Bayesian Models

Yuki Shiraito

POLSCI 798 Advanced Topics in Quantitative Methodology
University of Michigan

Winter 2019

Example: Predicting Election Outcomes

- The 538 problem:
  - Predicts the winner in each state/district
  - Uses aggregate data from polls
    - Sample size
    - Number of intended votes for each candidate

- Model:
  - For each district i,
  - p_i: Dem's actual vote share
  - X_{ij}: Intended votes for Dem in poll j
  - n_{ij}: Sample size of poll j
  - Generative process: $X_{ij} \overset{\text{indep.}}{\sim} \text{Binom}(n_{ij}, p_i)$ for $j = 1, \ldots, J$

- Goal of Bayesian inference:
  - Posterior distribution of p_i given $X_{i1}, \ldots, X_{iJ}$

Bayes' Theorem

- Bayes' theorem:
  - Parameter θ with prior $p(\theta)$
  - Data $\mathbf{X}$ with likelihood $p(\mathbf{X} \mid \theta)$
  - Likelihood = joint distribution of data
  - Posterior of θ given $\mathbf{X}$:

$$p(\theta \mid \mathbf{X}) = \frac{\overbrace{p(\theta)p(\mathbf{X} \mid \theta)}^{=p(\theta,\mathbf{X})}}{\underbrace{\int p(\theta)p(\mathbf{X} \mid \theta)d\theta}_{=p(\mathbf{X})}}$$

  - Conditional = joint / marginal
- $p(\mathbf{X})$ is constant for any θ:
  - Posterior is *proportional* to prior $\times$ likelihood

$$p(\theta \mid \mathbf{X}) \propto p(\theta)p(\mathbf{X} \mid \theta)$$

  - Worry about $p(\mathbf{X})$ only if necessary

Likelihood: Binomial Data

- What is the likelihood?
  - Joint distribution of data given parameters
  - Density if continuous; probability if discrete
- Binomial distribution $X_{ij} \sim \text{Binom}(n_{ij}, p_i)$
  - Probability function of X_{ij} evaluated at x_{ij}:
$$p\left(X_{ij} = x_{ij} \mid n_{ij}, p_i\right) = \binom{n_{ij}}{x_{ij}} p_i^{x_{ij}} (1 - p_i)^{n_{ij} - x_{ij}}$$
  - Probability that random variable X_{ij} takes value x_{ij}
  - For simplicity, will write $p\left(X_{ij} \mid n_{ij}, p_i\right)$
- Independence $\rightsquigarrow$ factorization of likelihood
  - Joint probability of independent Binomials $X_{i1}, \ldots, X_{iJ}$:
$$p\left(X_{i1}, \ldots, X_{iJ} \mid n_{i1}, \ldots, n_{iJ}, p_i\right) = \prod_{j=1}^{J} p\left(X_{ij} \mid n_{ij}, p_i\right)$$
$$= \left\{\prod_{j=1}^{J} \binom{n_{ij}}{X_{ij}}\right\} p_i^{\sum_{j=1}^{J} X_{ij}} (1 - p_i)^{\sum_{j=1}^{J} n_{ij} - \sum_{j=1}^{J} X_{ij}}$$

Prior: Beta-Binomial Model

- What is the prior?
  - Joint distribution of parameters
  - Chosen by you
  - Means to include information other than data

- The Beta distribution
  - Continuous distribution over interval $[0, 1]$
  - Probability density function of p_i:

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

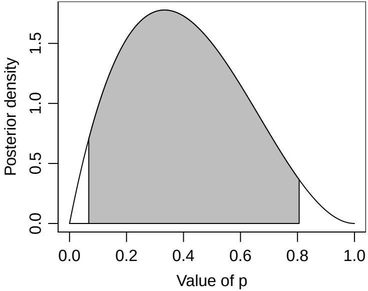
  - Commonly used for probability parameters
  - Uniform distribution if $\alpha = \beta = 1$
  - $B(\alpha, \beta) \equiv \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx$: Normalizing constant
- Beta-Binomial model
  1. Prior on actual vote share: $p_i \sim \text{Beta}(\alpha, \beta)$
  2. Model for data: $X_{ij} \overset{\text{indep.}}{\sim} \text{Binom}(n_{ij}, p_i)$

