STAT7630: Bayesian Statistics
Lecture Slides # 5

Chapter 6 Approximating the Posterior & Chapter 7 - MCMC under the Hood

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: September, 2024

Outline

Introduction

Gamma-Poisson Conjugate Family

Posterior Intervals

- Bayesian vs Frequentist Intervals

Normal-Normal Conjugacy

Conjugate Families Introduction

- **Conjugate prior:** A prior distribution is conjugate if the posterior is in the same family as the prior.
- More formally, a **conjugate prior** is a prior distribution that, combined with the data model, results in a posterior distribution with the same functional form as the prior, but with updated parameter values.
- **Example:** For a Beta prior in a Binomial likelihood, the posterior is also Beta.
- That is, in the Beta-Binomial model, the Beta prior is conjugate since the posterior is also a Beta distribution.

$$f(\theta|y) \propto f(\theta)L(\theta|y) \implies \text{Posterior: Beta}(\alpha + y, \beta + n - y)$$

- This property allows for straightforward posterior analysis.

Revisiting Choice of Prior

- When choosing a prior, we consider:
  - **Computational ease:** Is the posterior easy to compute?
  - **Interpretability:** Can we interpret the prior vs. data contribution?
- **Example: Beta-Binomial model**

$$\text{Posterior: Beta}(\alpha + y, \beta + n - y)$$

- The influence of the data depends on the sample size n and y relative to α and β.

Why are Conjugate Priors Nice?

- **Why conjugate priors?**
  - They make Bayesian analysis easier by simplifying the computation of the posterior.
  - Conjugate priors allow for tractable posterior distributions.
- Conjugate priors are advantageous because:
  1. They reduce the computational burden. The posterior can be derived without complex computations.
  2. Posterior models are easy to interpret. It's easier to understand how the prior and data contribute to the posterior.
- Next, we will explore an example of non-conjugate prior and then other Bayesian models with conjugate priors.

Non-Conjugate Priors

- Non-conjugate priors make posterior computation harder.
- Consider a non-conjugate prior for π:

$$f(\pi) = e - e^{\pi}, \quad \pi \in [0, 1]$$

- The resulting posterior:

$$f(\pi | y = 10) \propto (e - e^{\pi})\pi^{10}(1 - \pi)^{40}$$

- This posterior is messy and not easy to interpret or compute.

Example: Non-Conjugate Posterior

- A non-conjugate prior leads to complex posterior:

$$f(\pi|y=10) = \frac{(e-e^{\pi})\pi^{10}(1-\pi)^{40}}{\int_0^1 (e-e^{\pi})\pi^{10}(1-\pi)^{40} d\pi}$$

- This is hard to compute, requiring numerical integration.
- Conjugate priors would avoid this complexity.

Outline

Introduction

Gamma-Poisson Conjugate Family

Posterior Intervals

- Bayesian vs Frequentist Intervals

Normal-Normal Conjugacy

The Poisson Distribution

- The **Poisson distribution** is a widely used model for count data, where the possible values are nonnegative integers $(0, 1, 2, \ldots)$.
- It is parameterized by $\lambda > 0$. Given λ, the probability mass function (pmf) of a Poisson random variable $Y \mid \lambda$ is:

$$f(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- For a random sample of n independent counts $Y_1, Y_2, \ldots, Y_n$, the likelihood function is the product of the individual pdfs:

$$f(y_1|\lambda)f(y_2|\lambda)\cdots f(y_n|\lambda)$$

Poisson Data Model

- **Poisson model:** The number of independent events in a fixed time period.

$$Y|\lambda \sim \text{Pois}(\lambda)$$

- The Poisson probability mass function (pmf) is:

$$f(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- Mean and variance:

$$\mathbf{E}(Y|\lambda) = \text{Var}(Y|\lambda) = \lambda$$

Choice of Prior for the Poisson Model

- When modeling data with a **Poisson distribution**, a suitable prior for λ should have support on $(0, \infty)$, as $\lambda > 0$.
- The **Gamma distribution** is a good choice, as its support is $(0, \infty)$.
- In this class, we use a different parameterization of the Gamma distribution than in STAT 7600.
- Specifically, we use the Gamma distribution with:
  - **Shape parameter**: s
  - **Rate parameter**: r
- The Gamma pdf is:

$$f(\lambda) = \frac{r^s}{\Gamma(s)} \lambda^{s-1} e^{-r\lambda}, \quad \lambda > 0$$

- Note: The rate parameter r is the reciprocal of the scale parameter used in other parameterizations.

The Gamma/Poisson Bayesian Model

- If the data $Y_1, \dots, Y_n$ are iid **Poisson**(λ), then a **Gamma(s, r)** prior on λ is conjugate.
- **Likelihood**:

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum y_i}}{\prod_{i=1}^{n} y_i!}$$

- **Prior**:

$$f(\lambda) = \frac{r^s}{\Gamma(s)}\lambda^{s-1}e^{-r\lambda}, \quad \lambda > 0$$

- **Posterior** (using proportionality):

$$f(\lambda|\mathbf{y}) \propto \lambda^{\sum y_i + s - 1}e^{-(n+r)\lambda}, \quad \lambda > 0$$

- The posterior distribution is **Gamma**($\sum y_i + s, n + r$), confirming conjugacy!

