPLE

...

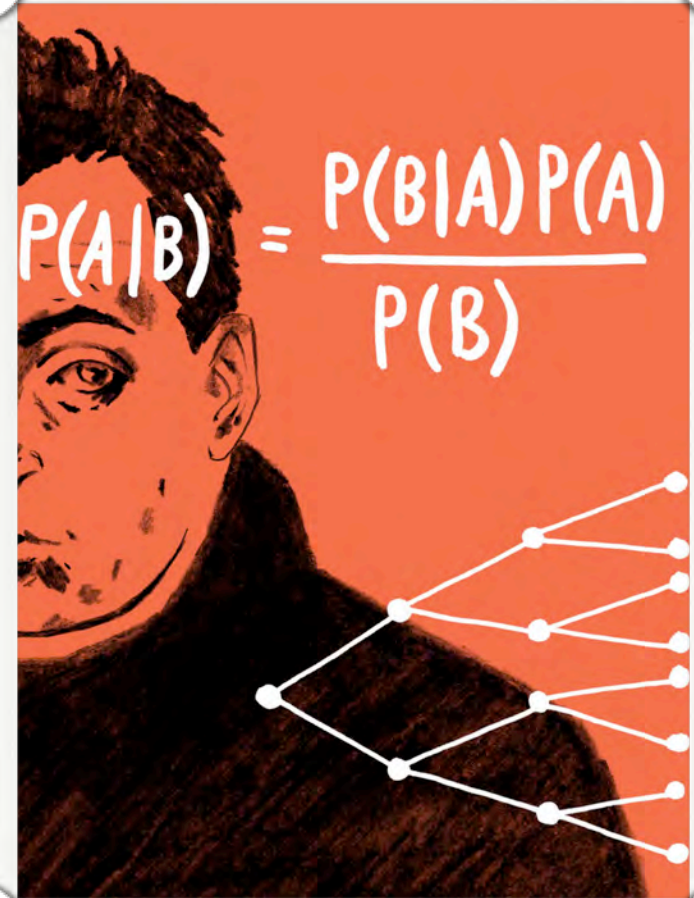
- ... everything is significant, even if it is meaningless
 - ➡ *So, NHST in very large samples is meaningless; focus on your estimates of effect size*
- *Note:* Although the significance test no longer matters, the effect size is a very good estimate in large samples

THE CONUNDRUM!

- Small \mathcal{N} = unreliable estimates
 - and low NHST sensitivity
- Large \mathcal{N} = reliable estimates
 - ... but NHST is rendered meaningless

GOING BAYESIAN

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



BAYESIAN HYPOTHESIS TESTING TERMINOLOGY

$$\begin{array}{c}
 \text{How well your data} \quad \text{Expectation for} \\
 \text{fit your model} \quad \text{posterior} \\
 \text{distribution} \\
 \text{“Posterior”} \quad \text{“Likelihood”} \quad \text{“Prior”} \\
 \boxed{P(Theory \mid Data)} = \frac{\boxed{P(Data \mid Theory)} \boxed{P(Theory)}}{\boxed{P(Data)}} \\
 \text{The question you have} \quad \text{“Marginal} \\
 \text{always wanted to test} \quad \text{Likelihood”} \\
 \text{Evidence}
 \end{array}$$

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

PRIOR DISTRIBUTIONS

$$P(\text{Theory})$$

- *An unconditional probability distribution representing a priori belief about a parameter*

- Commonly denoted by “ $P(\theta)$ ”

- Sometimes, $P(\text{Theory})$ is also expressed as a conditional statement:

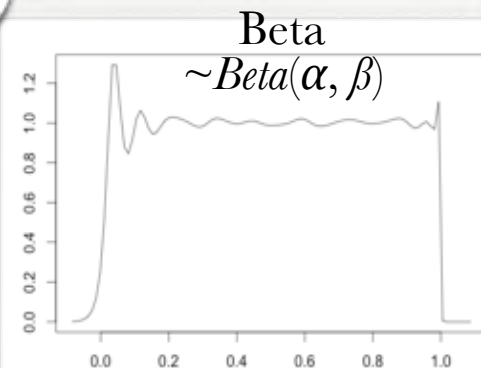
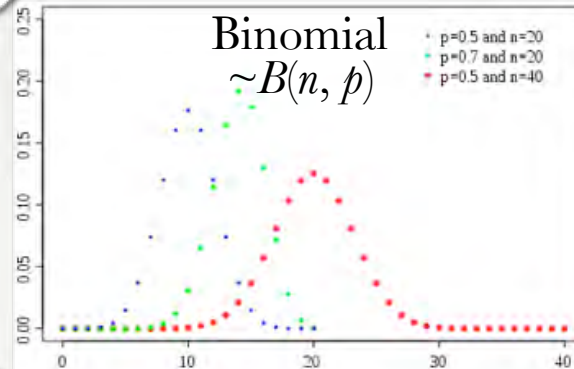
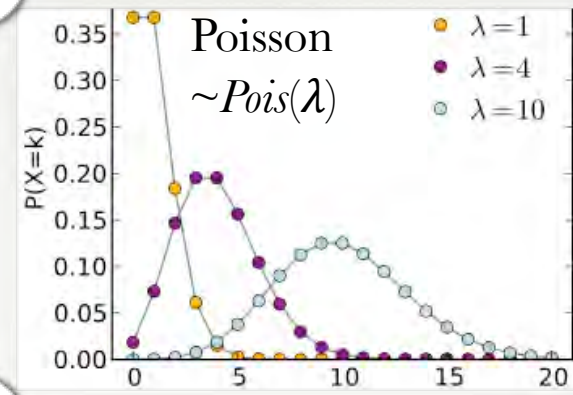
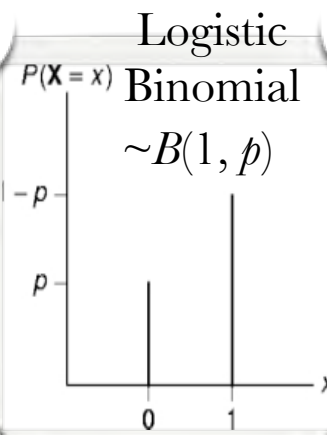
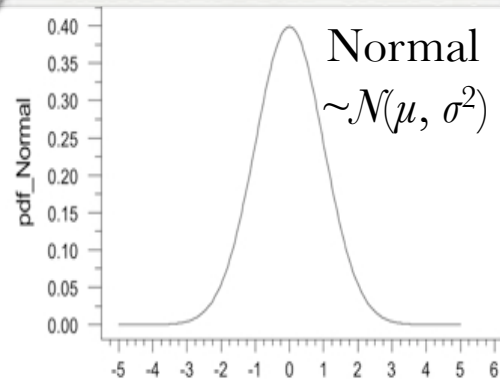
$$P(\text{Theory}) = P(\theta | M) = P(\text{Model Parameters} | \text{Theoretical Constructs})$$

