Introduction to Bayesian Statistics

James Swain
University of Alabama in Huntsville
ISEEM Department

Author Introduction

- James J. Swain is Professor of Industrial and Systems Engineering Management at UAH. His BA,BS, and MS are from Notre Dame, and PhD in Industrial Engineering from Purdue University.

- Research interests include statistical estimation, Monte Carlo and Discrete Event simulation and analysis.

- Celebrating 25 years at UAH with experience at Georgia Tech, Purdue, the University of Miami, and Air Products and Chemicals.

Outline

- Views of probability
- Likelihood principle
- Classical Maximum Likelihood
- Bayes Theorem
- Bayesian Formulation
- Applications
  - Proportions
  - Means
  - Approach to Hypothesis Tests
- Conclusion
- References

> “It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel.”
>
> Savage, 1954

Bayesian statistics are based on “subjective” rather than “objective” probability

Classical View of Probability

- Physical or Objective probability
  - Also known as “frequentist probability”
  - Description of physical random processes
    - Dice
    - Roulette Wheels
    - Radioactive processes

- Probability as the limiting ratio of repeated occurrences

- Philosophically --- Popper, von Mises, etc.

Subjective Probability

- Subjective or Evidential
  - Also called “Bayesian” probability
- Assessment of uncertainty
- Plausibility of possible outcomes
  - May be applied to uncertain but fixed quantities
  - Bayes analysis treats unknown parameters as random
- Sometimes elucidated in terms of wagers on outcomes

- Philosophically --- de Finetti, Savage, Carnap, etc.

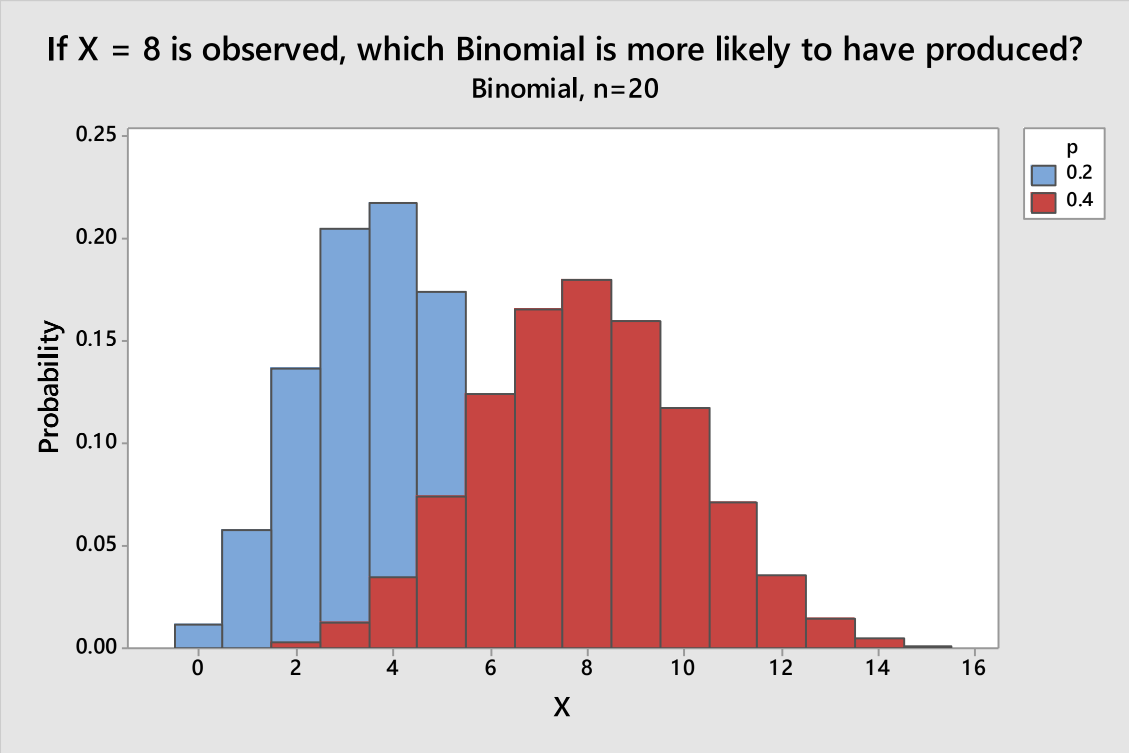
Note on Notation

- Samples: X for either single or multiple samples

- Distributions:
  - p(X) for either discrete or continuous
  - $N(\mu, \sigma^2)$ for normal with mean and variance parameters (sometimes φ)

