An Introduction to Bayesian Statistics for Psychological Research

Western Psychological Association 2024 Convention

April 28, 2024

Hyeri Hong, Ph.D.

Assistant Professor of Research and Statistics

California State University, Fresno

Email: hyerihong@mail.fresnostate.edu

Website: https://sites.google.com/mail.fresnostate.edu/prof-hyeri-hong?usp=sharing

ORCID:0000-0002-7576-2574

Researchgate: https://www.researchgate.net/profile/Hyeri-Hong

Alfonso J. Martinez

University of Iowa

Email: alfonso-martinez@uiowa.edu

Website: www.ajmquant.com

X: AlfonsoMPsych

Overview

- Core Faculty and Quantitative Fellow in Educational Leadership Doctoral Program and Research Center
- Ph.D. in Ed Measurement and Statistics from the University of Iowa
- Research Interests: Structural Equation Modeling, Bayesian Estimation, and Generalizability Theory + Equity
- P.I. for Steve's Scholars Grant Project (FUSD)
- Courses taught: Ed Statistics, Ed Measurement and Evaluation, Advanced Applied Quan Methods, Research in Education, etc.

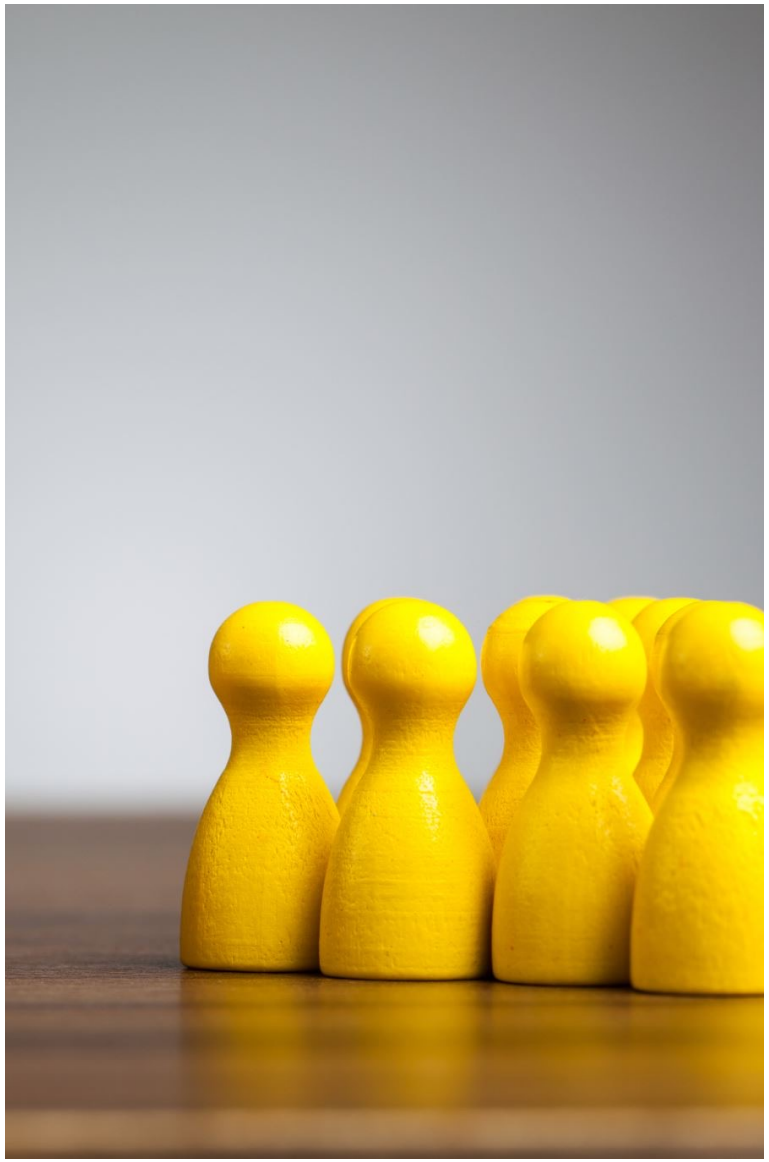
- PhD Candidate in Psychological and Quantitative Foundations at the University of Iowa
- Fresno State Alum (BA in Psychology)
- Research Interests: latent variable modeling, Bayesian statistics, estimation theory, machine learning and applications to psychological research, computational statistics

Part 1: Foundations

Presented by Hyeri Hong, Ph.D.

Outline of This Workshop

- Quiz – Thinking like Bayesian?
- Background
- Bayes vs. Frequentist
- Probability
- Credibility Interval
- Priors
- Markov Chain Monte Carlo
- Convergence
- Model Fit
- Pop vs Soda vs Coke example
- Analysis demonstration



Thinking like a Bayesian?

1. When flipping a fair coin, we say that "the probability of flipping Heads is 0.5." How do you interpret this probability?
    1. If I flip this coin over and over, roughly 50% will be Heads.
    2. Heads and Tails are equally plausible.
    3. Both a and b make sense.

Johnson, A. A., Ott, M. Q., & Dogucu, M. (2022). *Bayes rules!: An introduction to applied Bayesian modeling*. Chapman and Hall/CRC.

Thinking like a Bayesian?

2. Suppose that during a recent dentist's visit, you have a decayed wisdom tooth in the upper gum. If you only get to ask the dentist one question, which would it be?

1. What's the chance that I actually have the decayed wisdom tooth in my upper gum?
2. If in fact I don't have the decayed wisdom tooth, what's the chance that I would've gotten this test result?

Thinking like a Bayesian?

Bayesian: In light of **my test result** (my cavity), what is the chance that I actually have the cavity?

Frequentist: **Testing (Data collection) is repeatable**. You can get tested for the cavity over and over and over.

If I don't actually have the cavity (null hypothesis), what is the chance that I could be tested as having a cavity in my upper gum?