The measures you collect to quantify/operationalize the constructs

WHAT ARE PRIOR DISTRIBUTIONS?

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

- The expected *probability distribution* of your outcome variable!



LIKELIHOOD AND MARGINAL LIKELIHOOD

How well your data
fit your model

“Likelihood”

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

“Marginal
Likelihood”

Evidence

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

LIKELIHOOD

$$P(Data | Theory)$$

$$P\left(\begin{array}{|c|} \hline \text{ID} & \text{sex} & \text{bdate.month} & \text{bdate.year} \\ \hline 1 & 1 & 2 & 9 & 1981 \\ 2 & 2 & 1 & 7 & 1982 \\ 3 & 3 & 2 & 9 & 1983 \\ 4 & 4 & 2 & 2 & 1981 \\ 5 & 5 & 1 & 10 & 1982 \\ 6 & 6 & 2 & 1 & 1982 \\ 7 & 7 & 1 & 4 & 1983 \\ 8 & 8 & 2 & 6 & 1981 \\ \hline \end{array} \mid y = x_1 + x_2 \right)$$

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

$$P(\text{Theory}) = P(\text{Model Parameters} \mid \text{Theoretical constructs})$$

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

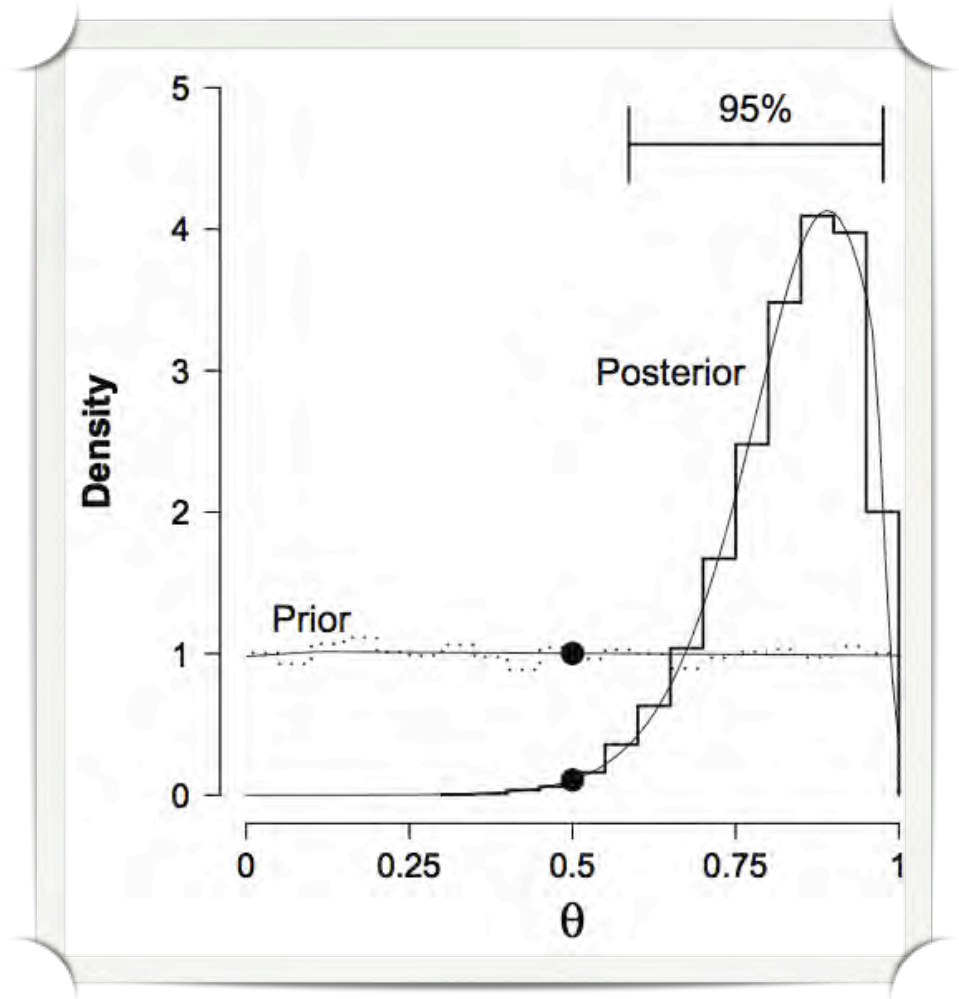
MARGINAL LIKELIHOOD

$$P(\text{Data})$$

- *Probability of your data, unconstrained by your theoretical model*
- Remember, $P(\text{Theory}) = P(\text{Model Parameters} \mid \text{Theoretical Constructs})$
 - *Marginal Likelihood = $P(\text{Data} \mid \text{Theoretical Constructs})$ after “marginalizing” out $P(\text{Model Parameters})$*
- It is typically ignored because it's constant across model comparisons

POSTERIOR DISTRIBUTIONS

- *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



GOING BAYESIAN: ROUTE #1

**BAYESIAN MODEL COMPARISON
AKA “BAYESIAN INFERENCE”**

HEALTH BEFORE WEALTH!