Gamma-Poisson Conjugate Family

- **Poisson distribution:** A common model for count data.

$$Y_i|\lambda \sim \text{Pois}(\lambda)$$

- The conjugate prior for λ is a **Gamma** distribution:

$$\lambda \sim \text{Gamma}(s, r)$$

- The resulting posterior is also Gamma:

$$\lambda|\mathbf{y} \sim \text{Gamma}\left(s + \sum y_i, r + n\right)$$

Properties of the Gamma (Mean)

- In the **shape/rate** parameterization, the mean of the Gamma(s, r) prior distribution is:

$$\mathbf{E}(\lambda) = \frac{s}{r}$$

- We select the hyperparameters s and r based on our prior beliefs about λ.
- The mean of the Gamma$(\sum y_i + s, n + r)$ posterior distribution is:

$$\mathbf{E}(\lambda|\mathbf{y}) = \frac{\sum y_i + s}{n + r}$$

- This posterior mean also serves as a **Bayesian estimator** of λ.

Properties of the Gamma (Variance)

- Once we have a good estimate of the prior mean of λ, how do we choose s and r for the prior?
- In the **shape/rate** parameterization, the variance of the $\text{Gamma}(s, r)$ prior distribution is:

$$\text{Var}(\lambda) = \frac{s}{r^2}$$

- The **prior variance** (or standard deviation) helps inform our choice of s and r.
- Visualizing the potential prior using the `plot_gamma()` function in the `bayesrules` package can assist in selecting appropriate values for the prior.

The Posterior Mean in the Gamma/Poisson Bayesian Model

- The posterior mean for λ is:

$$\hat{\lambda}_B = \frac{\sum y_i + s}{n + r} = \frac{\sum y_i}{n + r} + \frac{s}{n + r}$$

- This can be rewritten as:

$$\hat{\lambda}_B = \left(\frac{n}{n + r}\right)\left(\frac{\sum y_i}{n}\right) + \left(\frac{r}{n + r}\right)\left(\frac{s}{r}\right)$$

- As $n \to \infty$, the **data** receives more weight in the posterior mean.

Example: Fraud Risk Phone Calls

- The textbook presents an example involving data on the number of fraud risk phone calls per day, modeled by a **Poisson distribution**.
- The parameter of interest is λ, the mean number of fraud risk calls per day.
- **Prior belief**: The average number of calls per day is approximately 5.
- We choose s and r such that $s/r = 5$.
- Additionally, we believe λ is very likely to fall between 2 and 7.
- Let's plot several potential priors with $s/r = 5$ to explore possible choices (refer to the `R` examples).

Plot of Gamma priors with $s/r = 5$

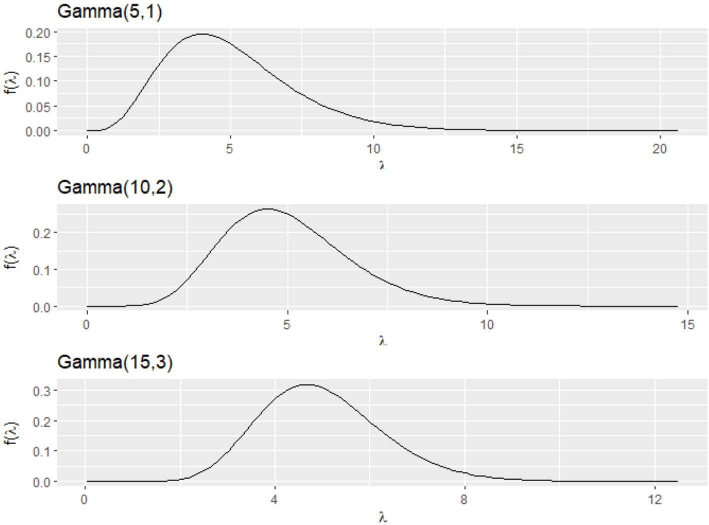


Figure 1: Gamma(s, r) priors with $s/r = 5$.

Example: Fraud Risk Phone Calls (Posterior Calculation)

- We choose $s = 10$ and $r = 2$, which align with our prior beliefs.
- That is, we use a Gamma(10, 2) prior: $\lambda \sim \text{Gamma}(10, 2)$
- Data collected ($n = 4$): 6, 2, 2, 1 ($\sum y_i = 11 \quad \text{and} \quad \bar{y} = 2.75$).
- The posterior distribution is:

$$\text{Gamma}\left(\sum y_i + s, n + r\right) = \text{Gamma}(11 + 10, 4 + 2) = \text{Gamma}(21, 6)$$

- A Bayesian estimate of λ is the posterior mean:

$$\frac{21}{6} = 3.5$$

- Compare this to the prior mean of 5 calls/day.
- Visualize the `R` plots to see how the data updated our prior beliefs.

Example: Fraud Risk Phone Calls (Posterior Calculation)

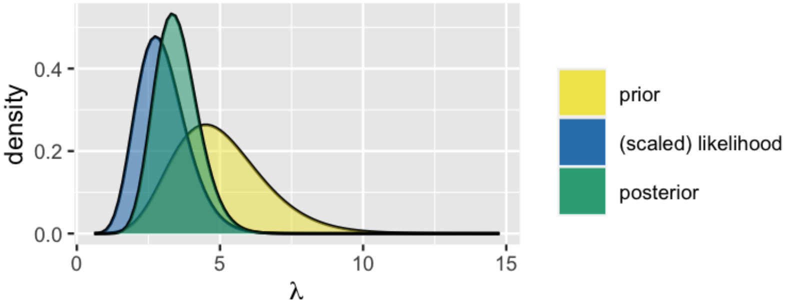


Figure 2: The Gamma-Poisson model of λ, the daily rate of fraud risk calls.