- Parameters:
  - For binomial examples, π is population proportion of success
  - For continuous case, it is simply π

Likelihood Principle



- Example: X=8 --- source?
- b(8:0.4, 20) = 0.18
- b(8;0.2, 20) = 0.02
- Either possible
- First more likely

Likelihood (Discrete Sample)

- Likelihood function $L(\pi : X)$
  - Note: π is unknown proportion of success

$$x_i = \begin{cases} 1 & \text{Success} \\ 0 & \text{Failure} \end{cases}$$

- Here, sample X is “given”

$$L(\pi : X) = \prod_{i=1}^{n} \pi^{\,x_i}(1 - \pi)^{1 - \,x_i} = \pi^{x}(1 - \pi)^{n - x}$$

- Parameter π is variable to solve

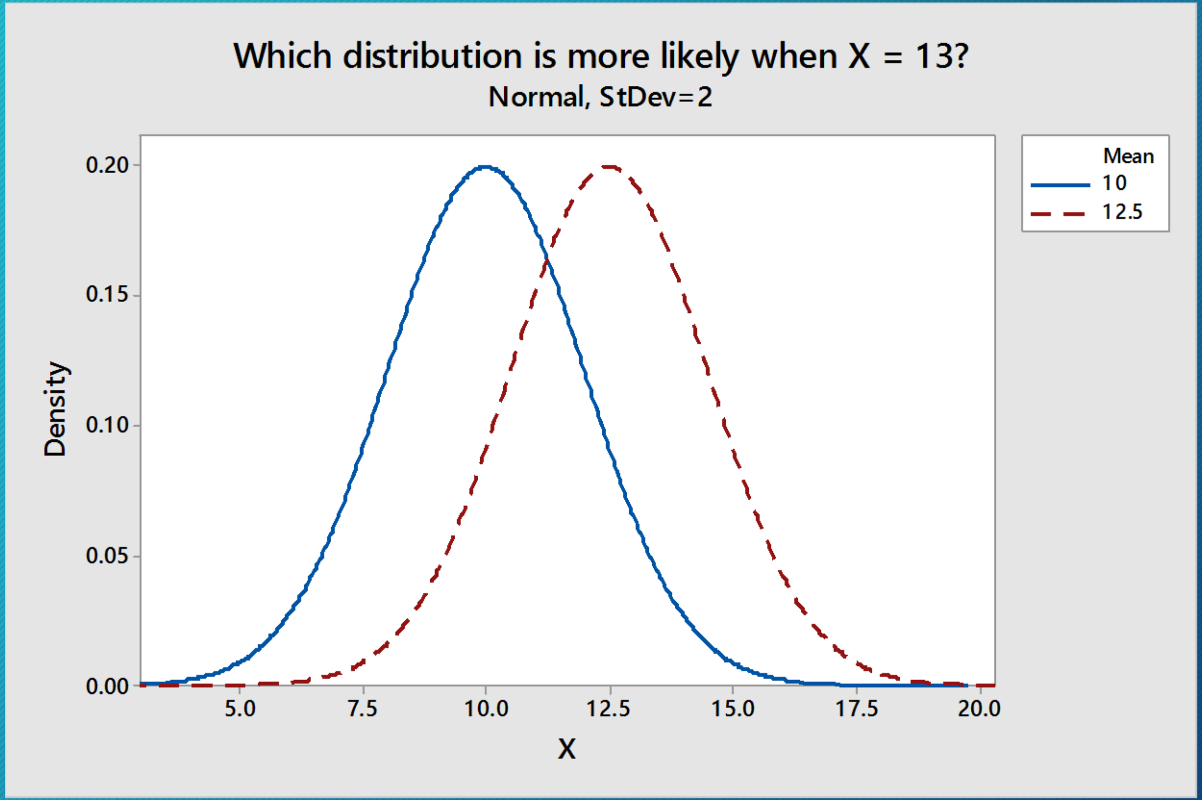
$$l(\pi : X) = \ln L(\pi : X) = x \ln \pi + (n - x) \ln 1 - \pi$$