Posterior Distribution

- Goal of Bayesian inference: Posterior distribution of p_i
- Bayes' Theorem:
  1. Proportional to prior $\times$ likelihood
     $$p(p_i \mid X_{i1}, \ldots, X_{iJ})$$
     $$\propto \frac{1}{B(\alpha, \beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1} \left\{ \prod_{j=1}^{J} \binom{n_{ij}}{X_{ij}} \right\} p_i^{\sum_{j=1}^{J} X_{ij}} (1-p_i)^{\sum_{j=1}^{J} n_{ij} - \sum_{j=1}^{J} X_{ij}}$$
  2. Proportional to terms with p_i only
     $$\propto p_i^{\alpha-1}(1-p_i)^{\beta-1} p_i^{\sum_{j=1}^{J} X_{ij}} (1-p_i)^{\sum_{j=1}^{J} n_{ij} - \sum_{j=1}^{J} X_{ij}}$$
  3. Beta kernel by rearranging
     $$= p_i^{\alpha + \sum_{j=1}^{J} X_{ij} - 1} (1-p_i)^{\beta + \sum_{j=1}^{J} n_{ij} - \sum_{j=1}^{J} X_{ij} - 1}$$
  4. Posterior is a Beta distribution
     $$p_i \mid X_{i1} \ldots, X_{iJ} \sim \text{Beta}\left(\alpha + \sum_{j=1}^{J} X_{ij}, \beta + \sum_{j=1}^{J} n_{ij} - \sum_{j=1}^{J} X_{ij} \right)$$
- Conjugacy: Beta prior leads to Beta posterior

Summary of Posterior Inference

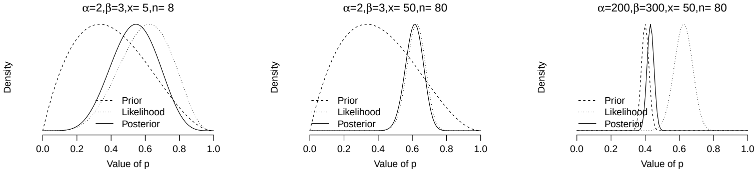
- Common summaries: Mean, median, mode, variance, etc.
- Credible intervals
  - Interval estimation in Bayesian context
  - $(1-\alpha) \times 100\%$ (central) credible interval:
    - Between $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior



    - "Center" of the posterior distribution
  - More intuitive interpretation than confidence intervals:
    - Probability that parameter is in the interval is $1-\alpha$
    - Does not rely on repeated samples

Prior as Additional Data

- Bayesian inference: Update the prior to the posterior



  - Bayesian update $\rightsquigarrow$ compromise between data and prior
  - More data $\rightsquigarrow$ data dominate prior
  - Too strong prior $\rightsquigarrow$ little update
- Prior parameters: Number of "additional data points"
  - Posterior mean in the Beta-Binomial model:

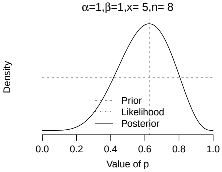
$$\mathbb{E}[p_i \mid X_{i1}, \ldots, X_{iJ}] = \frac{\alpha + \sum_{j=1}^{J} X_{ij}}{\alpha + \beta + \sum_{j=1}^{n} n_{ij}}$$

  - α: Pseudo number of successes in prior
  - β: Pseudo number of failures in prior
  - Holds for many other models with conjugate prior

Uniform Prior and the Maximum Likelihood Estimator

- The Uniform prior distribution
  - "No information" in the prior
  - Posterior proportional to likelihood only:

$$p(p_i \mid X_{i1}, \ldots, X_{iJ}) \propto 1 \times \prod_{j=1}^{J} p(X_{ij} \mid p_i)$$



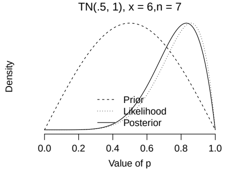
- Posterior mode = MLE

$$\widehat{p}_{\mathrm{MLE}} = \frac{\sum_{j=1}^{J} X_{ij}}{\sum_{j=1}^{J} n_{ij}}, \quad \widehat{p}_{\mathrm{mode}} = \frac{\overbrace{\alpha}^{=1} + \sum_{j=1}^{J} X_{ij} - 1}{\underbrace{\alpha + \beta}_{=1+1} + \sum_{j=1}^{J} n_{ij} - 2}$$