Outline

Introduction

Gamma-Poisson Conjugate Family

Posterior Intervals

- Bayesian vs Frequentist Intervals

Normal-Normal Conjugacy

Bayesian Inference: Posterior Intervals

- Simple summaries like the **posterior mean** $\mathbf{E}[\theta|\mathbf{y}]$ and **posterior variance** $\text{Var}[\theta|\mathbf{y}]$ are helpful for understanding θ.
- **Quantiles** of the posterior distribution $p(\theta|\mathbf{y})$, such as the **posterior median**, provide additional useful insights about θ.
- The ideal summary of θ is an **interval** (or region) with a specified probability of containing θ.
- Unlike a Bayesian posterior interval, a classical **confidence interval** does not directly provide this interpretation.

Bayesian Credible Intervals

- A **credible interval** (or more generally, a credible set) is the Bayesian counterpart to a confidence interval.
- A $100(1-\alpha)$ % credible set C is a subset of the parameter space Θ s.t.:

$$\int_C p(\theta|\mathbf{y})\, d\theta = 1-\alpha$$

- If Θ is a discrete set, the integral is replaced by a summation.

Quantile-Based Intervals

- If θ_L^* is the $\alpha/2$ posterior quantile and θ_U^* is the $1 - \alpha/2$ posterior quantile, then (θ_L^*, θ_U^*) forms a 100(1 - α)% **credible interval** for θ.
- Key relationships:

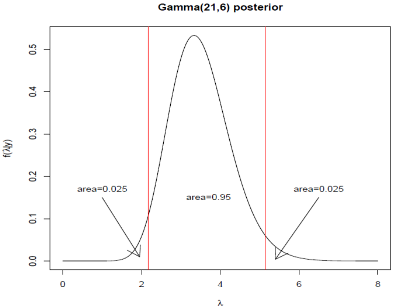
$$P[\theta < \theta_L^* | \mathbf{y}] = \frac{\alpha}{2} \quad \text{and} \quad P[\theta > \theta_U^* | \mathbf{y}] = \frac{\alpha}{2}$$

- Therefore, the credible interval satisfies:

$$P[\theta \in (\theta_L^*, \theta_U^*) | \mathbf{y}] = 1 - P[\theta \notin (\theta_L^*, \theta_U^*) | \mathbf{y}] = 1 - \alpha$$

Quantile-Based Intervals: Example

- The figure shows the **Gamma(21, 6)** posterior distribution.
- The interval between 2.17 and 5.15 represents the central 95% of the posterior distribution.
- This is a **95% credible interval** for λ, meaning the posterior probability that λ falls within this interval is 0.95.
- The tails on either side represent the remaining 5% of the distribution, split evenly with 2.5% in each tail.



Example 2: Quantile-Based Interval

- Consider an experiment with 10 flips of a coin, where the probability of heads is θ.
- We observe 2 heads in the experiment.
- The number of heads follows a **binomial distribution**:

$$p(y|\theta) = \binom{10}{y}\theta^{y}(1-\theta)^{10-y}, \quad y = 0, 1, \ldots, 10$$

- We assume a **uniform prior** for θ:

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1$$

Example 2: Quantile-Based Interval (Posterior)

- The posterior distribution is given by:

$$p(\theta|y) \propto p(\theta)L(\theta|y) = (1)\binom{10}{y}\theta^y(1-\theta)^{10-y}$$

- Simplifying:

$$p(\theta|y) \propto \theta^y(1-\theta)^{10-y}, \quad 0 \le \theta \le 1$$

- This is a **Beta distribution** for θ with parameters $y+1$ and $10-y+1$.
- Since $y=2$, the posterior is:

$$p(\theta|y=2) \sim \text{Beta}(3,9)$$

- The 0.025 and 0.975 quantiles of $\text{Beta}(3,9)$ are (0.0602, 0.5178), forming a **95% credible interval** for θ.

Example 2: Quantile-Based Interval (Posterior)

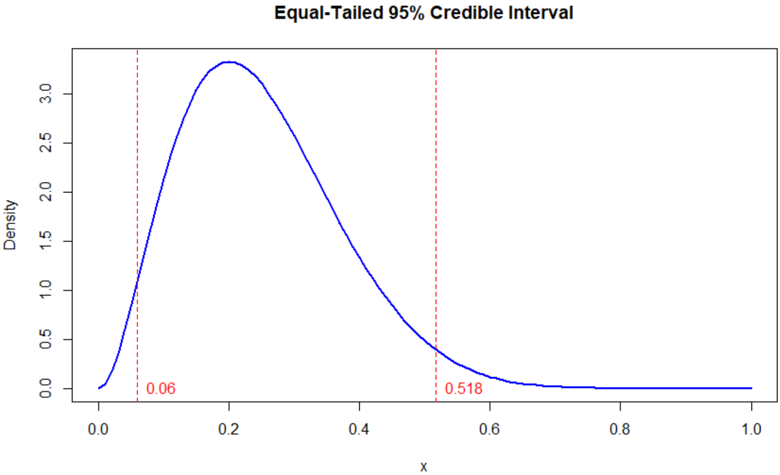


Figure 4: 95% equal tail credible interval for Beta$(3, 9)$ posterior.

HPD Intervals / Regions

- The **equal-tail credible interval** is most effective when the posterior distribution is symmetric.
- However, if the posterior distribution $p(\theta|\mathbf{y})$ is **skewed**, the equal-tail interval might not be the best choice.
- In such cases, the **Highest Posterior Density (HPD) interval** is preferred, as it identifies the region with the highest posterior probability density that covers a specified probability.

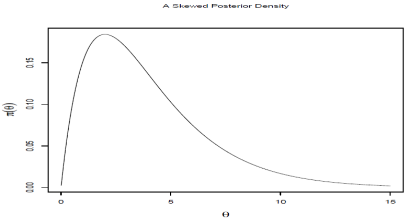


Figure 5: A skewed posterior distribution.