A STATISTICAL ODE TO
LORETA BONOMO GOULD



EXAMPLE DATA

- $N = 246$ online participants
- Variables of interest
 - “psychological.stress” = *Perceived Stress Scale*
 - “symptoms” = count of up to 12 physical symptoms from Hopkins Symptoms Checklist
 - “income” = household income ranging from $<\$20,000/\text{year}$ to $>\$200,000/\text{year}$ in increments of $\$10,000/\text{year}$
- All predictors have been mean-centered



BAYESIAN MODEL COMPARISON

- What matters more for your everyday psychological stress, health or wealth?

H1 Healthy people are less stressed out

Model 1: Psychological Stress = Physical symptoms

H2 Rich people are less stressed out

Model 2: Psychological Stress = -Wealth

LET'S COMPARE THEM

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

^{Stress = Symptoms}
^{Stress = Symptoms}

ID	sex	birthdate month	birthdate year
1	1	9	1981
2	2	1	1982
3	3	2	1983
4	4	2	1981
5	5	1	1982
6	6	2	1982
7	7	1	1983
8	8	2	1981

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

^{Stress = -Income}
^{Stress = -Income}

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

Bayes Factor

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 = $|BIC_2 - BIC_1|$

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*
 - $P(Data | Model)P(Model)$
 - 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 = $\frac{P(Data | Model\ 1)P(Model\ 1)}{P(Data | Model\ 2)P(Model\ 2)} \approx |BIC_2 - BIC_1|$
- Huge advantage:
 - Can be used for non-nested model comparison
 - But only for models with the same dependent variables!

(Schwarz, 1978)

BAYESIAN MODEL COMPARISON IN SPSS

- You must use the GENLIN procedure to get BIC
 - Syntax documentation for GENLIN: <http://tinyurl.com/d9w98wg>
- GENLIN psychological.stress WITH symptoms
/MODEL symptoms DISTRIBUTION=NORMAL
/PRINT FIT.
- GENLIN psychological.stress WITH income
/MODEL income DISTRIBUTION=NORMAL
/PRINT FIT.
- Manually calculate Bayes Factor: $|BIC_2 - BIC_1|$

BAYESIAN MODEL COMPARISON IN R

- `model.1 <- glm(psychological.stress
~ symptoms)`
- `model.2 <- glm(psychological.stress
~ income)`
- `bayes.factor <- abs(BIC(model.2) -
BIC(model.1))`
- `bayes.factor`

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*)

IS STRESS BETTER PREDICTED BY HEALTH OR WEALTH?

- $\text{BIC}(\text{Stress} = \text{Symptoms}) = 444.86$
- $\text{BIC}(\text{Stress} = \text{Income}) = 499.98$
- $\text{Bayes Factor} = 499.98 - 444.86 = 55.13$

REPORTING YOUR ANALYSIS

- *State the original analyses you ran when calculating BIC, indicating some measure of the overall quality of the models prior to comparing them*
 - “We tested the hypothesis that health is a more important factor for predicting psychological stress than wealth with Bayesian Inference (Raftery, 1995). Psychological stress was regressed on each predictor in two general linear models. Experiencing more physical symptoms predicted psychological stress, $b = 0.115$, $SE = 0.014$, $t(245) = 8.34$, $p < 0.001$. Income was negatively related to psychological stress, $b = -0.031$, $SE = 0.012$, $t(245) = -2.57$, $p = 0.011$. Thus, each model represents a viable candidate for model comparison.”

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 stress from daily symptoms had a smaller BIC, $BIC_{Health} = 445$, than the model predicting stress from income, $BIC_{Wealth} = 500$, suggesting that the health model is 55 times more likely than the wealth model. Thus, we found very strong evidence that health is more informative about a person’s psychological stress than wealth.”

GOING BAYESIAN: ROUTE #2

BAYESIAN DATA ANALYSIS

BAYESIAN DATA ANALYSIS

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

- Basic process:
 1. Build a probability density distribution for the posterior distribution
 2. Compare it to the probability density distribution you declared for your prior