- Why not always use the Uniform prior?
  - Unbounded parameter space
  - Non-conjugacy $\rightsquigarrow$ computational issues

Inference with Non-conjugate Prior

- Conjugate prior $\rightsquigarrow$ known family of posterior
- What if prior is not conjugate?
- Non-conjugate prior: Truncated Gaussian
$$p_i \sim \mathcal{TN}_{[0,1]}(.5, .25)$$
- Posterior:
$$p(p_i \mid X_{i1}, \ldots, X_{iJ})$$
$$\propto e^{-\frac{(p_i-.5)^2}{2\times .25}} p_i^{\sum_{j=1}^{J} X_{ij}} (1-p_i)^{\sum_{j=1}^{J}(n_{ij}-X_{ij})}$$
- Explicit form of the posterior density available $\rightsquigarrow$ evaluation of the density
- Problems:
  1. Intractable posterior normalizing constant
  2. Intractable posterior expectation
- Need for Monte Carlo approximation



Monte Carlo Methods: Overview

- Posterior $p(\theta \mid \mathbf{X})$ with a non-conjugate prior:
  - $q(\theta \mid \mathbf{X})$ is available, where $p(\theta \mid \mathbf{X}) \propto q(\theta \mid \mathbf{X})$
  - $q(\theta \mid \mathbf{X})$ is *unnormalized* because $\int q(\theta \mid \mathbf{X}) d\theta \neq 1$
  - $\int q(\theta \mid \mathbf{X}) d\theta$ is intractable $\Rightarrow$ hard to summarize $p(\theta \mid \mathbf{X})$
  - Posterior mean, variance, quantiles...
- Monte Carlo, or simulation-based inference:
  - Simulate a random sample from $p(\theta \mid \mathbf{X})$:

$$\theta^{(s)} \sim p(\theta \mid \mathbf{X}), \quad s = 1, \ldots, S$$

  - Approximate integrals by sums:

$$\mathbb{E}[\theta \mid \mathbf{X}] \approx \overline{\theta}_S \equiv \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}, \quad \mathbb{V}(\theta \mid \mathbf{X}) \approx \frac{1}{S} \sum_{s=1}^{S} \left(\theta^{(s)} - \overline{\theta}_S\right)^2$$

- How do we simulate from $p(\theta \mid \mathbf{X})$ if only $q(\theta \mid \mathbf{X})$ is available?

Generic Implementation: Stan

- Stan: "Black box" implementation
  - Can be used for virtually any applied Bayesian model
  - Only need objects and the model defined
- Interface with **R**: RStan
- Stan code for the Truncated Gaussian prior example:

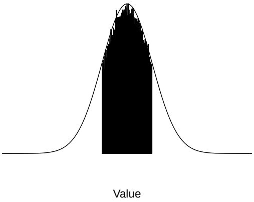
```
data {
    int<lower=0> x;      // trump count
    int<lower=0> n;     // sample size
}
parameters {
    real<lower=0,upper=1> p;    // posterior parameter
}
model {
    p ~ normal(.5, .5)T[0,1]; // truncated normal prior
    x ~ binomial(n,p); // likelihood function
}
```

- Output: Sample from the posterior distribution of parameters
- Easy to use, but sometimes hard to debug or improve speed

Rejection Sampling: Truncated Gaussian Samples

- Sampling from $\mathcal{TN}_{[0,1]}(\mu, \sigma^2)$: For draw s,
  1. Draw a *proposal* $\theta_p \sim \mathcal{N}(\mu, \sigma^2)$
  2. Accept θ_p as $\theta^{(s)}$ if $0 \leq \theta_p \leq 1$; return to Step 1 otherwise
- **R** code:

```
samplesize <- 100000L
posterior.draws <- numeric(samplesize)
for (s in 1:samplesize) {
    rejected <- TRUE
    while(rejected) {
        proposal <- rnorm(1, mean = .5, sd = .5)
        if (proposal >= 0 & proposal <= 1) {
            rejected <- FALSE
            posterior.draws[s] <- proposal
        }
    }
}
```