HPD Intervals / Regions

- Notice that values of θ around 1 have a much higher **posterior probability** than values around 7.5.
- However, in the **equal-tail interval**, 7.5 is included, while 1 is not!
- A more appropriate approach in this case is to construct an **HPD interval**, which includes the θ-values with the **highest posterior density**.

HPD Intervals / Regions: Definition

- **Definition:** A 100(1 - α)% **HPD region** for θ is a subset $C \subset \Theta$ defined as:

  $$C = \{\theta : p(\theta|\mathbf{y}) \geq k\}$$

  where k is the largest value such that:

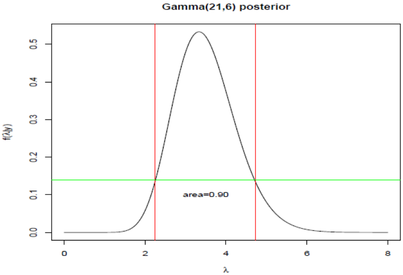
  $$\int_C p(\theta|\mathbf{y})\, d\theta = 1 - \alpha$$

- The value k can be visualized as a **horizontal line** over the posterior density. The intersections of this line with the posterior density define regions with probability $1 - \alpha$.

HPD Intervals / Regions: Example

- The figure shows the **Gamma(21, 6)** posterior distribution and the corresponding **95% HPD interval**.
- The values between $\theta_L^* = 2.25$ and $\theta_U^* = 4.72$ have the highest posterior density.
- This region contains 90% of the posterior probability:

$$P\{\theta_L^* < \theta < \theta_U^*\} = 0.90$$



Example 2: Coin Toss Example

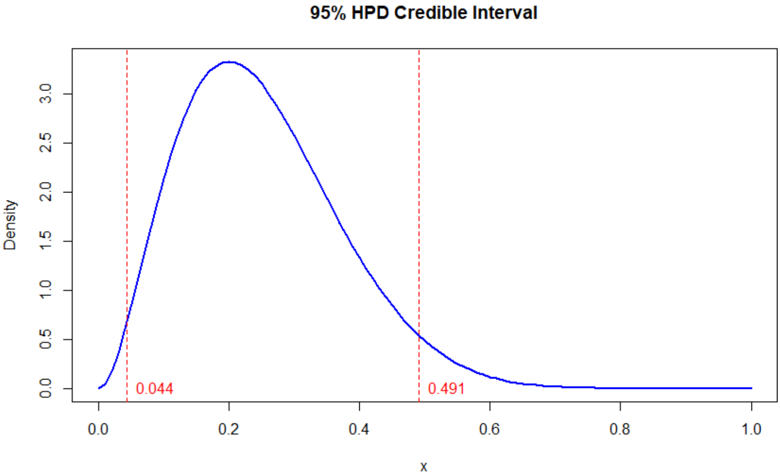
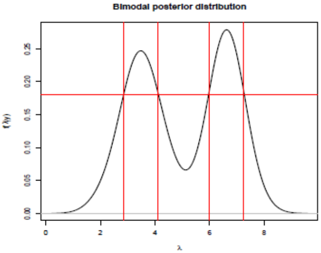


Figure 7: 95% HPD credible interval for Beta$(3, 9)$ posterior.

HPD Intervals / Regions: Unimodal vs. Multimodal

- The **HPD region** is an interval when the posterior distribution is **unimodal**.
- However, for a **multimodal** posterior, the HPD region may be a **discontiguous set**.
- **Example:** In a bimodal posterior distribution, the HPD region might consist of two separate intervals:

$$\{\theta : \theta \in (2.85, 4.1) \cup (6.0, 7.25)\}$$



Example 1 Revisited: HPD Interval

- Refer to Canvas for `R` code to find an **HPD interval** for λ in the fraud risk call example.
- The **90% quantile-based credible interval** for λ is $(2.167, 5.148)$.
- You can also use the `hpd()` function from the `TeachingDemos` package in `R` to calculate the HPD interval, yielding $(2.345, 4.844)$
- Check the `R` code for Example 2 (coin-flipping data) on Canvas.

Outline

Introduction

Gamma-Poisson Conjugate Family

Posterior Intervals

- Bayesian vs Frequentist Intervals

Normal-Normal Conjugacy

Bayesian vs Frequentist Coverage

Definition: A random interval $(L_f(\mathbf{Y}), U_f(\mathbf{Y}))$ has $100(1-\alpha)\%$ frequentist coverage for θ if, before the data are gathered:

$$P[L_f(\mathbf{Y}) < \theta < U_f(\mathbf{Y}) | \theta] = 1 - \alpha$$

(Pre-experimental $1-\alpha$ coverage)

Note: If we observe $\mathbf{Y} = \mathbf{y}$ and plug $\mathbf{y}$ into the confidence interval formula:

$$P[L_f(\mathbf{y}) < \theta < U_f(\mathbf{y}) | \theta] = \begin{cases} 0, & \text{if } \theta \notin (L_f(\mathbf{y}), U_f(\mathbf{y})) \\ 1, & \text{if } \theta \in (L_f(\mathbf{y}), U_f(\mathbf{y})) \end{cases}$$

(NOT Post-experimental $1-\alpha$ coverage)

Bayesian vs Frequentist Coverage

Definition: An interval $(L_b(\mathbf{y}), U_b(\mathbf{y}))$, based on the observed data $\mathbf{Y} = \mathbf{y}$, has $100(1-\alpha)\%$ Bayesian coverage for θ if:

$$P[L_b(\mathbf{y}) < \theta < U_b(\mathbf{y}) | \mathbf{y}] = 1 - \alpha$$

(Posterior (i.e. post-experimental) $1 - \alpha$ coverage)

The frequentist interpretation is less desirable when performing inference about θ based on a single interval.