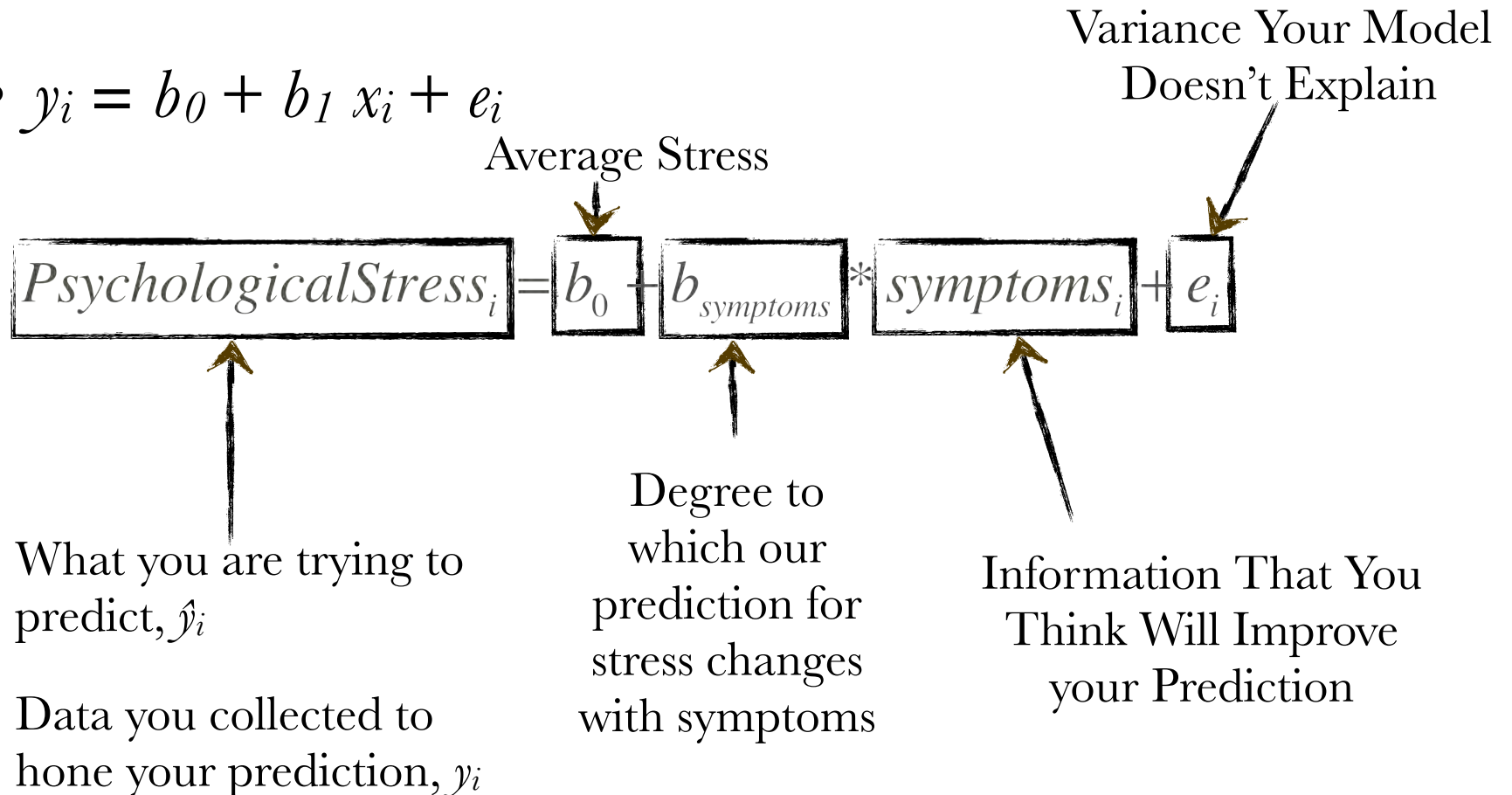
EXAMPLE

- Bayesian General Linear Modelling (rethinking regression)
 - Psychological stress as a function of physical health versus wealth

CLASSICAL PERSPECTIVE

- General Linear Model:

- $y_i = b_0 + b_1 x_i + e_i$



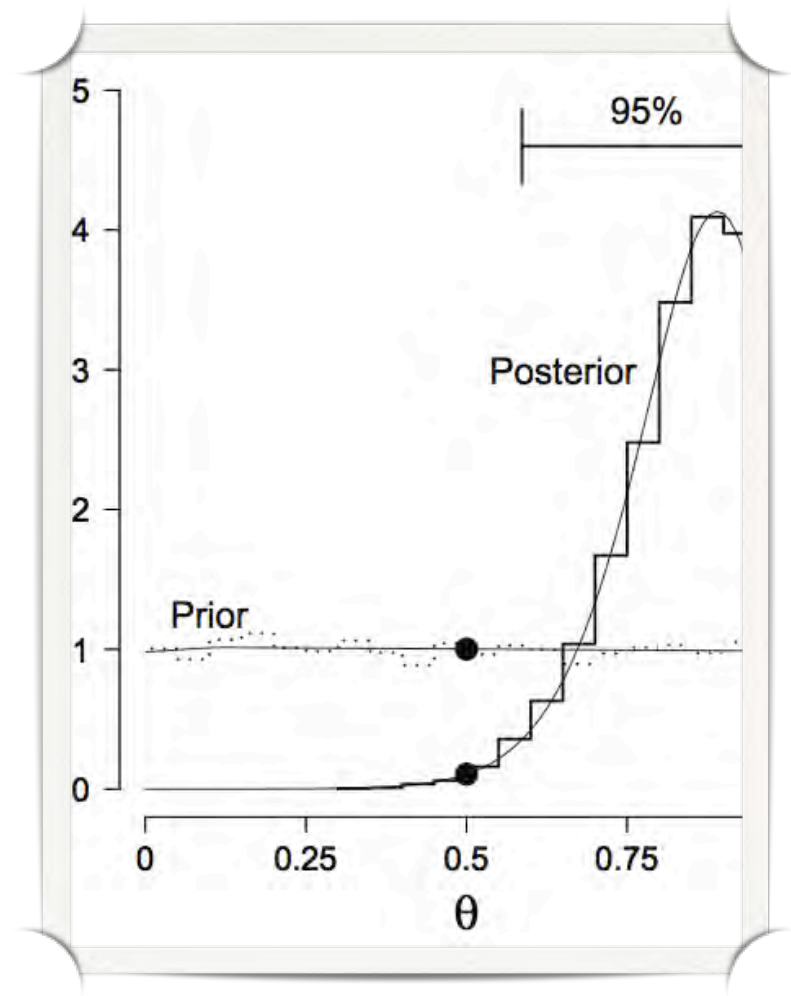
BAYESIAN PERSPECTIVE

Psychological Stress $\sim \mathcal{N}(\text{true mean, spread})$

- *Does the presence of physical symptoms predict feeling psychologically stressed?*
 - $\text{Psychological.stress}[i] \sim \mathcal{N}(\text{mean}[i], \text{spread})$
 - $\text{mean}[i] \leftarrow \text{intercept} + \text{slope} * \text{symptoms}[i]$