- Propose–accept/reject: Common in many Monte Carlo methods
- Acceptance rule:
  - Designed so that accepted draws follow the target distribution
  - Rejecting too many proposals $\rightsquigarrow$ inefficiency

Rejection Sampling: Theory

- Requirements for rejection sampling:
  - We can simulate random draws from proposal density $g(\theta)$
  - The importance ratio is bounded by a known constant:
  $$0 \leq \frac{q(\theta \mid \mathbf{X})}{g(\theta)} \leq M \Leftrightarrow 0 \leq \frac{q(\theta \mid \mathbf{X})}{Mg(\theta)} \leq 1$$
- Two steps for each draw:
  1. Draw a proposal: $\theta_p \sim g(\theta)$
  2. Accept θ_p with probability $q(\theta_p \mid \mathbf{X})/Mg(\theta_p)$
  3. If accepted, go to next draw; if rejected, return to 1
- $\mathcal{TN}(\mu, \sigma^2)$ example:
  - g is $\mathcal{N}(\mu, \sigma^2)$
  - $q(\theta_p \mid \mathbf{X})/Mg(\theta_p) = 1$ if $\theta_p \in [0, 1]$ and 0 otherwise
- Justification for the algorithm:
  - Binary acceptance indicator Z: $p(Z = 1 \mid \theta_p) = q(\theta_p \mid \mathbf{X})/Mg(\theta_p)$
  - Density of accepted draws:
  $$p(\theta_p \mid Z = 1) = \frac{g(\theta_p)p(Z = 1 \mid \theta_p)}{\int g(\theta)p(Z = 1 \mid \theta)d\theta} = \frac{q(\theta_p \mid \mathbf{X})}{\int q(\theta \mid \mathbf{X})d\theta} = p(\theta_p \mid \mathbf{X})$$

Importance Resampling

- Caveats of rejection sampling:
  - May be slow/inefficient
  - May be hard to find good g and M

- Importance resampling (a.k.a. SIR):
  1. Simulate R s.t. $R > S$ draws, $\{\theta_p^{(1)}, \ldots, \theta_p^{(R)}\}$, from $g(\theta)$
  2. Resample S draws from the R draws above
     - Probability of resampling proportional to $q\left(\theta_p^{(r)} \mid \mathbf{X}\right)/g\left(\theta_p^{(r)}\right)$
     - Resampling without replacement generally recommended
- No rejections nor need to find M
- Bad proposal $\rightsquigarrow$ approximation worse than rejection sampling

- Posterior with a conjugate prior can be used as a proposal
  - $\text{Beta}(7, 8)$ as a proposal for the posterior with the $\mathcal{TN}$ prior

Example: Randomized Experiment

- Health savings experiment
  - Dupas, Pascaline, and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review*, 103 (4): 1138-71.
  - Randomized field experiment in Kenya
  - Outcome: Amount spent on preventive health products
  - Treatment: Simple locked box similar to a piggy bank
  - Does the treatment increase health investment?
- Causal inference with randomized experiment
  - Potential outcomes: $(Y_i(0), Y_i(1))$
  - Population average treatment effect (PATE): $\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$
  - Ignorability: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i$
- Normal-Normal model for Bayesian inference on PATE
  - Data: $Y_i(0) \sim \mathcal{N}(\mu_0, \sigma_0^2),\ Y_i(1) \sim \mathcal{N}(\mu_0 + \tau, \sigma_1^2),\ T_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$
  - Conjugate prior: $\mu_0 \sim \mathcal{N}(v, \lambda^2),\ \tau \sim \mathcal{N}(0, \kappa^2)$
  - For now, assume σ_0 and σ_1 are known