Bayesian Picture

Incoming information

(5-star Korean restaurant)

Data (Not tasty and very spicy but Expensive Bibimbab)

Updated information (3-star restaurant)

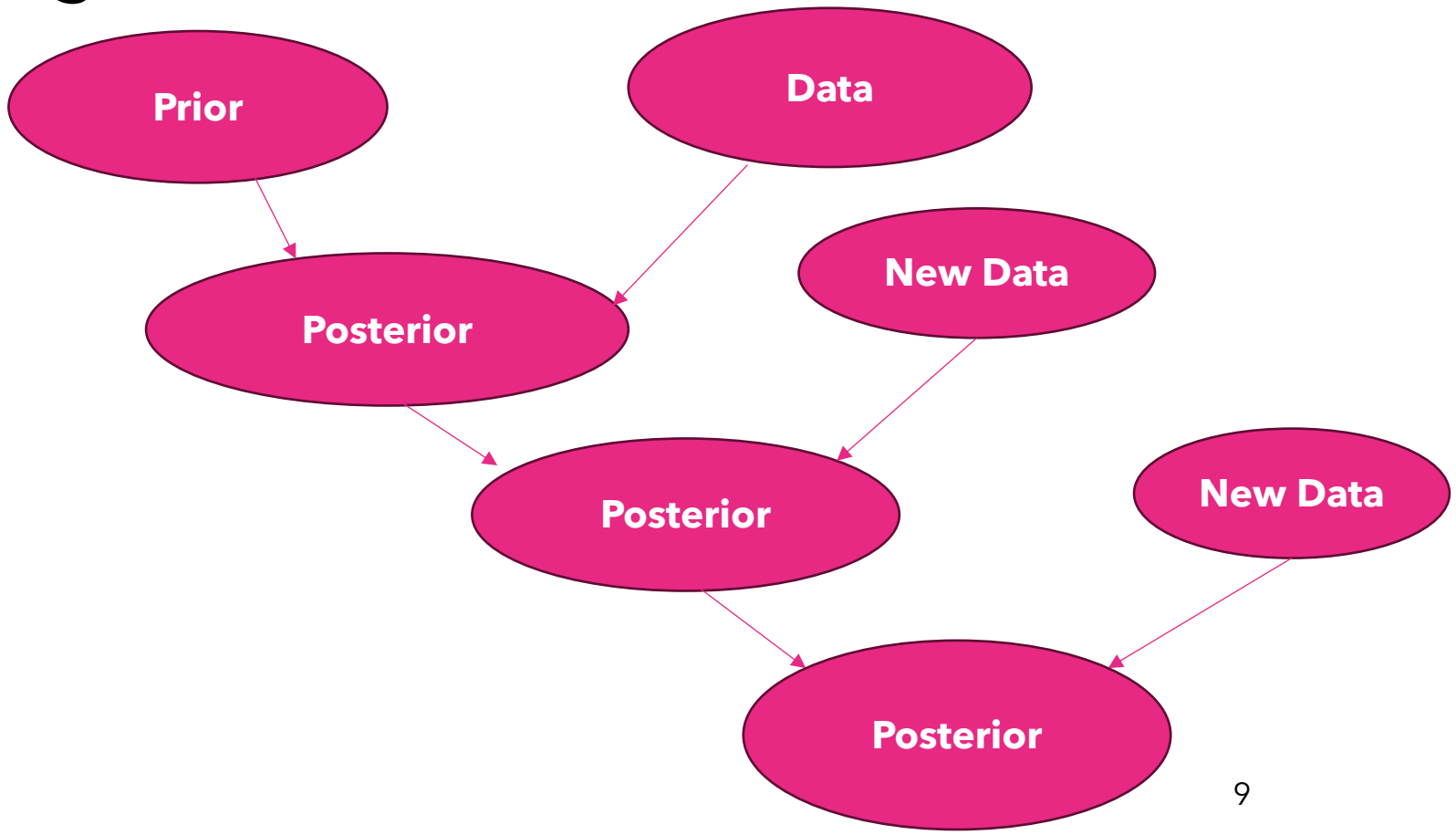
New data (Tasty Bulgogi)

Updated information (4-star restaurant)

Overview

Foundations (focus on conceptual ideas)

Background



Why Bayesian?
Benefits of Bayesian Methods

1

Expand the range of testable hypotheses; **no null hypothesis significance testing**

2

Combine prior findings with new data, producing results that are **automatic meta-analyses**

3

Use prior findings, data from **small-sample studies** is less problematic

4

Allow estimating models when traditional estimation fails because of **model complexity**

Bayesian vs. Non-Bayesian (Frequentists) Approaches

Frequentist: Maximum Likelihood, P values, Confidence Intervals, The Null Hypothesis Significance Testing

Frequentist: **inhibits small-sample research and causes convergence issues**

(negative variance)

Bayesian: Prior, Likelihood, Posterior, Credibility intervals, Small-Sample Research, Convergence

Bayesian vs. Non-Bayesian (Frequentists) Approaches

	Frequentist	Bayesian
Probability	**The relative frequency of an event in a hypothetical infinite series of events** (how often something happens in an infinite series of observations)	**Degrees of belief or degrees of knowledge**
Example 1: There is a .50 probability that a fair coin will land heads.	If the coin were flipped a hypothetical **infinite** number of times, heads would be seen half of the time.	Someone's belief is evenly **divided between the coin landing heads or tails**

Bayesian vs. Non-Bayesian (Frequentists) Approaches

Probability	Frequentist	Bayesian
Example 2: Regression: the effect of an environment (x) on personality (y) $\boldsymbol{Y = a + bx + e}$ **b:** the Effect of Environment	Treat **the data** in Y as **random** Treat **the population parameters** as **fixed** so that the null hypothesis must be treated as a single value (b = 0) to compute a p value for the observed data	Treat **the parameters** as **random** Treat **observed data** in Y as **fixed** or **constant** (not being random)
	Using the p values, assuming the null hypothesis that b = 0 is true, how is the observed effect of an environment unlikely to happen or statistically significant? Estimate **a sampling distribution** for the estimated effect	**Uncertainty** in the effect of the environment is quantified by estimating its probability across a range of values, allowing direct probabilistic statements about the environment based on **observed data**

Bayesian vs. Non-Bayesian (Frequentists) Approaches

Probability	Frequentist	Bayesian
The probability that the environment has an impact on personality	**Confidence Interval**: We are 95 % confident that **the true** average effect of environment on personality ranges between 0.1 and 0.3	**Credibility Interval**: The probability the effect is between 0.1 and 0.3 is 95% given **the observed data**