- Solution is MLE (max. likelihood estimator)

$$\frac{dl}{d\pi} = 0 \quad \xrightarrow{yields} \quad p = \hat{\pi} = \frac{x}{n}$$

Likelihood for continuous

- For sample x = 13
- Normal(12.5, 4)
  - f(13: 12.5, 4) = 0.19
- Normal (10, 4)
  - f(13: 10, 4) = 0.065



MLE for Continuous (Normal)

- Multiple samples (n)

$$L\left(\mu,\sigma^2 : X\right) = \prod_{i=1}^{n} \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

- Likelihood L

$$l\left(\mu,\sigma^2 : X\right) = \ln L\left(\mu,\sigma^2 : X\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_{i=1}^{n}(\,x_i-\mu)^2$$

- Parameters chosen to max L using Ln L

- Estimation (MLE)

$$\frac{dl}{d\mu} = 0 \qquad\qquad \hat{\mu} = \bar{x}$$

$$\frac{dl}{d\sigma^2} = 0 \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\,x_i-\bar{x})^2}{n} = \frac{n\;s^{\;\;2}}{n-1}$$

Classical Statistics: Summary

- Theory is well developed
  - E.g., most cases asymptotically unbiased and normal
  - Variance parameters based on log likelihood derivatives

- Confidence Intervals
  - Confidence level means ?

- Hypothesis Tests
  - Pesky p-value

“There’s no theorem like Bayes’ theorem
Like no theorem we know
Everything about it is appealing
Everything about it is a wow
Let out all that a priori feeling
You’ve been concealing right up to now!”

George E. P. Box

Bayes Theorem useful, controversial (in its day), and the bane of introductory probability students. Sometimes known as “inverse probability”

Bayes Theorem I: Sets and Partition of S

- Consider a sample space S
- Event $B \subset S$
- Partition of sample space $A_1, A_2, \ldots, A_k$
  - Exhaustive
  - Mutually exclusive

$$S = A_1 \ \cup A_2 \cup \ldots \cup A_k$$
$$A_i \cap A_j = \emptyset$$

$$B = B \cap S = (B \cap A_1) \ \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k)$$

- Decomposition of B into subsets $B \cap A_i$

- By Axioms of probability

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i)$$

Bayes Theorem II: Conditional Probability

- Conditional probability

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

- Multiplication rule

$$P(A_i \cap B) = P(A_i)\, P(B|A_i)$$

- Total probability

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(A_i)P(B|A_i)$$

Bayes III: The Theorem

- Direct result of previous definitions
  - Multiplication rule
  - Total Probability
- Conditionals are “reversed” or “inverted”

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)\,P(B|A_i)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}$$

Quick Bayes Theorem Example

- Let C = “conforming”
- A part is sampled, tested, and found to be conforming: $P(A_3|C)$ = ?

Vendor	P(C | Vendor) (%)	Vendor Proportion (%)
A_1	88	30
A_2	85	20
A_3	95	50

Bayes Statistical Formulation

- Parameter θ uncertainty
- Prior distribution $p(\theta)$
- Conditional on sample: Likelihood
- Posterior distribution $p(\theta|X)$
- Posterior proportional to numerator

$$p(\theta|X) = \frac{L(\theta|X)p(\theta)}{p(X)}$$

$$p(X) = \begin{cases} \sum_{All\ i} L(\theta_i|X)\, p(\theta_i) & \textit{Discrete case} \\ \int L(\theta|X)\, p(\theta)\, d\theta & \textit{Continuous Case} \end{cases}$$

$$p(\theta|X) \propto \; L(\theta|X)\, p(\theta)$$

Proportionality

- Denominator
    - Complex
    - A constant
- Key information is numerator
- Standardize numerically if needed
- Bayes methods frequently computer intensive

$$p(\theta|X) = \frac{L(\theta|X)p(\theta)}{p(X)}$$

$$p(\theta|X) \propto \; L(\theta|X)\, p(\theta)$$

Updating Posterior

- Two independent samples X, Y
- Posterior from X is prior to Y
- Updated posterior