Bayesian Update of Gaussian Mean

- Given unit i in the control group, $Y_i \equiv Y_i(T_i) = Y_i(0)$
  1. Posterior proportional to the prior times the likelihood:
  $$p\left(\mu_0, \tau \,\middle|\, Y_i, T_i = 0, \sigma_0^2\right) \propto \underbrace{e^{-\frac{(\mu_0 - \nu)^2}{2\lambda^2}}}_{\text{prior on } \mu_0} \underbrace{e^{-\frac{\tau^2}{2\kappa^2}}}_{\text{prior on } \tau} \underbrace{e^{-\frac{(Y_i - \mu_0)^2}{2\sigma_0^2}}}_{\text{likelihood of } Y_i} \underbrace{(1-p)}_{\text{likelihood of } T_i = 0}$$
  2. Factorization $\rightsquigarrow$ μ_0 is independent of τ *a posteriori*:
  $$\propto \underbrace{e^{-\frac{(\mu_0 - \nu)^2}{2\lambda^2}} e^{-\frac{(Y_i - \mu_0)^2}{2\sigma_0^2}}}_{\propto p\left(\mu_0 \middle| Y_i, T_i = 0, \sigma_0^2\right)} \quad \underbrace{e^{-\frac{\tau^2}{2\kappa^2}}}_{\propto p\left(\tau \middle| Y_i, T_i = 0, \sigma_0^2\right)}$$
  3. No treated obs $\rightsquigarrow$ no updates on τ: $p\left(\tau \middle| Y_i, T_i = 0, \sigma_0^2\right) = p(\tau)$
  4. Posterior distribution of mean control outcome:
  $$p\left(\mu_0 \,\middle|\, Y_i, T_i = 0, \sigma_0^2\right) \propto e^{-\left(\frac{(\mu_0 - \nu)^2}{2\lambda^2} + \frac{(Y_i - \mu_0)^2}{2\sigma_0^2}\right)} \propto e^{-\frac{(\mu_0 - \widehat{\nu})^2}{2\widehat{\lambda}^2}}$$
  $$\therefore \mu_0 \mid Y_i, T_i = 0, \sigma_0 \sim \mathcal{N}\left(\widehat{\nu}, \widehat{\lambda}^2\right), \quad \widehat{\nu} = \frac{\frac{1}{\lambda^2}\nu + \frac{1}{\sigma_0^2} Y_i}{\frac{1}{\lambda^2} + \frac{1}{\sigma_0^2}}, \quad \widehat{\lambda}^2 = \frac{1}{\frac{1}{\lambda^2} + \frac{1}{\sigma_0^2}}$$
- $\widehat{\nu}$: Weighted average of the prior mean and the data
- $1/\widehat{\lambda}^2$: Sum of the inverse variances

Shrinkage and the Bias-Variance Tradeoff

- Shrinkage
  - Y_i is "shrunk" toward v:

$$\widehat{v} = Y_i - (Y_i - v)\,\frac{\sigma_0^2}{\lambda^2 + \sigma_0^2}$$

  - Larger (smaller) variance of data $\rightsquigarrow$ more (less) shrinkage
  - Larger (smaller) prior variance $\rightsquigarrow$ less (more) shrinkage
  - $\widehat{v}$ is *biased* from the frequentist perspective:

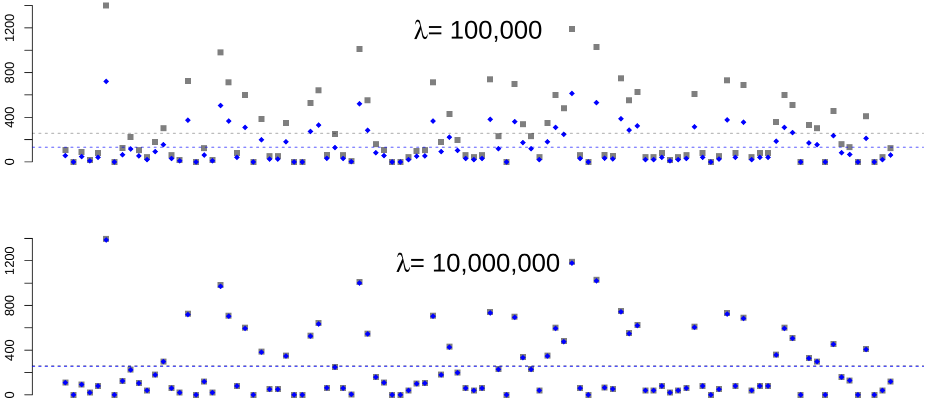
$$\mathbb{E}_Y\left[\widehat{v} \mid \mu_0\right] = \mu_0 - (\mu_0 - v)\,\frac{\sigma_0^2}{\lambda^2 + \sigma_0^2}$$