Bayesian

PROBABILITY

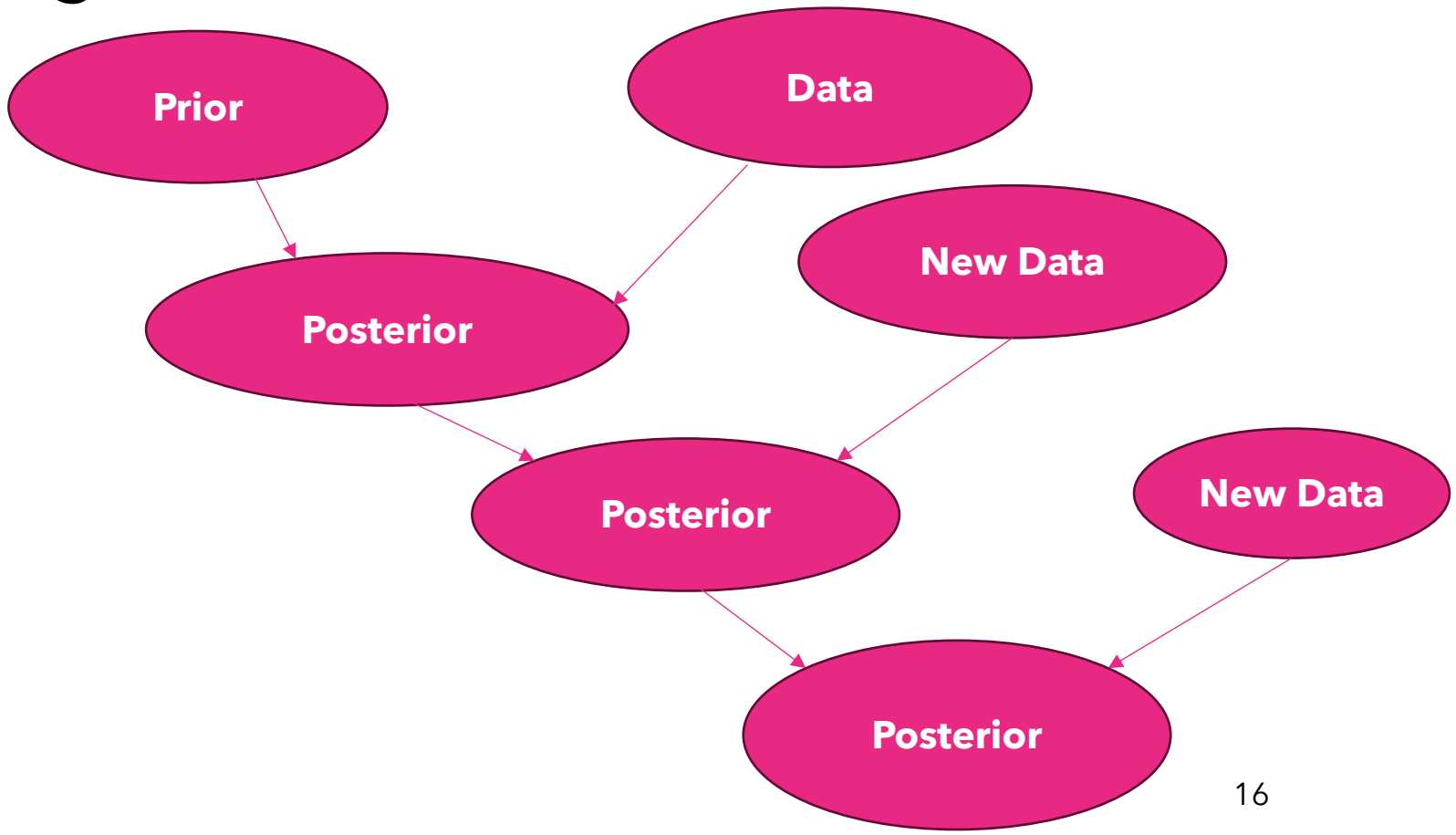
ESTIMATION

INFERENCE

Bayesian Overview

Foundations (focus on conceptual ideas)

Background



Ingredients in the Bayesian Soup : The prior, likelihood, and posterior

Bayes' theorem tells us **how to combine prior distribution with likelihood to construct posterior**

$$P(\theta \mid y) \propto P(y \mid \theta) \times P(\theta)$$

Posterior distribution — Likelihood Function — Prior distribution

The posterior distribution of the parameters given the data is proportional to the likelihood of the data given the parameters times the prior distribution of the parameters before observing the data

Ingredients in the Bayesian Soup : The prior, likelihood, and posterior

$$P(\theta \mid y) \propto P(y \mid \theta) \times P(\theta)$$

Posterior distribution — Likelihood Function — Prior distribution

Bayesian Estimation and Inference : θ **(**random variables)

Bayes' rule estimates $P(\theta \mid y)$= **Posterior Probability** of parameters in θ given observed data in y or **Bayesian Posterior**

= parameters' probability **after** observing the data

= **Results** of a Bayesian analysis

Ingredients in the Bayesian Soup : The prior, likelihood, and posterior

$$P(\theta \mid y) \propto P(y \mid \theta) \times P(\theta)$$

Posterior distribution — Likelihood Function — Prior distribution

are proportional to (i.e., ∝)

multiplied (i.e., weighted) by

Likelihood: the probability of the data as informed by the parameters, $P(y \mid \theta)$, Information contained in the data

Prior: Prior information **before** collecting the data, $P(\theta)$

Example

Regression: the effect of an Environment (x) on personality (y)

$$Y = a + bx + e$$

b: the effect of an environment

$$P(\theta \mid y) \propto P(y \mid \theta) \times P(\theta)$$

Bayes estimation θ: Regression slope b

$P(\theta \mid y)$: Results – The probability that the environment has an impact on personality

$\propto$: proportional to

$P(y \mid \theta)$: the likelihood - the observed data

$\times$: multiplied by

$P(\theta)$: the prior

Regression: the effect of an Environment (x) on personality (y)

$$P(\theta \mid y) \propto P(y \mid \theta) \times P(\theta)$$

- $P(\theta \mid y)$: a probability density function where density describes the probability associated with the range of all possible b values
- Which values are **most probable** for b, given the data?
- The median or the peak as the most probable estimate