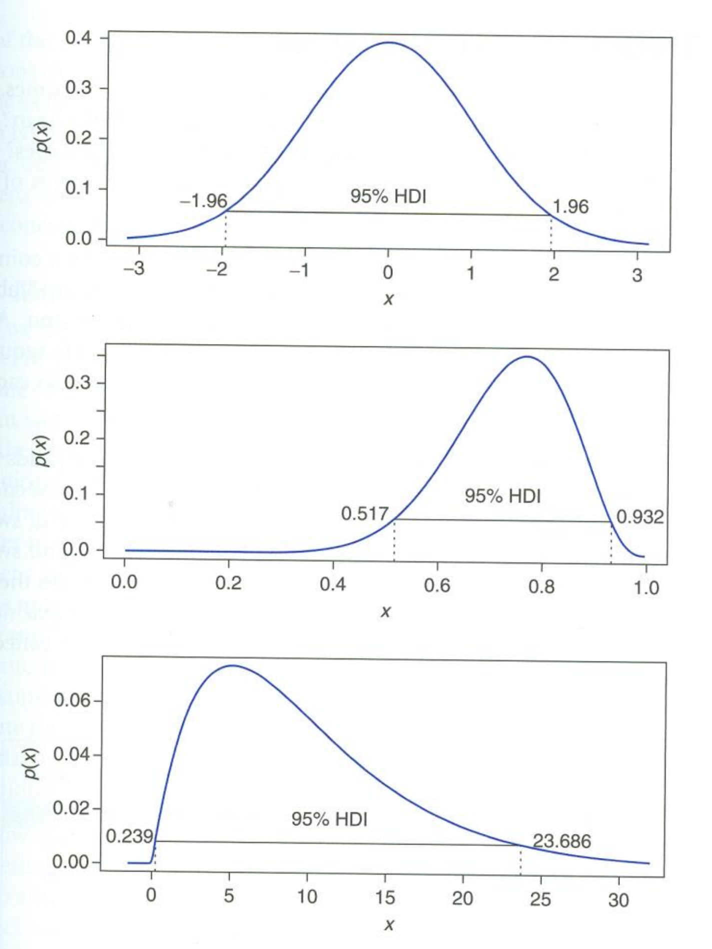
$$p(\theta|X) = \frac{L(\theta|X)\,p(\theta)}{p(X)}$$

$$p(\theta|X,Y) = \frac{L(\theta|Y)\,p(\theta|X)}{p(Y)} = \frac{L(\theta|Y)\,L(\theta|X)\,p(\theta)}{p(Y)p(X)}$$

High Density Intervals (HDI)

- Summarizing Posterior Distribution
- High Density Intervals
- Probability content (e.g., 95%)
- Region with largest density values
  - Similar for symmetric (e.g., normal)
  - May differ for skewed (e.g., Chi-squared)
  - Special tables may be required
  - Source: Krusche, p. 41

- Special tables for HDI (e.g., inverse log F)



Bayes Analysis: Binomial Sample (Beta prior)

- Estimating proportion of success π

$$x_i = \begin{cases} 1 & \mathit{Success} \\ 0 & \mathit{Failure} \end{cases}$$

- Noted: Likelihood

$$L(\pi\colon X) = \prod_{i=1}^{n} \pi^{\,x_i}(1-\pi)^{1-\,x_i} = \pi^{x}(1-\pi)^{n-x}$$