Frequentist Coverage for Bayesian Intervals

Hartigan (1966) showed that for standard posterior intervals, an interval with $100(1-\alpha)\%$ Bayesian coverage will have:

$$P[L_b(\mathbf{Y}) < \theta < U_b(\mathbf{Y}) | \theta] = (1-\alpha) + \epsilon_n,$$

where $|\epsilon_n| < \frac{a}{n}$ for some constant a.

That is,

Frequentist coverage (of the Bayesian interval) $\rightarrow 1{-}\alpha$ as $n \rightarrow \infty$.

Note: Many classical confidence interval methods only achieve $100(1-\alpha)\%$ frequentist coverage asymptotically, as well.

Outline

Introduction

Gamma-Poisson Conjugate Family

Posterior Intervals

Bayesian vs Frequentist Intervals

Normal-Normal Conjugacy

The Normal-Normal Model

- Why is the **normal distribution** so frequently used to model data?
- Many quantities in nature are approximately normally distributed.
- The **Central Limit Theorem (CLT)** suggests that any variable that is a sum of independent components will be approximately normal.
- Additionally, when sampling from a normal population, $\bar{Y}$ (sample mean) and S^2 (sample variance) are independent.
- If beliefs about the mean are independent of beliefs about the variance, using a **normal model** is often appropriate.

Why Normal Models?

- The **normal model** is analytically convenient due to its properties, including being part of the **exponential family** and having sufficient statistics $\bar{Y}$ and S^2.
- Inference about the population mean based on a normal model remains **correct as** $n \to \infty$, even if the data are not truly normal.
- By assuming a **normal likelihood**, we can obtain a wide range of **posterior distributions** by choosing different priors.

A Conjugate Analysis with Normal Data (Variance Known)

- **Simple scenario:** Assume data $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, with μ unknown and σ^2 known.
- For example, **Normal model:** Continuous data such as hippocampal volumes.
- The goal is to make inference about μ.
- The **likelihood function** is:

$$L(\mu|\mathbf{y}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

- The parameter of interest, μ, can take values from $-\infty$ to ∞.
- A **conjugate prior** for μ is $\mu \sim N(\delta, \tau^2)$, with:

$$p(\mu) = (2\pi\tau^2)^{-1/2} e^{-\frac{1}{2\tau^2}(\mu-\delta)^2}$$

A Conjugate Analysis with Normal Data (Variance Known): Posterior

- The posterior distribution is obtained by combining the likelihood and the prior: $p(\mu|\mathbf{y}) \propto p(\mu)L(\mu|\mathbf{y})$
- Substituting in the expressions:

$$\propto e^{-\frac{1}{2\tau^2}(\mu-\delta)^2} \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} = e^{-\frac{1}{2\tau^2}(\mu-\delta)^2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

- Collecting and simplifying the exponent:

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 + \frac{1}{\tau^2}(\mu-\theta)^2\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\mu + \mu^2) + \frac{1}{\tau^2}(\mu^2 - 2\mu\delta + \delta^2)\right]\right)$$

- This expression shows the posterior distribution as a product of Gaussian terms from the likelihood and the prior.

A Conjugate Analysis with Normal Data (Variance Known): Posterior

- The posterior distribution is given by:

$$p(\mu|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\cdot\frac{1}{\sigma^2\tau^2}\left[\tau^2\sum y_i^2-2\tau^2 n\bar{y}\mu+n\tau^2\mu^2+\sigma^2\mu^2-2\sigma^2\mu\delta+\sigma^2\delta^2\right]\right)$$

- Simplifying further:

$$\propto \exp\left(-\frac{1}{2}\cdot\frac{1}{\sigma^2\tau^2}\left[\mu^2(\sigma^2+n\tau^2)-2\mu(\delta\sigma^2+\tau^2 n\bar{y})+(\delta^2\sigma^2+\tau^2\sum y_i^2)\right]\right)$$

- Finally, the posterior can be written as:

$$\propto \exp\left(-\frac{1}{2}\left[\mu^2\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)-2\mu\left(\frac{\delta}{\tau^2}+\frac{n\bar{y}}{\sigma^2}\right)+k\right]\right)$$

where k is a constant.

A Conjugate Analysis with Normal Data (Variance Known): Final Posterior

- The posterior distribution simplifies to:

$$p(\mu|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\left(\mu^2 - 2\mu\left(\frac{\delta}{\tau^2}+\frac{n\bar{y}}{\sigma^2}\right)\right)+k\right]\right)$$

- Further simplifying:

$$p(\mu|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\left(\mu - \frac{\delta/\tau^2 + n\bar{y}/\sigma^2}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)^2\right]\right)$$

- Thus, the posterior distribution is normally distributed as

$$\mu|\mathbf{y} \sim N\left(\frac{\delta/\tau^2 + n\bar{y}/\sigma^2}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)$$

A Conjugate Analysis with Normal Data (Variance Known): Posterior Summary

- The posterior distribution for μ is a **normal distribution** with:
  - **Mean**: $\dfrac{\delta/\tau^2 + n\bar{y}/\sigma^2}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$
  - **Variance**: $\left(\dfrac{1}{\tau^2} + \dfrac{n}{\sigma^2}\right)^{-1} = \dfrac{\tau^2\sigma^2}{\sigma^2 + n\tau^2}$
- The **precision** is the reciprocal of the variance:
  - $\frac{1}{\tau^2}$ is the **prior precision**.
  - $\frac{n}{\sigma^2}$ is the **data precision**.
  - $\frac{1}{\tau^2} + \frac{n}{\sigma^2}$ is the **posterior precision**.