- Prior distributions introduce bias. Why do we use them?
- Bias-variance tradeoff
  - Y_i is unbiased for μ_0, but its sampling variance is σ_0^2
  - $\widehat{v}$ is biased for μ_0, but its sampling variance is:

$$\mathbb{V}_Y\left(\widehat{v} \mid \mu_0\right) = \frac{\lambda^4}{\left(\lambda^2 + \sigma_0^2\right)^2}\sigma_0^2 < \sigma_0^2$$

  - Unbiasedness $\hookleftarrow\!\hookrightarrow$ variance reduction
  - Prior as an example of regularization

Shrinkage and Uninformative Prior Distributions



- Posterior mean (blue) given each observation (gray)
  - Shrinkage with $\lambda = 100,000$
  - No shrinkage with $\lambda = 10,000,000$
- $\lambda \uparrow \rightsquigarrow$ less informative prior $\rightsquigarrow$ less shrinkage
- Uninformative (improper) prior: $\lambda \to \infty \Leftrightarrow p(\mu_0) \propto 1$

Sufficient Statistics

- Given all units in the control group, $Y_i = Y_i(0), i = 1, \ldots, N_0$
  1. Posterior with the likelihood of $(Y_1, \ldots, Y_{N_0})$:

$$p\left(\mu_0 \,\middle|\, Y_1, \ldots, Y_{N_0}, \sigma_0^2\right) \propto e^{-\frac{(\mu_0 - \nu)^2}{2\lambda^2}} \underbrace{\prod_{i=1}^{N_0} e^{-\frac{(Y_i - \mu_0)^2}{2\sigma_0^2}}}_{\text{likelihood of } Y_1, \ldots, Y_{N_0}}$$

$$= e^{-\frac{1}{2}\left(\frac{(\mu_0 - \nu)^2}{\lambda^2} + \frac{1}{\sigma_0^2}\sum_{i=1}^{N_0}(Y_i - \mu_0)^2\right)} \Rightarrow \mu_0 \,\middle|\, Y_1, \ldots, Y_{N_0}, \sigma_0^2 \sim \mathcal{N}\left(\hat{\nu}_C, \hat{\lambda}_C^2\right)$$

  2. Posterior mean and variance (will be derived in BK's section):

$$\hat{\nu}_C = \frac{\frac{1}{\lambda^2}\nu + \frac{1}{\sigma_0^2/N_0}\overline{Y}_C}{\frac{1}{\lambda^2} + \frac{1}{\sigma_0^2/N_0}}, \quad \hat{\lambda}_C^2 = \frac{1}{\frac{1}{\lambda^2} + \frac{1}{\sigma_0^2/N_0}}, \quad \text{where } \overline{Y}_C \equiv \frac{1}{N_0}\sum_{i=1}^{N_0} Y_i$$

     - $\hat{\nu}_C$: Weighted average of the prior mean and the sample mean
     - $N_0 \uparrow \rightsquigarrow$ weight on $\overline{Y}_C \uparrow$: Sample mean dominates given large data
- $\overline{Y}_C$ is called a sufficient statistic for μ_0
  - Identical to observing one data point of $\overline{Y}_C \sim \mathcal{N}(\mu_0, \sigma_0^2/N_0)$

    $\Rightarrow$ Suffices to know $\overline{Y}_C$ for the posterior

Bayesian Inference on the Treatment Effect

- Given all units in the control and the treatment groups
  1. Control group: $Y_i = Y_i(0), i = 1, \ldots, N_0$
  2. Treatment group: $Y_i = Y_i(1), i = N_0 + 1, \ldots, N_0 + N_1$
- Joint posterior distribution of μ_0 and τ
  1. Posterior proportional to the prior and the likelihood:

$$p\left(\mu_0, \tau \,\middle|\, \mathbf{Y}, \sigma_0^2, \sigma_1^2\right) \propto e^{-\frac{(\mu_0 - \nu)^2}{2\lambda^2}} e^{-\frac{\tau^2}{2\kappa^2}} \times \underbrace{e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{N_0}\left(Y_i - \mu_0\right)^2}}_{\text{control group}} \underbrace{e^{-\frac{1}{2\sigma_1^2}\sum_{i=N_0+1}^{N_0+N_1}\left\{Y_i - \left(\mu_0 + \tau\right)\right\}^2}}_{\text{treatment group}}$$