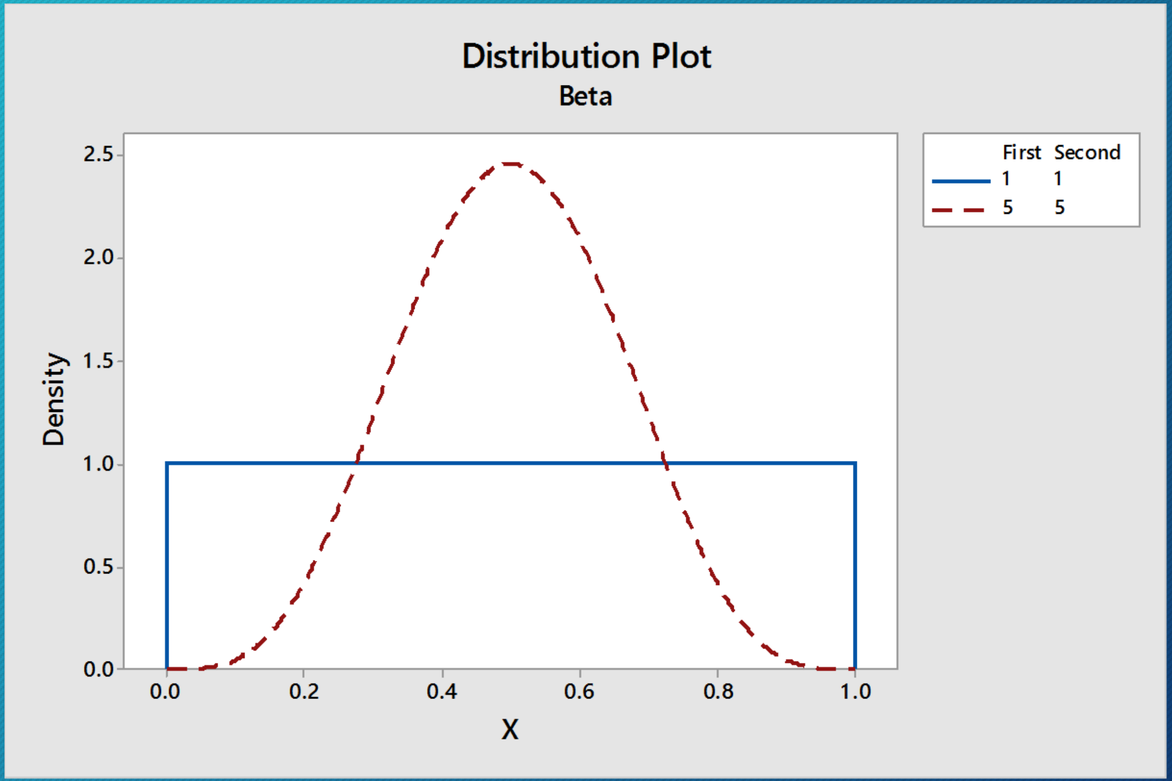
- What prior? Beta distribution is typical

Beta Prior for Proportion Success

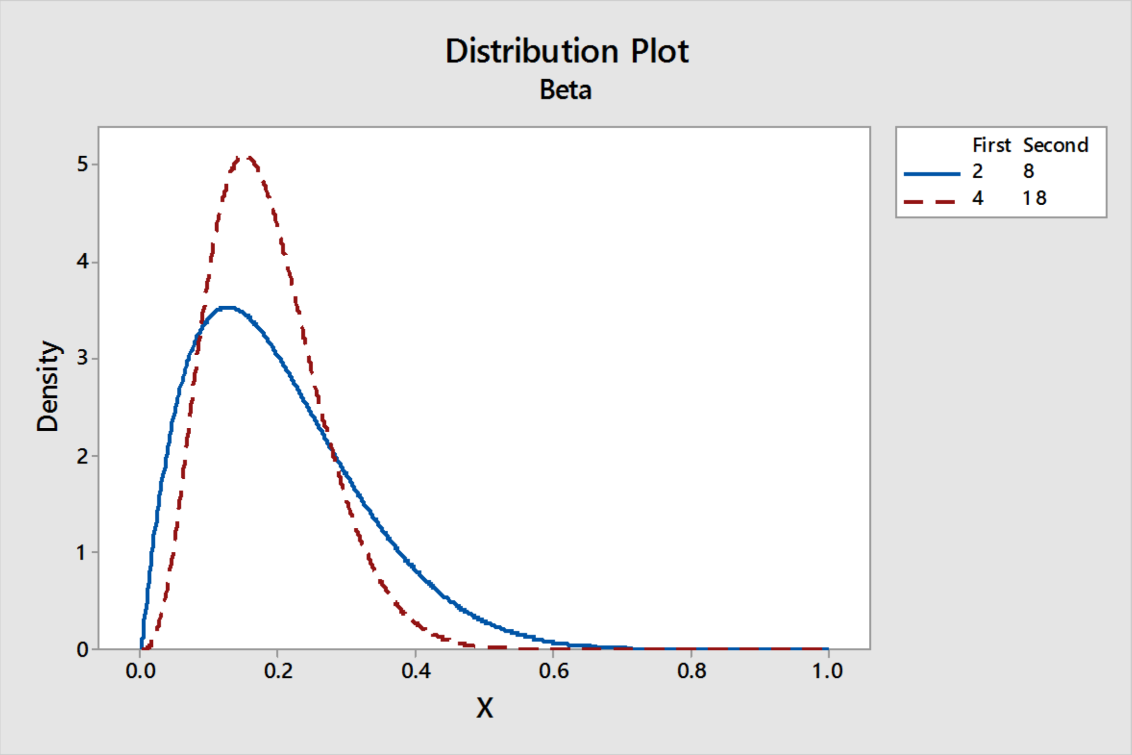
- Uncertainty about proportion success (π)
- Similarity to likelihood
- “Uninformative” prior