A Conjugate Analysis with Normal Data (Variance Known): Posterior Mean

- The posterior mean $\mathbf{E}[\mu|\mathbf{y}]$ is:

$$\mathbf{E}[\mu|\mathbf{y}] = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\delta + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\bar{y}$$

- This is a weighted average of the **prior mean** δ and the **sample mean** $\bar{y}$.
- If the **prior is highly precise** (small τ^2), more weight is placed on δ.
- If the **data are highly precise** (large n), more weight is placed on $\bar{y}$.
- As $n \to \infty$, $\mathbf{E}[\mu|\mathbf{y}] \approx \bar{y}$; and $\text{Var}[\mu|\mathbf{y}] \approx \frac{\sigma^2}{n}$ when τ^2 is large.
- This shows that for large τ^2 and n, Bayesian and frequentist inference about μ will be nearly identical.

A Conjugate Analysis with Normal Data (Mean Known)

- Now assume $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, with μ known and σ^2 unknown.
- We aim to make inference about σ^2.
- The **likelihood function** is:

$$L(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2\right]\right)$$

- Let $W = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2$, which is the **sufficient statistic**.
- The conjugate prior for σ^2 is the **inverse gamma distribution**.
- If $Y \sim \text{Gamma}(\alpha, \beta)$, then $1/Y \sim \text{Inverse Gamma}(\alpha, \beta)$.
- The prior for σ^2 is:

$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right), \quad \sigma^2 > 0$$

where $\alpha > 0$ and $\beta > 0$.

A Conjugate Analysis with Normal Data (Mean Known)

- The prior **mean** and **variance** of σ^2 are:

$$\mathbf{E}(\sigma^2) = \frac{\beta}{\alpha - 1} \text{ for } \alpha > 1; \;\; \text{Var}(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \text{ for } \alpha > 2$$

- The posterior distribution for σ^2 is: $p(\sigma^2|\mathbf{y}) \propto p(\sigma^2) L(\sigma^2|\mathbf{y})$
- Substituting the likelihood and prior:

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{n}{2\sigma^2} w} (\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}$$

$$\propto (\sigma^2)^{-(\alpha + \frac{n}{2} + 1)} \exp\left(-\frac{\beta + \frac{n}{2} w}{\sigma^2}\right)$$

- Hence, the posterior distribution is an **Inverse Gamma**:

$$\sigma^2|\mathbf{y} \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{n}{2} w\right)$$

where $w = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$. **Conjugate!**

- $\implies \sigma^2|\mathbf{y} \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2\right)$

A Conjugate Analysis with Normal Data (Mean Known): Choosing Prior Parameters

- How do we choose the prior parameters α and β?
- The parameters can be determined from the prior mean, $\mathbf{E}(\sigma^2) = m$ and variance of σ^2, $\text{Var}(\sigma^2) = s_p^2$:

$$\alpha = \frac{m^2}{s_p^2} + 2$$

$$\beta = m\left(\frac{m^2}{s_p^2} + 1\right)$$

- Thus, by making reasonable **guesses** about m and s_p^2, we can determine α and β for the inverse gamma prior.

A Model for Normal Data (Mean and Variance Both Unknown)

- When $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$ with both μ and σ^2 unknown, the conjugate prior for μ depends explicitly on σ^2.
- The prior for σ^2 is:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

- The prior for $\mu|\sigma^2$ is:

$$p(\mu|\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2/s_0}(\mu-\delta)^2}$$

- The parameter s_0 represents the analyst's confidence in the prior specification.
- When s_0 is large, it indicates strong confidence in the prior belief about μ.

A Model for Normal Data (Mean and Variance Both Unknown): Joint Posterior

- The joint posterior distribution for (μ, σ^2) is:

$$p(\mu, \sigma^2|\mathbf{y}) \propto p(\sigma^2)p(\mu|\sigma^2)L(\mu, \sigma^2|\mathbf{y})$$

- Substituting the likelihood and priors:

$$\propto (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2-\frac{1}{2\sigma^2/s_0}(\mu-\delta)^2}$$

$$\propto (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} \exp\left(-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\mu)^2 + \frac{1}{s_0}(\mu-\delta)^2\right)\right)$$

- Expanding the squares in the exponent:

$$= (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} e^{-\frac{1}{2\sigma^2}\left(\sum y_i^2-2n\bar{y}\mu+n\mu^2\right)-\frac{1}{2\sigma^2/s_0}\left(\mu^2-2\mu\delta+\delta^2\right)}$$

A Model for Normal Data (Mean and Variance Both Unknown): Joint Posterior

- Simplifying:

$$= (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{1}{2}} \exp\left(-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}\left(\sum y_i^2 - n\bar{y}^2\right)\right)$$

$$\times(\sigma^2)^{-1} \exp\left(-\frac{1}{2\sigma^2}\left((n+s_0)\mu^2 - 2(n\bar{y}+\delta s_0)\mu + (n\bar{y}^2 + s_0\delta^2)\right)\right)$$

- The second part is a normal kernel for μ.

A Model for Normal Data (Mean and Variance Both Unknown): Posterior for σ^2

- To obtain the posterior for σ^2, we integrate out μ from the joint posterior:

$$p(\sigma^2|\mathbf{y}) = \int_{-\infty}^{\infty} p(\mu, \sigma^2|\mathbf{y})\, d\mu$$

- This results in:

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{1}{2}} \exp\left(-\frac{1}{\sigma^2}\left[\beta + \frac{1}{2}\left(\sum y_i^2 - n\bar{y}^2\right)\right]\right)$$

  since the term involving μ integrates to a normalizing constant.