  2. Joint posterior distribution: Bivariate Gaussian

$$\left(\begin{array}{c} \mu_0 \\ \tau \end{array}\right) \Bigg| \mathbf{Y}, \sigma_0^2, \sigma_1^2 \ \sim \mathcal{N}\left(\widehat{\Lambda}\,\widehat{\upsilon}, \widehat{\Lambda}\right), \quad \text{where}$$

$$\widehat{\upsilon} \equiv \left(\begin{array}{c} \frac{\nu}{\lambda^2} + \frac{N_0}{\sigma_0^2}\overline{Y}_C + \frac{N_1}{\sigma_1^2}\overline{Y}_T \\ \frac{N_1}{\sigma_1^2}\overline{Y}_T \end{array}\right), \quad \widehat{\Lambda} \equiv \left(\begin{array}{cc} \frac{1}{\lambda^2} + \frac{N_0}{\sigma_0^2} + \frac{N_1}{\sigma_1^2} & \frac{N_1}{\sigma_1^2} \\ \frac{N_1}{\sigma_1^2} & \frac{N_1}{\sigma_1^2} + \frac{1}{\kappa^2} \end{array}\right)^{-1}$$

- Sufficient statistics: $\overline{Y}_C$ and $\overline{Y}_T \equiv \frac{1}{N_1}\sum_{i=N_0+1}^{N_0+N_1} Y_i$
- Derivation too tedious $\rightsquigarrow$ Bayeisian linear regression

Conditional Posterior Distributions

- Goal: Joint posterior of all parameters
- Sometimes too tedious even with a conjugate prior
- Conditional posterior distributions:
  - Posterior of some parameters conditional on the others
  - Easier to derive; the other parameters as known constants

- Conditional posterior distributions of τ and μ_0:
  1. Conditional posterior of τ given μ_0

$$\tau \,\Big|\, \mathbf{Y}, \sigma_0^2, \sigma_1^2, \mu_0 \sim \mathcal{N}\left(\frac{\frac{(\bar{Y}_T - \mu_0)}{\sigma_1^2/N_1}}{\frac{1}{\kappa^2} + \frac{1}{\sigma_1^2/N_1}}, \left(\frac{1}{\frac{1}{\kappa^2} + \frac{1}{\sigma_1^2/N_1}} \right)^2 \right)$$

  2. Conditional posterior of μ_0 given τ

$$\mu_0 \,\Big|\, \mathbf{Y}, \sigma_0^2, \sigma_1^2, \tau \sim \mathcal{N}\left(\frac{\frac{v}{\lambda^2} + \frac{(\bar{Y}_T - \tau)}{\sigma_1^2/N_1} + \frac{\bar{Y}_C}{\sigma_0^2/N_0}}{\frac{1}{\lambda^2} + \frac{1}{\sigma_1^2/N_1} + \frac{1}{\sigma_0^2/N_0}}, \left(\frac{1}{\frac{1}{\lambda^2} + \frac{1}{\sigma_1^2/N_1} + \frac{1}{\sigma_0^2/N_0}} \right)^2 \right)$$

- From conditionals to joint: Markov chain Monte Carlo algorithms

Summary

- Fundamentals of Bayesian inference
  - Goal: Joint posterior distribution of all unknown parameters
  - Posterior: Prior times the likelihood
- Implementation of Bayesian inference
  - Posterior summaries: Mean, variance, credible intervals
  - Conjugate prior $\rightsquigarrow$ known family of posterior
  - Posterior hard to derive $\rightsquigarrow$ Monte Carlo methods
- Interpretation of Bayesian inference
  - Compromise between prior and data
  - Shrinkage toward prior: Variance reduction

- Readings for review: **BDA3** Chs. 1-2, and 10