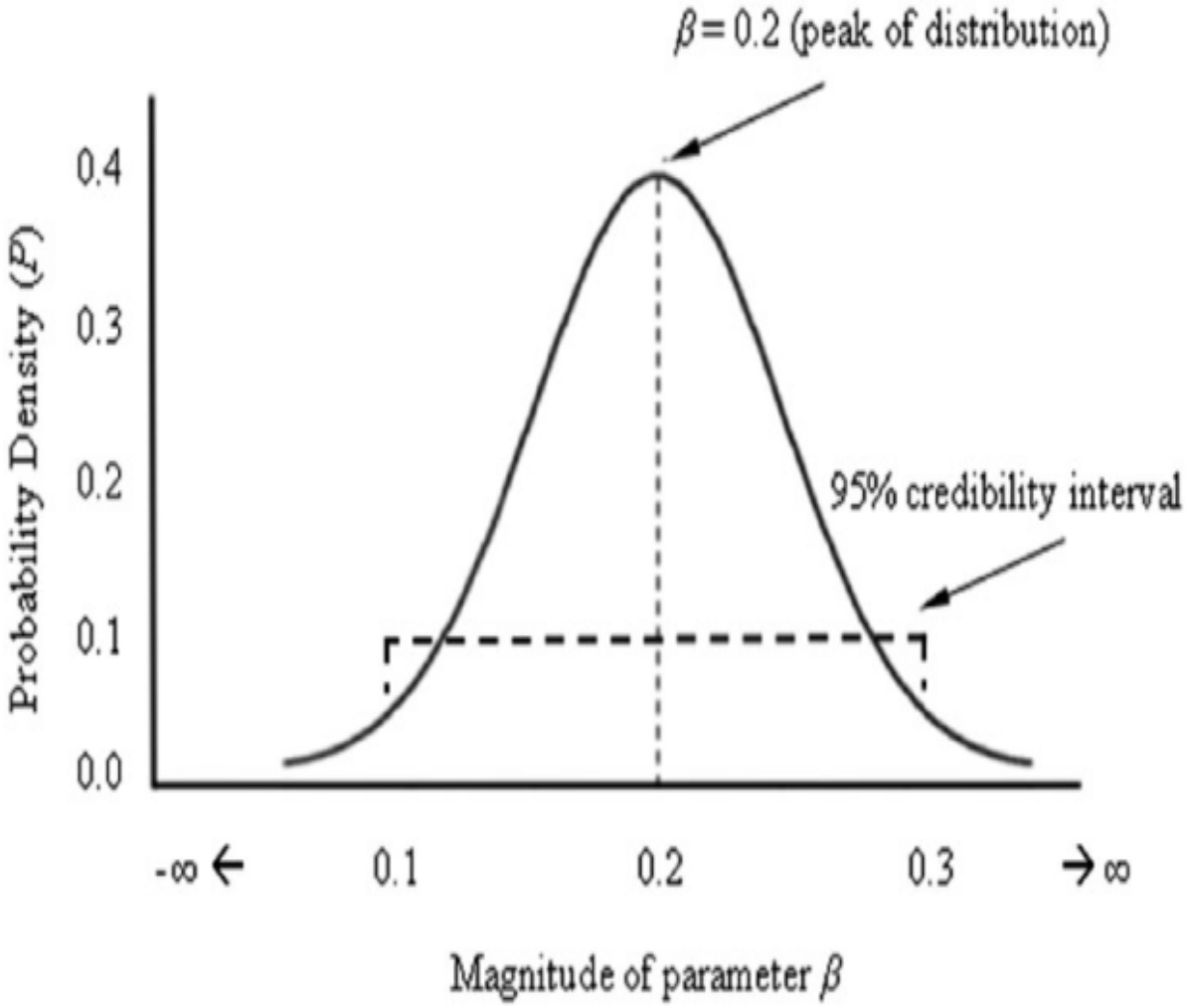
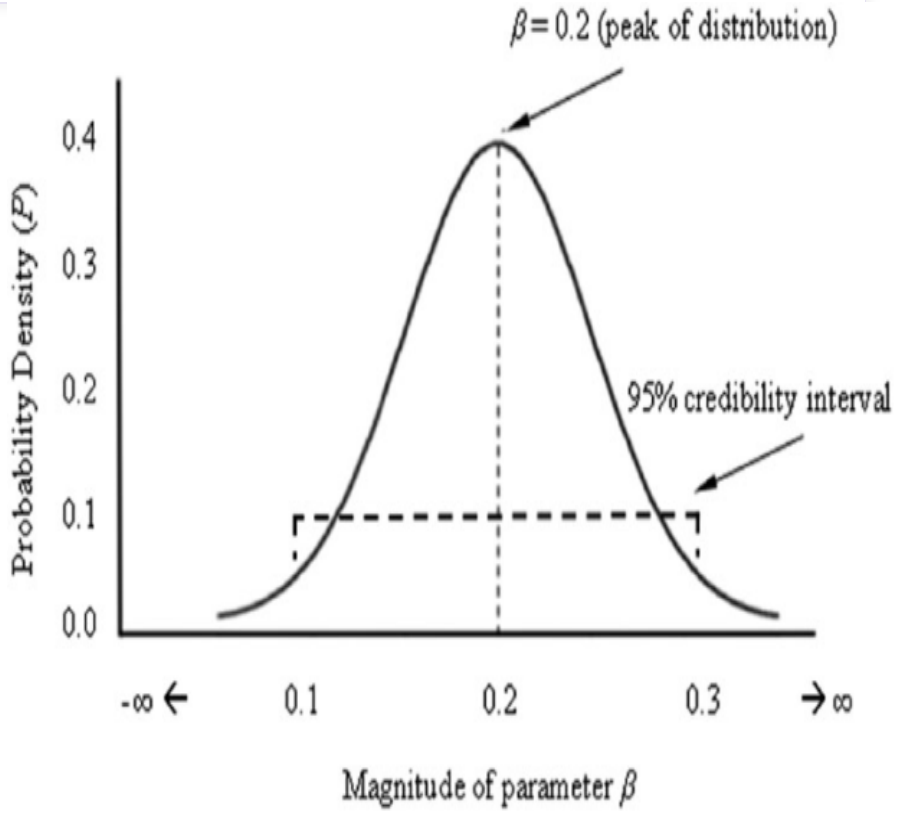


Figure 1. Hypothetical Probability Distribution for a Regression Coefficient β

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420.

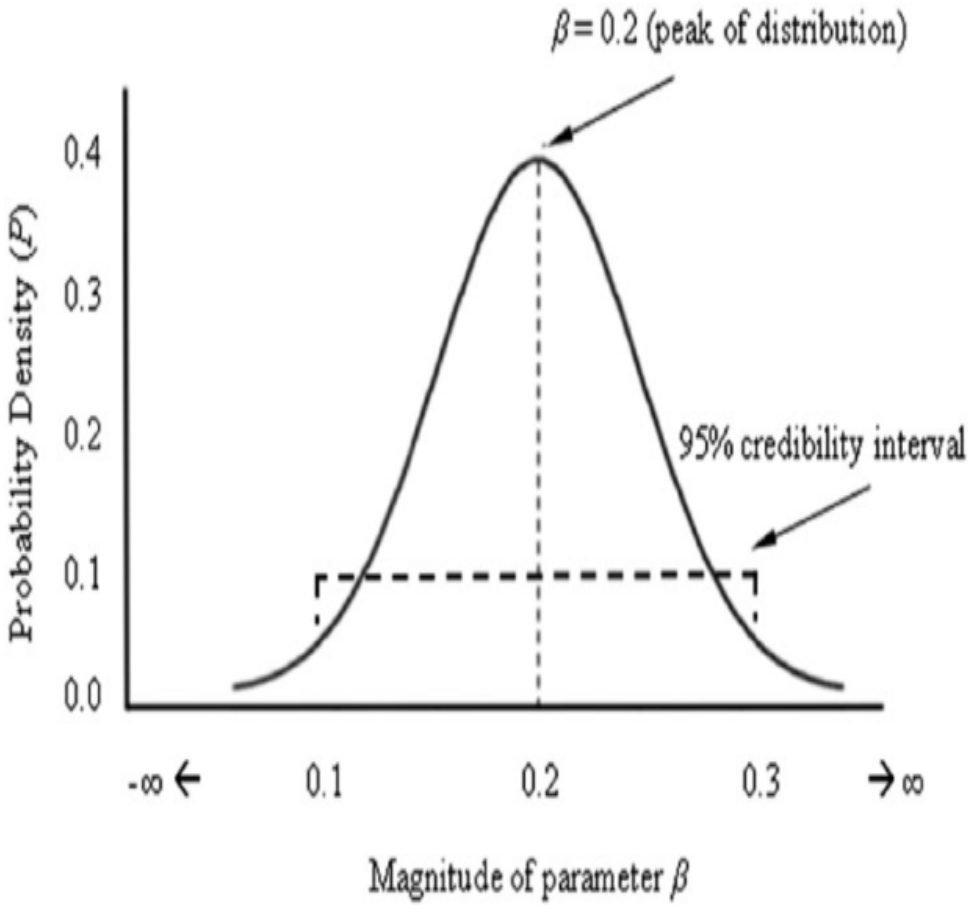
Regression: the effect of an Environment (x) on personality (y)