- Hence, since $-\alpha - \frac{n}{2} - \frac{1}{2} = -\left(\alpha + \frac{n}{2} - \frac{1}{2}\right) - 1$, we see that the posterior for σ^2 is inverse gamma.

$$\sigma^2|\mathbf{y} \sim \mathrm{IG}\left(\alpha + \frac{n}{2} - \frac{1}{2}, \beta + \frac{1}{2}\sum(y_i - \bar{y})^2\right)$$

A Model for Normal Data (Mean and Variance Both Unknown): Posterior for μ

- The posterior distribution for μ given σ^2 and $\mathbf{y}$ is:

$$p(\mu|\sigma^2, \mathbf{y}) = \frac{p(\mu, \sigma^2|\mathbf{y})}{p(\sigma^2|\mathbf{y})}$$

- After simplification, the posterior is:

$$p(\mu|\sigma^2, \mathbf{y}) \propto \sigma^{-2} \exp\left(-\frac{1}{2\sigma^2}\left[(n+s_0)\mu^2 - 2(n\bar{y} + \delta s_0)\mu + (n\bar{y}^2 + s_0\delta^2)\right]\right)$$

- This simplifies further to:

$$p(\mu|\sigma^2, \mathbf{y}) \propto \sigma^{-2} \exp\left(-\frac{1}{2\sigma^2/(n+s_0)}\left[\mu^2 - 2\frac{n\bar{y} + \delta s_0}{n+s_0}\mu + \frac{n\bar{y}^2 + s_0\delta^2}{n+s_0}\right]\right)$$

- Clearly, $\mu|\sigma^2, \mathbf{y}$ follows a normal distribution:

$$\mu|\sigma^2, \mathbf{y} \sim N\left(\frac{n\bar{y} + \delta s_0}{n+s_0}, \frac{\sigma^2}{n+s_0}\right)$$

A Model for Normal Data (Mean and Variance Both Unknown): Limiting Cases

- The **conditional posterior mean** is a weighted average of the sample mean $\bar{y}$ and the prior mean δ:

$$\left(\frac{n}{n+s_0}\right)\bar{y}+\left(\frac{s_0}{n+s_0}\right)\delta$$

- The **relative sizes** of n and s_0 determine the weighting of $\bar{y}$ (the sample mean) and δ (the prior mean):
  - When n is large, more weight is placed on $\bar{y}$.
  - When s_0 is large, more weight is placed on δ.
- As $s_0 \rightarrow 0$, the posterior distribution for $\mu|\sigma^2, \mathbf{y}$ approaches:

$$\mu|\sigma^2, \mathbf{y} \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

A Model for Normal Data (Mean and Variance Both Unknown): Marginal Posterior for μ

- The marginal posterior for μ is obtained by integrating out σ^2:

$$p(\mu|\mathbf{y}) = \int_0^\infty p(\mu, \sigma^2|\mathbf{y})\, d\sigma^2$$

- Substituting the joint posterior:

$$p(\mu|\mathbf{y}) \propto \int_0^\infty (\sigma^2)^{-\alpha-\frac{n}{2}-\frac{3}{2}} \exp\left(-\frac{2\beta + (s_0+n)(\mu-\delta)^2}{2\sigma^2}\right) d\sigma^2$$

- Letting $A = 2\beta + (s_0 + n)(\mu - \delta)^2$ and making the substitution $z = \frac{A}{2\sigma^2}$, so $\sigma^2 = \frac{A}{2z}$ and $d\sigma^2 = -\frac{A}{2z^2}\, dz$, we transform the integral for further simplification.

A Model for Normal Data (Mean and Variance Both Unknown): Marginal Posterior for μ

- After substitution, the marginal posterior for μ becomes:

$$p(\mu|\mathbf{y}) \propto \int_0^\infty \left(\frac{A}{2z}\right)^{-\alpha-\frac{n}{2}-\frac{3}{2}} \frac{A}{2z^2} e^{-z}\, dz$$

- Simplifying: $p(\mu|\mathbf{y}) \propto \int_0^\infty \left(\frac{A}{2z}\right)^{-\alpha-\frac{n}{2}-\frac{1}{2}} \frac{1}{z} e^{-z}\, dz$
- Factor out terms that don't depend on z:

$$\propto A^{-\alpha-\frac{n}{2}-\frac{1}{2}} \int_0^\infty z^{\alpha+\frac{n}{2}+\frac{1}{2}-1} e^{-z}\, dz$$

- The integrand is the kernel of a **Gamma distribution**, so the integral is a constant.
- Thus, the marginal posterior for μ is proportional to:

$$p(\mu|\mathbf{y}) \propto A^{-\alpha-\frac{n}{2}-\frac{1}{2}} = \left(2\beta + (s_0 + n)(\mu - \delta)^2\right)^{-\alpha-\frac{n}{2}-\frac{1}{2}}$$

A Model for Normal Data (Mean and Variance Both Unknown): Marginal Posterior for μ

- The marginal posterior for μ simplifies to:

$$p(\mu|\mathbf{y}) \propto \left(2\beta + (s_0 + n)(\mu - \delta)^2\right)^{-\frac{2\alpha+n+1}{2}}$$

- This can be rewritten as:

$$p(\mu|\mathbf{y}) \propto \left[1 + \frac{(s_0 + n)(\mu - \delta)^2}{2\beta}\right]^{-\frac{2\alpha+n+1}{2}}$$