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



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*
 - 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

OPENBUGS (AKA WINBUGS)

- *Software for Bayesian Hypothesis Testing*
- Needs you to provide at least 2 things:
 1. Model specification
 2. Data specially formatted as a “list”
- You can optionally provide:
 - “Script” commands to automate the process
 - Seed or “initialization” values

MODEL SPECIFICATION

- There is a normal distribution of stress for an individual, from which each observation of an individual's stress originates ...* It is assumed that the real data came from a Normal distribution with some true population mean and variance

$$\text{stress}_i \sim N(\text{mean}_i, \text{spread})$$

- Each person's stress is predicted by the true human level of stress, modified by health*

$$\text{mean}_i = \text{intercept} + \text{symptoms.slope} * \text{symptoms}_i$$

Weakly-informative priors for model parameters

$$\begin{aligned} \text{intercept} &\sim N(0, .0001) \\ \text{symptoms.slope} &\sim N(0, .0001) \end{aligned}$$

WinBUGS represents the spread of the Normal distribution with “precision” instead of variance, which is calculated as $\text{precision} = 1/\text{variance}$

In the general linear model, model estimates are normally-distributed

SPECIFYING A MODEL IN OPENBUGS

- Use the OpenBUGS syntax to specify your model
- Two symbols to note:
 - \sim indicates how a variable is distributed
 - \leftarrow indicates how a variable is calculated from other model parameters

ANALYSIS STEPS IN OPENBUGS

1. Specify your model and check it
2. Load data
3. Compile your model with 3 parallel MCMC chains
4. Give the chains starting values or generate them randomly
5. Update your model with 1000 burn-in samples
6. Identify the parameters you want to record
7. Update your model with the remaining samples (e.g., 9000 or 99000)
8. Save the chains (“coda”) and evaluate the output

CONVERGENCE DIAGNOSTICS

- *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)

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 k th interval

HOW TO OBTAIN THESE IN SPSS?

- Manually ... (see slides at end of this presentation)
- In a nutshell:
 - Load coda for a given parameter into SPSS
 - Calculate \hat{R}
 - Find autocorrelation within the chains

HOW TO OBTAIN THESE IN R?

- Many, many available packages
 - Of particular note: coda, BRugs, BayesFactor, R2OpenBUGS
- `library(coda)`

```
symptoms.slope.coda <-  
read.openbugs( "symptoms_slope_" )  
  
gelman.diag(symptoms.slope.coda)  
  
autocorr.diag(symptoms.slope.coda)
```

REPORTING YOUR ANALYSIS

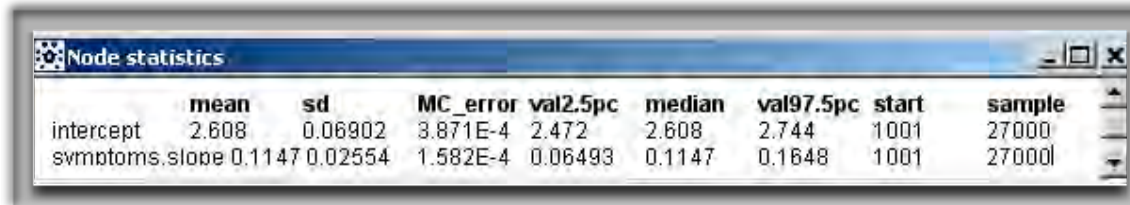
- *Begin by describing the analysis*
 - “We tested the hypothesis that daily health symptoms predict psychological stress with a General Bayesian Linear Model (Smith, 1973). Using OpenBUGS 3.2.2, the likelihood of this hypothesis was estimated through Monte Carlo Markov Chain (MCMC) sampling. Three MCMC chains were estimated for 10,000 iterations, discarding the first 1000 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. Psychological stress was assumed to be normally distributed, and the precision of this distribution was assumed to come from the gamma distribution, $\Gamma(.001, .001)$. The mean of psychological stress was modeled 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 time were assumed to be normally-distributed around a mean of zero, $\mathcal{N}(0, .001)$.”

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, $\hat{R} = 1$, $Upper\ 95\% CI_R = 1$, and the slope for health symptoms, $R = 1$, $Upper\ 95\% CI_R = 1$. The chains also showed low evidence of autocorrelation for either the intercept, $Lag_1 = -0.005$, $Lag_5 = -0.005$, $Lag_{10} = -0.003$, $Lag_{50} = 0.002$, or the slope for symptoms, $Lag_1 = -0.002$, $Lag_5 = 0.010$, $Lag_{10} = 0.012$, $Lag_{50} = -0.003$. Together, these diagnostic criteria suggest that the linear model converged on a solution that should be able to make reliable predictions.”