- The probability associated with the effect of an environment on personality is highest at b = .20, with 95% of the probability distribution falling inside **a credibility interval** between b = .10 and b = .30 showing **the most probable range of values for the effect**



Bayesian inference: Credibility Interval

- Examine the range of parameter estimates that captures 95% or 99% of the posterior probability distribution
- Shows the **most probable range of values** for the effect
- **"There is a 95% chance that the effect of an environment on personality ranges between 0.10 and 0.30."**
- Traditional confidence intervals: an infinite number of replications of a study
- Take the peak of the posterior distribution as the Bayesian estimate of a parameter



The Prior Distribution:
How will different priors influence the posterior conclusions?

(a) Informative priors based on **previous findings and theoretical predictions**

(b) Diffuse, non-informative, or uninformative priors based on **no prior knowledge or belief, or a desire to eliminate the influence of a prior distribution during estimation**

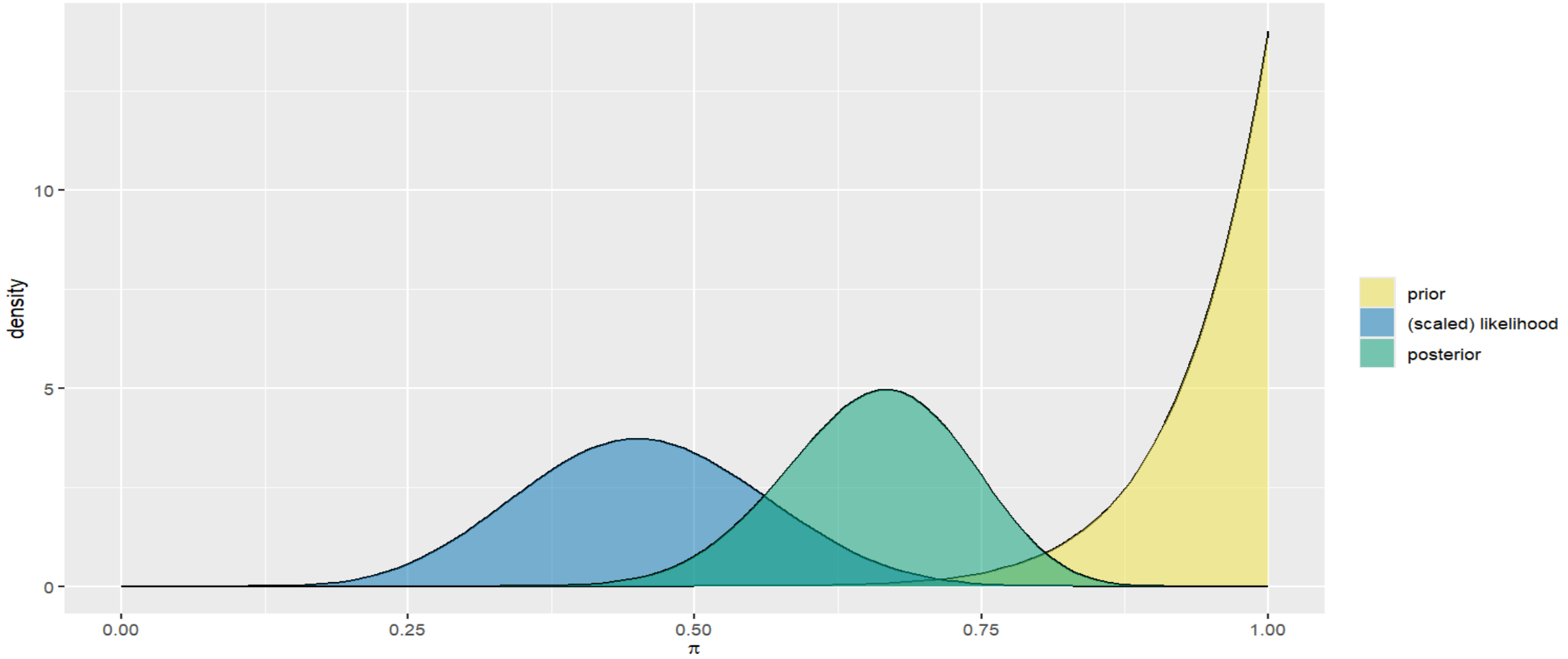
(c) Empirical priors based on **observed data**

Informative Priors

- Map **previous findings or theory** onto parameters in θ before conducting a new study
- The more **certain** the prior information, **the *smaller*** the prior variability
- Findings from a meta-analysis of the effect of feedback on performance could be used to specify a prior distribution for a standardized regression coefficient
- Example: Based on Kluger and DeNisi's (1996) finding, a Cohen's d of $\mu_d = .20$ and variance $\sigma_d^2 = .97$, the d to r transformation leads to a Bayesian prior with a mean effect $\mu_\beta = .20$ and a variance $\sigma_\beta^2 = .44$.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Probability Density Distributions of Informative Priors, Likelihood, and Posteriors



Informative Priors: Advantages

Do not have	Facilitate	Allow
to estimate parameters from scratch; **past research (findings)** can inform the current research (e.g., meta-analysis)	**small-sample research**, which involves a lot of uncertainty due to sampling error variance	estimating parameters with information from observed data (likelihoods) that can be **supplemented or augmented** with prior distributions

Uninformative (Vague, Diffuse, or Flat) Priors

- When a study is exploratory, there may be **little to no prior knowledge** that can be used for estimation.
- **Prior knowledge** may be **diffuse** because of contradictory findings or competing theories
- The prior is less informative (i.e., **nearly flat**)
- **A prior distribution with a huge variance**

- e.g., a normal distribution for an effect β with a mean $\mu_\beta = 0$ and variance $\sigma_\beta^2 = 10^{10}$.
- The default setting in some statistics programs (e.g., Mplus)