$$p(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Beta Prior for Proportion Success



- Mean and variance formulas

$$E[\Pi] = \frac{\alpha}{\alpha + \beta}$$

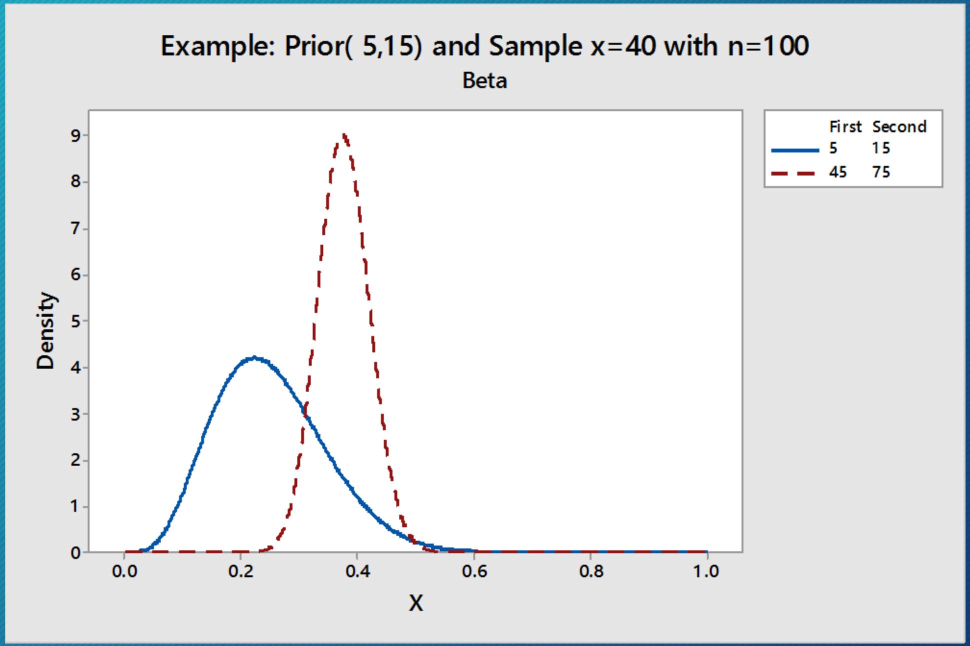
$$\text{Var}(\Pi) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Example: Small Poll

- Prior
  - “Worth” sample 20
  - Estimated Mean 25%

$$\alpha_0 = \pi_0\, n_0$$
$$\beta_0 = (1 - \pi_0)\, n_0$$
$$\pi_0 = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

- Sample
  - Sample size 100
  - Vote favorable 40 (40%)



$$\pi_1 = \frac{x + \alpha_0}{n + \alpha_0 + \beta_0} = \frac{x}{n}\,\frac{n}{n + \alpha_0 + \beta_0} + \frac{\alpha_0}{\alpha_0 + \beta_0}\,\frac{\alpha_0 + \beta_0}{n + \alpha_0 + \beta_0}$$