- This represents a (scaled) **noncentral t-distribution kernel** with:
  - **Noncentrality parameter**: δ
  - **Degrees of freedom**: $n + 2\alpha$

Bayesian Analysis for Normal Data Model

Case	Prior	Posterior
σ^2 known, μ unknown	$\mu \sim N(\delta, \tau^2)$	$\mu \vert \mathbf{y} \sim N\left(\frac{\delta/\tau^2 + n\bar{y}/\sigma^2}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2}\right)$
μ known, σ^2 unknown	$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\sigma^2 \vert \mathbf{y} \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(y_i - \mu)^2\right)$
μ and σ^2 both unknown	$\mu \vert \sigma^2 \sim N(\delta, \sigma^2/s_0)$ $\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\sigma^2 \vert \mathbf{y} \sim \text{IG}\left(\alpha + \frac{n-1}{2}, \beta + \frac{1}{2}\sum(y_i - \bar{y})^2\right)$ $\mu \vert \sigma^2, \mathbf{y} \sim N\left(\frac{n\bar{y} + s_0\delta}{n + s_0}, \frac{\sigma^2}{n + s_0}\right)$ $\mu \vert \mathbf{y} \sim t_{nc}(\delta, n + 2\alpha)$

Table 1: Conjugacy Table for Normal Data Model

Example 1: Midge Data

- **Example 1:** $Y_1, \ldots, Y_9$ represent a random sample of midge wing lengths (in mm), assumed to be iid $N(\mu, \sigma^2)$.
- **Example 1(a):** If $\sigma^2 = 0.01$ is known, we aim to make inference about μ. (See `R` example)
- The **Bayesian point estimate** for the population mean midge wing length is the **posterior mean**: 1.806 mm.
- A **95% credible interval** for μ is $(1.741, 1.871)$, meaning there is a 95% posterior probability that the population mean midge wing length lies between 1.741 and 1.871 mm.

Example 1: Midge Data (Part 1b)

- **Example 1(b):** Make inference about both μ and σ^2, both unknown. (See `R` example)
- This involves selecting the hyperparameters α and β for the inverse gamma prior on σ^2.
- The **95% credible interval** for σ^2 is $(0.012, 0.028)$, with a **posterior median** of 0.0188.
- To approximate the **posterior distribution** for μ:
  - Randomly generate values from the posterior distribution of σ^2.
  - For each generated σ^2, generate values from the posterior distribution of $\mu|\sigma^2$.
- The **95% credible interval** for μ is $(1.727, 1.90)$, with a **posterior median** of 1.81 mm.

Example 2: Brain Data

- The textbook provides an example of Bayesian inference on the **mean hippocampal volume** in a population of college football players with a history of concussions.
- **Example 2:** $Y_1, \ldots, Y_{25}$ represent a random sample of hippocampal volumes (in cm^3) for these football players. Assume the Y_i's are iid $N(\mu, \sigma^2)$.
- **Example 2(a):** If $\sigma = 0.5$ (i.e., $\sigma^2 = 0.25$), we aim to make inference about μ. We assume a prior distribution for $\mu \sim N(6.5, 0.4^2)$.
- The **posterior mean** is 5.78 cm^3.
- With **posterior probability 0.95**, the mean hippocampal volume of the population of concussed players is between 5.59 and 5.97 cm^3.

Prior Elicitation for Normal Priors (A Brief Yet Fascinating Interlude)

- A challenge is putting "expert opinion" into a form where it can be used as a prior distribution.

Strategies:

- Requesting guesses for several quantiles (maybe $\{0.1, 0.25, 0.5, 0.75, 0.9\}$?) from a few experts.
- For a **normal prior**, note that a quantile $q(\alpha)$ is related to the z-value $\Phi^{-1}(\alpha)$ by:

$$q(\alpha) = \text{mean} + \Phi^{-1}(\alpha) \times (\text{std. dev.})$$

- Via regression on the provided $\left[q(\alpha), \Phi^{-1}(\alpha)\right]$ values, we can get estimates for the mean and standard deviation of the normal prior. See the relevant `R` code on Canvas.

Extension to Location-Scale Family Priors

- Many distributions can be expressed as part of the **location-scale family**, where a random variable X is modeled as:

$$X = \theta + \tau Z$$

- θ is the location parameter, τ is the scale parameter, and Z is a standard distribution.
- Common examples include Normal, Cauchy, Laplace (Double-Exponential), Exponential, and *t*-distributions.

Eliciting Priors: A Quantile-Based Approach

- Similar to normal priors, we can ask experts for quantile estimates at several probabilities α, such as $\{0.1, 0.25, 0.5, 0.75, 0.9\}$.
- For location-scale distributions, the quantiles follow the general form:

  $$q(\alpha) = \theta + \tau \times F^{-1}(\alpha)$$

  where $F^{-1}(\alpha)$ is the quantile function of the standard version of the distribution (e.g., standard normal, Cauchy, etc.).
- Use the elicited quantiles $q(\alpha)$ and known quantile functions $F^{-1}(\alpha)$ to perform regression on pairs $\left[q(\alpha), F^{-1}(\alpha)\right]$.
- This allows estimation of the location parameter θ and scale parameter τ.

Prior Elicitation for Normal Priors (Cont.)

- Another strategy asks the expert to provide a “predictive modal value” (most “likely” value) for the parameter.
- Then a rough 67% interval is requested from the expert.
- With a **normal prior**, the length of this interval is twice the prior standard deviation and the modal value is the mean.
- For a prior on a Bernoulli probability, the “most likely” probability of success is often “clear”.