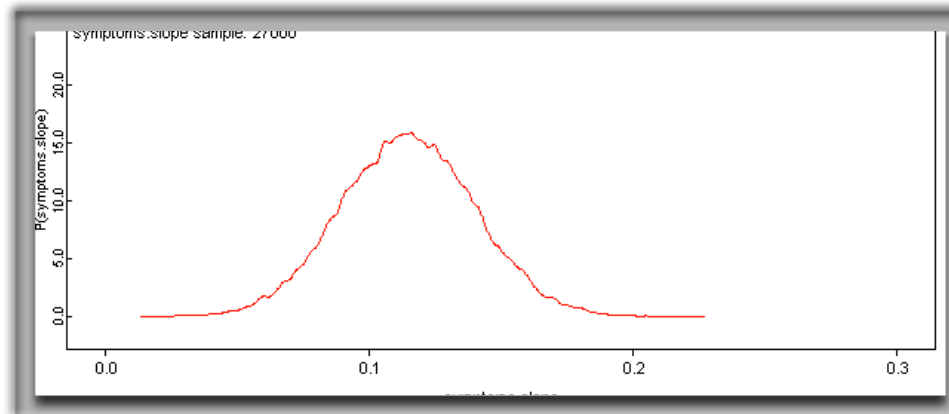
CAN YOU PREDICT PSYCHOLOGICAL STRESS FROM PHYSICAL HEALTH?

- Note the DIC for later model comparison (565.1)
- The “Highest Posterior Density” or HPD interval for the slope does not include 0



	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
intercept	2.608	0.06902	3.871E-4	2.472	2.608	2.744	1001	27000
symptoms.slope	0.1147	0.02554	1.582E-4	0.06493	0.1147	0.1648	1001	27000

- The posterior distribution was reliably above 0



AMOUNT OF SUPPORT FOR EACH HYPOTHESIS

- Bayesian Hypothesis Testing can tell you the probability that your hypothesis is true, given the data
 - *Bayes Factor*
 - This is where the true “hypothesis test” occurs



QUANTIFYING EVIDENCE FOR YOUR HYPOTHESIS

- If the priors for each model are the same, then Bayes Factor can be converted to a probability that one hypothesis is true
- [If Priors Are the Same] Posterior probability of your theory, given the data:

$$P(Hyp. | D) = \text{Bayes Factor} / (\text{Bayes Factor} + 1)$$

WHAT DIFFERENCE DOES HEALTH MAKE IN STRESS?

- $H_1: b_{\text{Symptoms}} = 0$

$$P(H_1 | D) = 0.311$$

- $H_2: b_{\text{Symptoms}} = .2 \text{ } SD_{\text{Stress}} = .14$

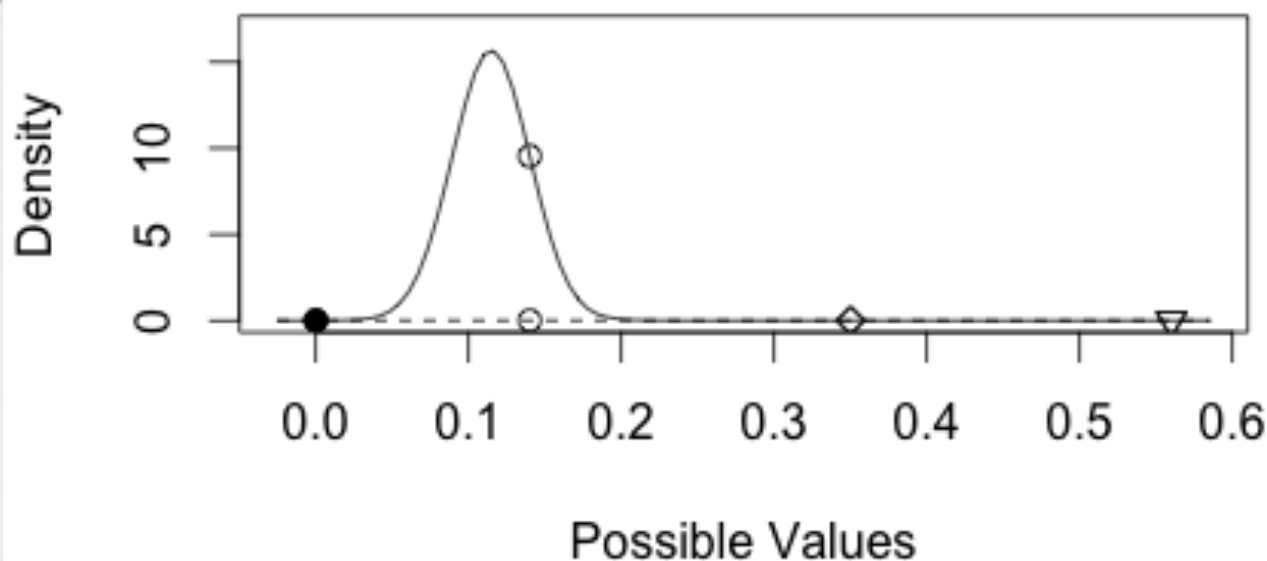
$$P(H_2 | D) = 0.9996$$

- $H_3: b_{\text{Symptoms}} = .5 \text{ } SD_{\text{Stress}} = .35$

$$P(H_3 | D) = 1.03\text{e-}5$$

- $H_4: b_{\text{Symptoms}} = .8 \text{ } SD_{\text{Stress}} = .56$

$$P(H_4 | D) = 1.60\text{e-}14$$



BAYESIAN HYPOTHESIS TEST

- Record and save the coda for the prior distributions of all parameters
- Follow the procedures of Wagenmakers, Lodewyckx, Kuriyal, & Grasman (2010)
 - Or, use my patch that implements their method in R:
 - `source("http://page.gould.com/scripts/r/bayes.test.r")`
 - Syntax is: `bayes.test (prior.coda, posterior.coda, hypothesized.values)`, where `prior.coda` and `posterior.coda` are MCMC lists and `hypothesized values` is a numerical value or vector
 - `bayes.test(symptoms.slope.prior.coda, symptoms.slope.coda, .5) #Single hypothesis`
 - `bayes.test(symptoms.slope.prior.coda, symptoms.slope.coda, c(.5, .62)) #Multiple hypotheses`