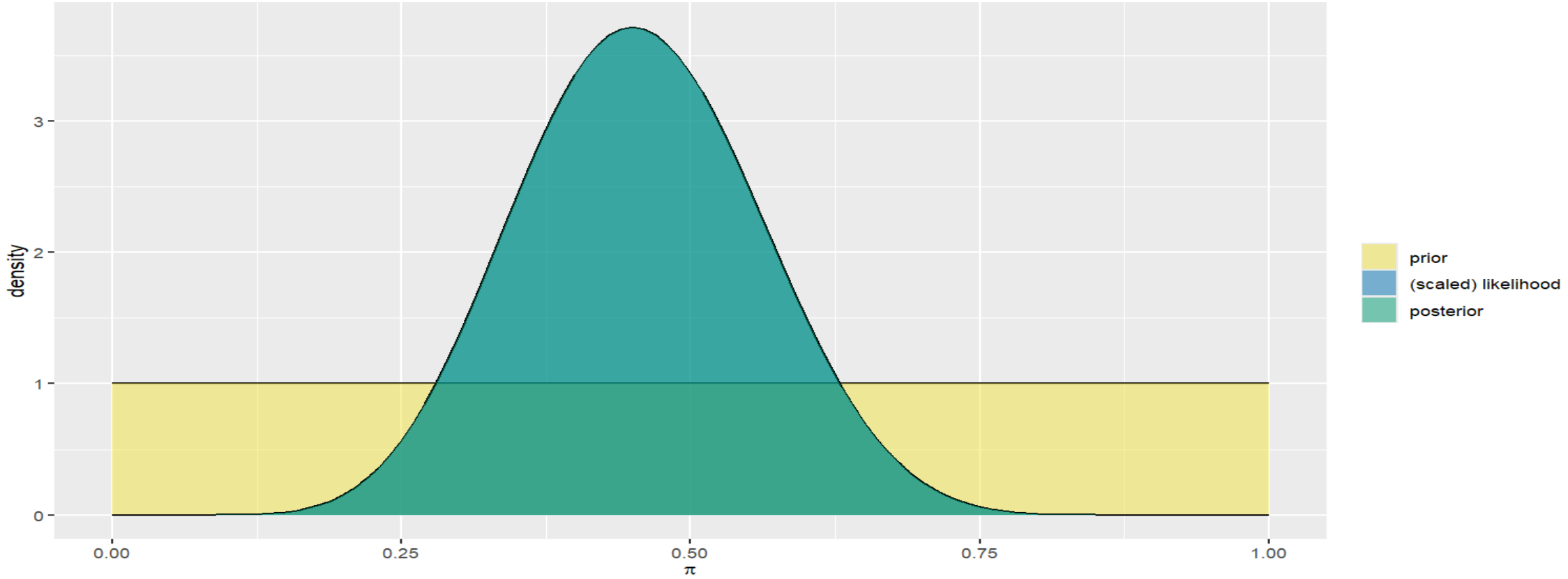
Uninformative (Vague, Diffuse, or Flat) Priors

- **Eliminate the importance of priors** in the estimation process to rely as much as possible on the likelihood (the data)

- Specify prior probabilities that allow **the data (the likelihood) to dominate** the estimation of posteriors through the likelihood

- No parameter values are more probable than others.

Probability Density Distributions of Uninformative Priors, Likelihood, and Posteriors



Uninformative (Vague, Diffuse, or Flat) Priors

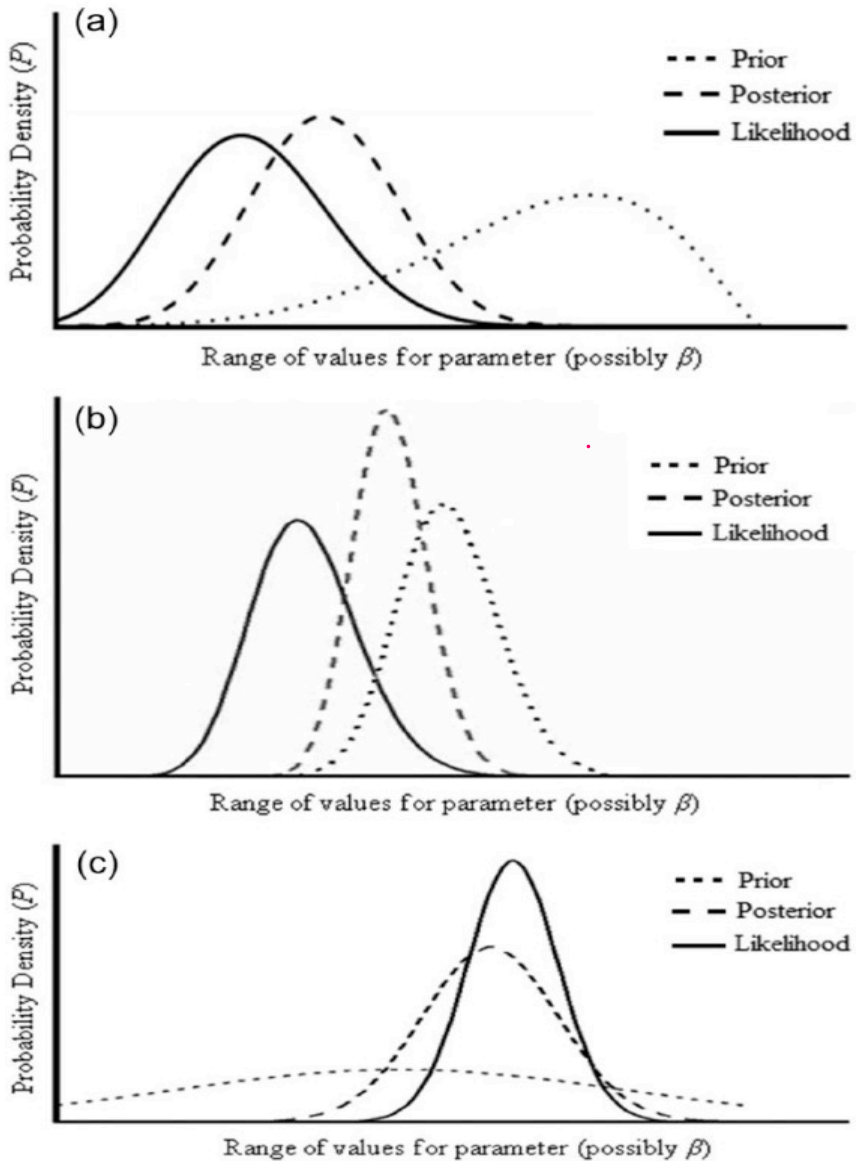
- Uninformative priors allow Bayesian estimation to mimic frequentist estimation, **because prior information does not influence results**
- **Fundamentally different from** frequentist estimation
- The "inverse" Bayesian interpretation: treat model parameters as random variables

Empirical Priors

- Empirical priors come from *empirical Bayes estimation*, where prior distributions are estimated from a **data set itself**
- Advantage: **using all observed data to estimate parameters** that are associated with only a subset of the data, as in multilevel modeling, where all of the data are used to estimate a group's mean
- **In single-level models**, empirical priors use the same observed data to estimate priors and likelihoods, thereby making an "undesirable double use of the data"
- **Priors that are similar to likelihoods lead to narrower posteriors**

Empirical Priors

- **Priors similar to likelihoods: Narrower posteriors**



Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420.