Conjugate Priors

- Prior / Posterior distributions that are similar

$$p(\pi) \sim \pi^{\alpha-1}(1-\pi)^{\beta-1}$$

- Conjugate: Posterior/Prior same form

$$L(\pi : X) = \pi^{x}(1-\pi)^{n-x}$$

- Identification of distributional constants simplified

$$p(\pi|X) \sim \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}$$

Simple Normal-Normal Case for Mean

- Simplified Case
  - Fixed variance
  - Mean μ has prior parameters μ_0, φ_0
- Derivation steps omitted (Ref: Lee)

$$\mu \sim N(\mu_0, \varphi_0)$$
$$x \sim N(\mu, \varphi)$$

$$p(\mu|x) \propto \exp\left(-\frac{\mu^2(\varphi_0^{-1} + \varphi^{-1})}{2} + \mu\left(\frac{\mu_0}{\varphi_0} + \frac{x}{\varphi}\right)\right)$$

Results: Normal-Normal

- Summarized results

- Posterior mean weighted value

$$\varphi_1 = \frac{1}{\varphi_0^{-1} + \varphi^{-1}}$$

$$\mu_1 = \varphi_1 \left(\frac{\mu_0}{\varphi_0} + \frac{x}{\varphi} \right)$$

$$\mu \sim N(\mu_1, \varphi_1)$$

$$x \sim N(\mu, \varphi)$$

$$\mu_1 = \mu_0 \frac{\varphi_0^{-1}}{\varphi_0^{-1} + \varphi^{-1}} + x \frac{\varphi^{-1}}{\varphi_0^{-1} + \varphi^{-1}}$$

Example: Carbon Dating I

- Ennerdale granophyre
- Earliest dating: K/Ar in 60’s, 70’s
- Prior: 370 M-yr, ± 20 M-yr

- Later Rb/Sr 421 ± 8 (M-yr)

- Posterior estimate

Carbon Dating II

- Substitute Alternative Expert Prior
- 400 M-yr, ± 50 M-yr

- Revised posterior

- Note posterior roughly same

Normal-Normal (Multiple Samples)

- Similar arrangement
- Multiple samples
- Variances are constant

$$\mu \sim N(\mu_0, \varphi_0)$$

$$x_i \sim N(\mu, \varphi)$$

$$L(\mu|x) \propto \exp(-\frac{\mu^2\left(\varphi_0^{-1} + n\varphi^{-1}\right)}{2} + \mu\left(\frac{\mu_0}{\varphi_0} + \frac{\sum x_i}{\varphi}\right)$$

Results: Multiple Samples

- Similar analysis
- As n increases, posterior converges to $\bar{x}$

$$\varphi_1 = \frac{1}{\varphi_0^{-1} + n\varphi^{-1}}$$

$$\mu_1 = \varphi_1\left(\frac{\mu_0}{\varphi_0} + \frac{\bar{x}}{\varphi/n}\right)$$

$$\mu \sim N(\mu_1, \varphi_1)$$
$$x \sim N(\mu, \varphi/n)$$

$$\mu_1 = \mu_0 \frac{\varphi}{\varphi + n\varphi_0} + \bar{x}\,\frac{n\varphi_0}{\varphi + n\varphi_0}$$

Example: Men’s Fittings

- Modified from Lee (p. 46)
- Experience: Chest measurement N(38, 9)
- Shop data 39.8, ±2 on n = 890 samples

Normal: Prior in Variance

- Default prior in mean uniform
- Improper for infinite range

$$p(\varphi) = \frac{1}{\varphi}$$

- Variance
  - Uniformity in value (like mean)?
  - Transform: uniform in log φ

$$p(\varphi) \propto \varphi^{-\frac{\nu}{2}-1} \exp\left(-\frac{S_0}{2\varphi}\right)$$

- Conjugate prior in Variances
  - Inverse Chi-squared distribution

$$p(\varphi|X) \propto \varphi^{-\frac{\nu+n}{2}-1} \exp\left(-\frac{S_0+S}{2\varphi}\right)$$