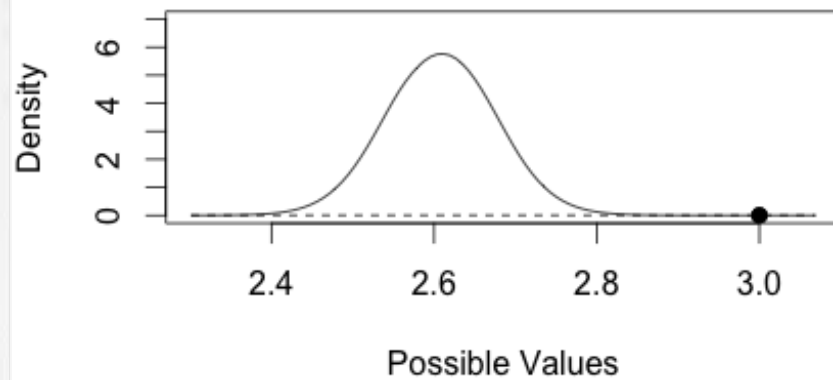
REPORTING YOUR ANALYSIS

^

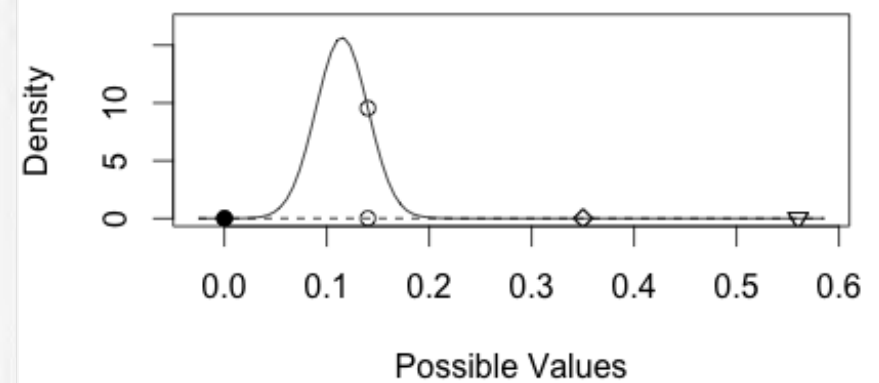
- Finally, report all the interesting stuff you found
 - “The most common posterior values for the intercept were below the midpoint of the scale, $MED = 2.61$, 95% HPD [2.47, 2.74]. We tested the hypothesis that the sample’s average stress was at the midpoint of the scale by calculating the Savage-Dickey Ratio (Dickey & Lientz, 1970) as implemented in Wagenmakers, Lodewyckx, Kuriyal, & Grasman (2010). Given the data, there is only a 0.012 probability that people are moderately stressed. The slope for daily health symptoms did not include zero and was positive, $MED = 0.115$, 95% HPD [0.065, 0.165], suggesting that daily health symptomatology improves the prediction of psychological stress. Using the Savage-Dickey method, we tested the hypotheses that having one symptom greater than the mean predicts no difference in stress, would predict 0.2 SD_{stress} change in stress ($b_{symptoms} = 0.14$), 0.5 SD_{stress} change ($b_{symptoms} = 0.35$), or 0.8 SD_{stress} amount of difference ($b_{symptoms} = 0.56$). As shown in Figure 2, there was a relatively large amount of support for the hypothesis that health has a small effect on stress levels, $P(b_{symptoms} = 0.14 | Data) = 0.999$, only a mild likelihood of no effect, $P(b_{symptoms} = 0 | Data) = 0.311$, and extremely small chances of either a medium effect, $P(b_{symptoms} = 0.35 | Data) = 1.02e-5$, or large effect, $P(b_{symptoms} = 0.56 | Data) = 1.60e-14$. There is a 99.9% chance that the effect of health on stress is small.”

FIGURES

Posterior and Prior Distributions



Posterior and Prior Distributions



IS KNOWING A PERSON'S HEALTH MORE INFORMATIVE ABOUT THEIR STRESS LEVEL THAN KNOWING THEIR INCOME?

- Finally, doing some model comparison
- Same basic process, but using *Deviance Information Criterion* (DIC) to compare models
 - Interpreted almost identically to BIC, including inferential cutoffs

IS KNOWING A PERSON'S HEALTH MORE INFORMATIVE ABOUT THEIR STRESS LEVEL THAN KNOWING THEIR INCOME?

- ```
model {
 for (i in 1:N) {
 psychological.stress[i] ~ dnorm(mean[i], spread)
 mean[i] <- intercept + income.slope * income[i]
 }
 intercept ~ dnorm(0, .0001)
 income.slope ~ dnorm(0, .0001)
 spread ~ dgamma(.01, 100)
}
```

# CAN YOU PREDICT PSYCHOLOGICAL STRESS FROM PHYSICAL HEALTH?

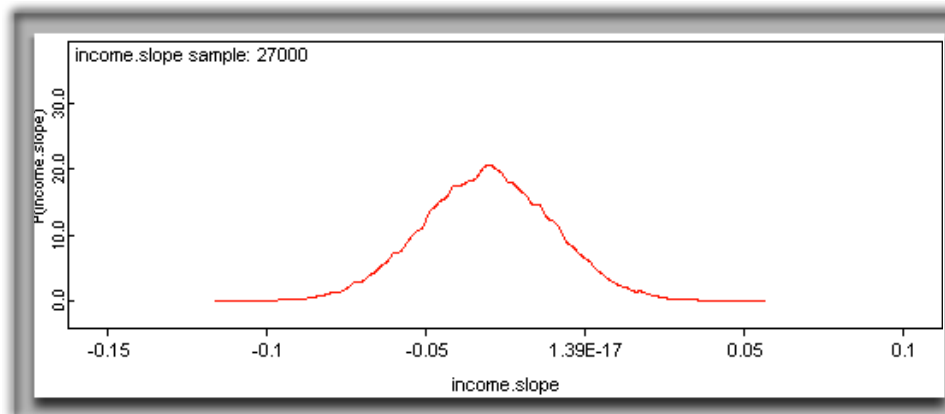
- Note the DIC for later model comparison (594.2)
- The “Highest Posterior Density” or HPD interval for the slope includes 0