The Choice of Prior and Sample Size

- The choice of a prior distribution can be debated
- **The impact of the prior in determining posteriors is** important **for smaller sample sizes** and **diminishes as sample sizes increase**

The Choice of Prior and Sample Size

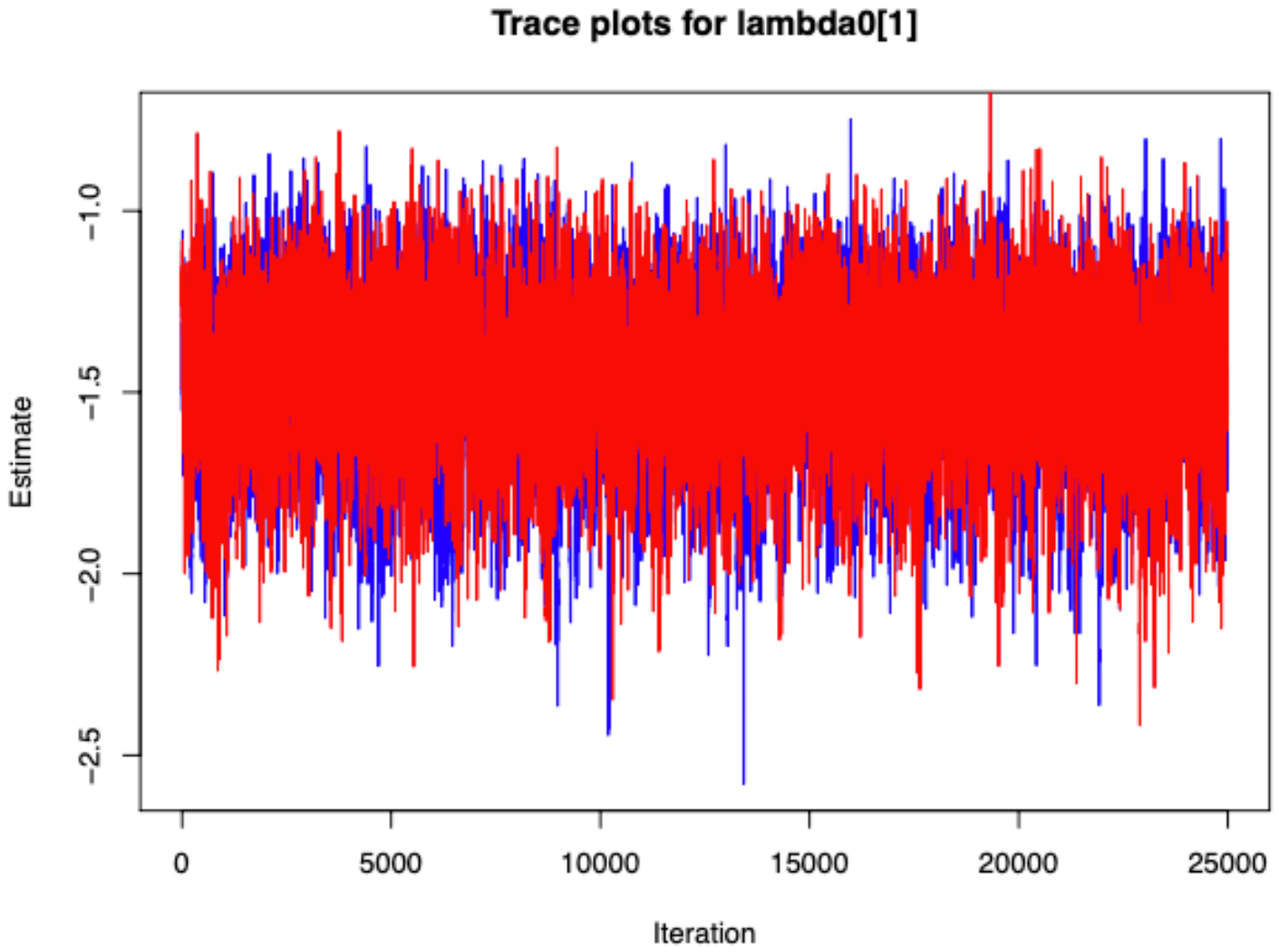
- With **large sample sizes**, we need not work hard to formulate a prior distribution
- Even when empirical or informative priors are used, results from Bayesian analysis will **converge with frequentist estimation** as **sample sizes increase** yet still provide the advantage of the probabilistic interpretation of parameters
- A study of **prior dependence**: The sensitivity of posterior distribution results to priors by using different priors for the same analysis and examining differences in posteriors

Bayesian Estimation: Markov Chain Monte Carlo (MCMC)

- Monte Carlo: The most common method
- Simulation with sampling, generating, and drawing
- Chain: The random values drawn by being linked and taking place sequentially
- The application of Markov Chains to simulate probability models
- Allows specifying many types of priors

Bayesian Estimation: Markov Chain Monte Carlo (MCMC)

- **An iterative process**: A prior distribution is specified
- Posterior values are estimated to build up and define the posterior distribution
- MCMC is carried out from at least two starting points (i.e., at least two "chains")



Assessing Convergence
The potential scale reduction (PSR)

- Convergence can be evaluated by calculating the potential scale reduction (PSR)
- **The PSR** : $\frac{\textbf{The ratio of total variance across chains}}{\textbf{Pooled variance within chain}}$
- Once this ratio reflects very little variance between chains when compared to within-chain variance (i.e., PSR < 1.05), estimation is halted because different iterative processes (i.e., different chains) yield equivalent results

Bayesian inference: Model Fit and Comparisons

- **Absolute Model Fit:**

Posterior predictive model checking (PPMC)

-Posterior predictive p-value (PPp)

- **Relative Model Fit**

Deviance Information Criterion (DIC)

- **Approximate Model Fit**

Bayesian comparative fit index (BCFI)

Bayesian Tucker–Lewis index (BTLI)

Bayesian root mean square error of approximation (BRMSEA)

Model Fit and Comparisons: Absolute Model Fit Posterior Predictive Model Checking (PPMC)

- Answers the questions
- **"Do the estimated parameters in the model produce data that look like the observed data?"**
- **"Does my model fit my data well?" : Simulated data based on the model will resemble the observed data**
- Samples posterior estimates of model parameters
- Generate a data set that is the same size as the observed data set
- The probability of the observed data and the probability of the generated data are then each estimated with χ^2 values, and the latter χ^2 value is subtracted from the former
- This is done over many iterations, which creates a distribution of **the χ^2 differences**

Model Fit and Comparisons: Absolute Model Fit Posterior Predictive Model Checking (PPMC)

- **Positive average differences** in χ^2 values: **poor model fit**, meaning the observed data have larger χ^2 values than generated data.
- **The average χ^2 difference** between observed and generated data sets **equals zero:** When the model conforms to the data, the observed and generated data are equally likely
- **Two sources of uncertainty**

- uncertainty in sample data (by comparing observed versus generated data)
- uncertainty in parameters themselves (by sampling parameters from the posterior distribution).