Conjugate Prior for Normal Variance

- Inverse chi-squared distribution
- Exact match not critical (likelihood will dominate)

$$\varphi \sim S_0\, \chi_v^{-2}$$

$$E[\varphi] = \frac{S_0}{\nu - 2}$$

$$Var[\varphi] = \frac{2S_0^2}{(\nu - 2)^2(\nu - 4)}$$

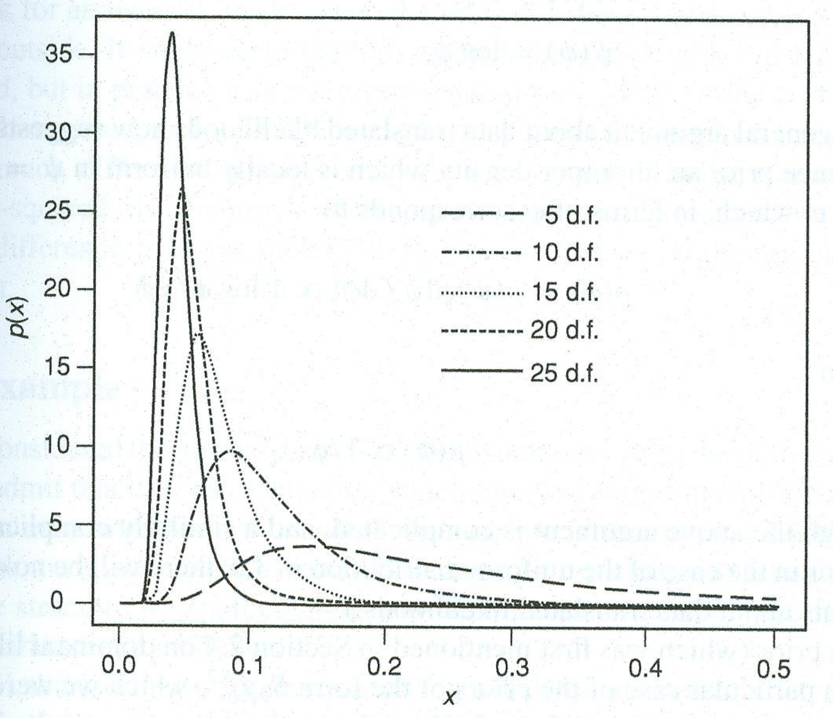


Figure 2.1 Examples of inverse chi-squared densities.

General Approach Hypothesis Tests

- Hypothesis Tests: Choice between two alternatives

$$H_0\!: \theta \in \Theta_0 \quad H_1\!: \theta \in \Theta_1$$

- Alternatives disjoint and exhaustive of course

$$\pi_0 = \theta \in \Theta_0 \quad \pi_1 = \theta \in \Theta_1$$

- Priors

$$p_0 = P(\theta \in \Theta_0 | X) \quad p_1 = P(\theta \in \Theta_1 | X)$$

- Posterior

$$\pi_0 + \pi_1 = 1 \quad p_0 + p_1 = 1$$

Hypothesis Tests --- General

- Prior “odds”

$$\frac{\pi_0}{\pi_1} = \frac{\pi_0}{1 - \pi_0}$$

- Posterior “odds”

$$\frac{p_0}{p_1} = \frac{p_0}{1 - p_0}$$

- Bayes Factor B
  - Odds favoring H_0 against H_1

$$B = \frac{\left(\frac{p_0}{p_1}\right)}{\left(\frac{\pi_0}{\pi_1}\right)}$$

- Two simple alternatives
  - Larger B denotes favor for H_0

$$B = \frac{p(X|\theta_0)}{p(X|\theta_1)}$$

Summary and Conclusion

- Bayesian estimation
  - Since parameters random, provides probability statements about values
  - Provides mechanism for non-sample information
  - Avoids technical difficulties with likelihood along

- More advanced analysis computationally intensive
  - Tied to computer software
  - Open source R with application BUGS

- Machine learning heavily dependent on Bayesian methods

References

- Lee (2012), *Bayesian Statistics: An Introduction*, 4Ed., Wiley

- Lindley (1972), *Bayesian Statistics, A Review*, SIAM.

- Krusche (2011), *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*, Academic Press/Elsevier.