A screenshot of a 'Node statistics' window from a statistical software package. It displays a table of summary statistics for two model parameters: 'intercept' and 'income.slope'. The columns include mean, standard deviation (sd), Monte Carlo error (MC\_error), 2.5th percentile (val2.5pc), median, 97.5th percentile (val97.5pc), start of the chain, and the total sample size (sample).

|              | mean     | sd      | MC_error | val2.5pc | median   | val97.5pc | start | sample |
|--------------|----------|---------|----------|----------|----------|-----------|-------|--------|
| intercept    | 2.608    | 0.07122 | 4.437E-4 | 2.488    | 2.608    | 2.748     | 1001  | 27000  |
| income.slope | -0.03063 | 0.02044 | 1.214E-4 | -0.07098 | -0.03057 | 0.009427  | 1001  | 27000  |

- The posterior distribution included 0



# COMPARING TWO MODELS

- *Model with the smallest DIC will make the best short-term predictions*
- Conventions for evaluating DICs same as BICs
- E.g.,  $DIC_{\text{Health}} = 565.1$ ,  $DIC_{\text{Wealth}} = 594.2$ ,  
 $\text{Difference}_{\text{Wealth-Health}} = 29.1$

# REPORTING YOUR BAYESIAN HYPOTHESIS TEST

- Report all the same things as before for the alternative model
  - [...]
- Report the results of the model comparison
  - “The models predicting psychological stress from health versus wealth were compared by taking the difference in the Deviance Information Criterion (DIC). The model predicting stress from health had the smaller DIC,  $DIC_{\text{Health}} = 565$ , than the model predicting stress from wealth,  $DIC_{\text{Wealth}} = 594$ , suggesting that the health model is 29.1 times more likely than the wealth model. Thus, we conclude that health is more important for psychological stress than wealth”

# **BEST PRACTICES AND CONCLUSIONS**

# BAYESIAN VALUES

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

# 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”

# 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  $\sim \mathcal{N}$  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





# 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  $\approx \mathcal{N}(\mu = -0.5)$ )*
- **Non-informative Priors (“Objective Bayesians”)**
  - *Usually, uniform distributions that includes all values of a parameter (e.g.,  $-1 \leq \text{true correlation} \leq +1$ , 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

# THE “PROPRIETY” OF PRIORS

- **Improper Priors**

- *A probability distribution that integrates to infinity*
- e.g., Unbounded, continuous uniform distribution,  $U(-\infty, +\infty)$ , 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*



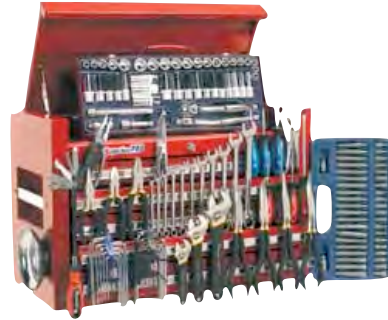
# BAYESIAN HYPOTHESIS TESTING: PROS

- 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

# 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

# BOTTOM LINE



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

# WHERE TO GO FROM HERE

1. Run the examples yourself, following the steps in the distributed slides

2a. Bayesian Inference

- Immediately begin calculating BIC and using it to compare hypotheses

2b. Bayesian Data Analysis

- Think of how you would analyze your data as regression, then apply it to the second example

3. Check out the readings on the last slide

# !!THANK YOU!!

- Workshop Materials:
  - <http://page-gould.com/bayesian/>
- Questions? Comments?
  - [elizabeth.page-gould@utsc.utoronto.ca](mailto:elizabeth.page-gould@utsc.utoronto.ca)
- 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*. (3rd Edition available September 2013)
- Good introduction to WinBUGS: <http://zoe.bme.gatech.edu/~bv20/bmed2803/Bank72/chapter02.pdf>

# **APPENDIX**

## **CALCULATING DIAGNOSTIC CRITERIA WITH SPSS**



# HOW TO CALCULATE COMMON DIAGNOSTIC CRITERIA IN SPSS

- Load all coda for a given parameter into SPSS
- Estimate:
  - Gelman-Rubin  $\hat{R}$
  - Autocorrelation

# GELMAN-RUBIN DIAGNOSTICS

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}$$

$$\hat{V} = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B$$

$$B = \frac{n}{M-1} \sum_{j=1}^M (\bar{\theta}_j - \bar{\theta})^2$$

$$W = \frac{1}{M} \sum_{j=1}^M s_j^2$$

- , where  $M$  = number of chains,  $n$  = length of chains,  $\theta_k$  = values predicted at all  $k$  iterations of the MCMC sampling, and  $s_j^2$  is the variance of each chain

# GELMAN-RUBIN DIAGNOSTICS

- Estimate the variance of the stationary distribution by first calculating:
  - Between-chain variance (B) is the sum of squared deviations of each chain mean from the mean of all chains, multiplied by  $n/M - 1$
  - Within-chain variance (W) is the mean variance of all the chains
- Calculate these things and then just plug in the rest

# CALCULATE AUTOCORRELATIONS IN SPSS

- Let's say you saved coda in variables called "intercept.chain.1," "intercept.chain.2," and "intercept.chain.3"
- `ACF VARIABLES = intercept.chain.1  
intercept.chain.2 intercept.chain.  
3.`
- Report the median autocorrelation across chains