- Frequentist approaches assume fixed parameters and do not include this latter source of **uncertainty**

Model Fit and Comparisons: Absolute Model Fit Posterior Predictive Model Checking (PPMC)

- Quantify misfit from PPMC
- The χ^2 difference values: **posterior predictive p values (PPP)**
- PPP: The proportion of times that the observed data are more probable than the generated (simulated) data (i.e., the proportion of times observed data have a smaller $\chi 2$ than the generated data).
- **PPP values of .50: Good model fit**
- **Small PPP values (e.g., <.05): poor model fit**
- The observed data fit better than generated data very infrequently (e.g., less than 5% of the time)
- Your data are far off from their predictive distribution

Model Fit and Comparisons:
Relative Model Fit
Deviance Information Criterion

- **Deviance Information Criterion** (DIC; B.Muthén, 2010).

: Late 1990s and early 2000s

- DIC rewards models that strike a balance between parsimony and fit

- **Smaller DIC values indicate better models**
- Have issues when parameters are discrete
- Not fully Bayesian

Model Fit and Comparisons
Approximate Fit Indices: BCFI, BTLI, BRMSEA

- RMSEA, CFI, and TLI in SEM: Mplus Bayesian framework
- **Bayesian comparative fit index (BCFI)**: > .95 excellent fit
- **Bayesian Tucker–Lewis index (BTLI; Garnier-Villarreal & Jorgensen, 2020):** > .95 excellent fit
- **Bayesian root mean square error of approximation (BRMSEA; Hoofs et al., 2018)**
- Smaller the better
- Effective in large samples (N > 1000)
- Based on discrepancies between actual and replicated data at each MCMC iteration in a similar way to the PPMC technique
- Intended to be used with large sample sizes (N > 100 or 200)
- Issues with overfitting

Bayes Rules Example: Pop vs Soda vs Coke

- Our word choices can reflect where we live
- In Korea, we use the different words for Korean pancake
- Suppose you're watching an interview of somebody that lives in the United States
- Pop vs Soda dataset in the bayesrules package (Dogucu, Johnson, and Ott 2021)
- **374,250** responses to a volunteer survey conducted at popvssoda.com

Letting **A** denote **the event** that a person uses the word "pop,"
The following regional **likelihoods**: Percentage of people that use the term "pop.

L(M|A)=0.6447 : 64.47% of people in the Midwest
L(N|A)=0.2734 : 27.34 % of people in the Northeast
L(S|A)=0.0792 : 7.92% of people in the South
L(W|A)=0.2943 : 29.43% of people in West

Johnson, A. A., Ott, M. Q., & Dogucu, M. (2022). *Bayes rules!: An introduction to applied Bayesian modeling*. Chapman and Hall/CRC.

Bayes Rules Example: Pop vs Soda vs Coke

Without knowing anything about this person, U.S. Census figures provide **prior information** about the region in which they might live:
the Midwest (M), Northeast (N), South (S), or West (W).

This **prior model**

region	M	N	S	W	Total
Probability	0.21	0.17	0.38	0.24	1

Based on population statistics only,
P(M) = 0.21: there's a 21% prior probability that the interviewee lives in the Midwest
P(N) = 0.17: there's a 17% prior probability that the interviewee lives in the Northeast
P(S)=0.38: there's a 38% prior probability that the interviewee lives in the South
P(W) = 0.24: there's a 24% prior probability that the interviewee lives in the West

The South is the most populous region and the Northeast the least (P(S)>P(N))

Bayes Rules Example: Pop vs Soda vs Coke

Weighing the **prior** information about regional populations with the **data** that the interviewee used the word "pop," what results can we get?

For example, because 38% of people live in the South but that "pop" is not used a lot in that region, what's the **posterior probability** that the interviewee lives in the South?

A: the event that a person uses the word "pop"
S: South

$$P(S|A)=\frac{P(S)*L(S|A)}{P(A)}=\frac{\text{the prior probability } P(S)*\text{likelihood } L(S|A)}{\text{marginal distribution of A}}$$

= Posterior Prob of S given A

$$\mathbf{P(A)}=L(M|A)P(M)+L(N|A)P(N)+L(S|A)P(S)+L(W|A)P(W)$$
$$=0.6447\cdot 0.21+0.2734\cdot 0.17+0.0792\cdot 0.38+0.2943\cdot 0.24\approx \mathbf{0.2826}$$

-> There's a 28.26% chance that a person in the U.S. uses the word "pop"

Bayes Rules Example: Pop vs Soda vs Coke

$$P(S|A) = \frac{P(S)*L(S|A)}{P(A)} = \frac{\text{the prior probability } P(S)*\text{likelihood } L(S|A)}{\text{marginal distribution of A}}$$

= Posterior Prob of S given A

P(S) = .38 : Prior probability that the interviewee lives in the South
L(S|A) = .0792 : Likelihoods (Percentage of people in the south that use the term pop)
P(A) = .2826 : Marginal Distribution

$$P(S|A) = \frac{.38*.0792}{.2826} \approx 0.1065$$

- a roughly **10.65% posterior** chance that the interviewee lives in the South

Bayes Rules Example: Pop vs Soda vs Coke

We can similarly update our understanding of the interviewee living in the Midwest, Northeast, or West.

The resulting posterior model of region alongside the original prior.

Region	M	N	S	W	Total
Prior probability	0.21	0.17	0.38	0.24	1
Posterior probability	**0.4791**	0.1645	**0.1065**	0.2499	1

Upon hearing the interviewee use “pop,”

most likely that they live in the Midwest and least likely that they live in the South

Demo : Bayesian Structural Equation Model

Hong, H., Vispoel, W. P., & Martinez, A. J. (2024). Applying SEM, Exploratory SEM, and Bayesian SEM to Personality Assessments. *Psych*, *6*(1). https://www.mdpi.com/2624-8611/6/1/7

- IPIP-NEO-120 Agreeableness and Mplus 8.10

Vispoel, W. P., Lee, H., Xu, G., & Hong, H. (2022). Expanding bifactor models of psychological traits to account for multiple sources of measurement error. *Psychological assessment*, *34*(12), 1093. DOI: 10.1037/pas0001170

- BFI2 data and Blavaan Package in R