ECON5120 Bayesian Data Analysis

Week 1: Introduction and examples

Santiago Montoya-Blandón

January 14, 2026

Outline

Bayesian vs Frequentist

- Deep fundamental question: *how to assign probabilities to events?*

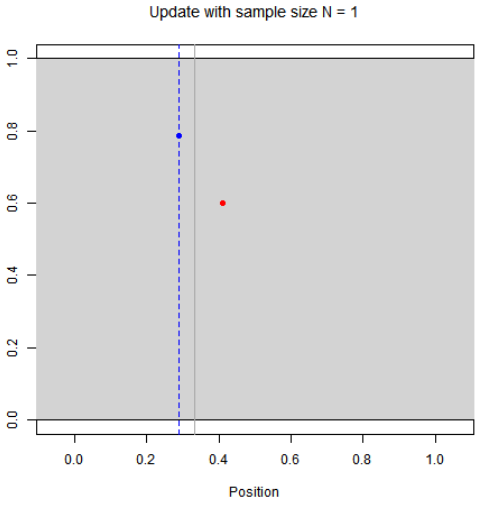
Bayesian vs Frequentist

- Deep fundamental question: *how to assign probabilities to events?*
- Bayesian and Frequentist approaches are two such possible answers
- Frequentist approach (current dominant way of thinking) $\Rightarrow$ Probabilities are relative frequencies from repeated experiments
- Bayesian paradigm $\Rightarrow$ Any *uncertain* event can be assigned a probability: start from a prior belief and update it *optimally* to any upcoming information

Bayesian vs Frequentist

- Deep fundamental question: *how to assign probabilities to events?*
- Bayesian and Frequentist approaches are two such possible answers
- Frequentist approach (current dominant way of thinking) $\Rightarrow$ Probabilities are relative frequencies from repeated experiments
- Bayesian paradigm $\Rightarrow$ Any *uncertain* event can be assigned a probability: start from a prior belief and update it *optimally* to any upcoming information
- Both approaches are mathematically rigorous descriptions of probabilities (satisfy Kolmogorov's probability axioms)
- There are deep methodological and philosophical debates between Bayesian and Frequentist outside the scope of the course
- We will take a pragmatic approach: exploring the statistical benefits of both approaches without addressing these concerns

Example: Bayes' Table



Bayesian vs Frequentist

- An economic model is described by parameters θ and is fit to data y
- Simplified way to understand difference between approaches:

Frequentists: θ is unknown yet fixed; y is random due to sampling variation

Bayesians: θ is uncertain and thus treated as random; y is taken as fixed

Bayesian vs Frequentist

- An economic model is described by parameters θ and is fit to data y
- Simplified way to understand difference between approaches:

Frequentists: θ is unknown yet fixed; y is random due to sampling variation

Bayesians: θ is uncertain and thus treated as random; y is taken as fixed

- Example: Returns to education. Data $y_i = wage_i$ and $x_i = educ_i$ for individuals $i = 1, \ldots, n$
- Economic model is a (normal) wage regression:

$$y_i = \theta_0 + \theta_1 x_i + u_i,\ u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \Rightarrow y_i \mid x_i, \theta_0, \theta_1, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$$

Bayesian vs Frequentist

- An economic model is described by parameters θ and is fit to data y
- Simplified way to understand difference between approaches:

Frequentists: θ is unknown yet fixed; y is random due to sampling variation
Bayesians: θ is uncertain and thus treated as random; y is taken as fixed

- Example: Returns to education. Data $y_i = wage_i$ and $x_i = educ_i$ for individuals $i = 1, \ldots, n$
- Economic model is a (normal) wage regression:

$$y_i = \theta_0 + \theta_1 x_i + u_i,\ u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \Rightarrow y_i \mid x_i, \theta_0, \theta_1, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$$

- Structural interpretation of parameters does not depend on Frequentist or Bayesian: θ_0 is average wage for those with no education. What about θ_1?
- Fixed or Random returns to education θ_0 and θ_1?

Bayesian Analysis

- All of Bayesian analysis relies on *Bayes' rule* (the *optimal* way of updating uncertainty given new information)

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

- The four components are the *likelihood* $p(y \mid \theta)$, *prior* $p(\theta)$, *marginal likelihood* $p(y)$ and *posterior* $p(\theta \mid y)$
- As $p(y)$ does not depend on θ, we can write

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta) \qquad (a \propto b \Leftrightarrow a = c \cdot b, c \in \mathbb{R})$$

- The posterior $p(\theta \mid y)$ summarizes everything we can know about θ given data y

Bayesian Analysis: Bayes' rule components

1. **Likelihood:** $p(y \mid \theta) \Rightarrow$ Model describing data y in terms of parameters θ
2. **Prior:** $p(\theta) \Rightarrow$ Knowledge about θ prior to seeing data y *encoded as a probability distribution*
3. **Posterior:** $p(\theta \mid y) \Rightarrow$ Main object of interest in Bayesian analysis, holds all information about parameters θ contained in data y
4. **Marginal likelihood:** $p(y) \Rightarrow$ marginal distribution of the data

- Using Bayes' theorem, we can also express

$$p(y) = \int p(y \mid \theta) p(\theta) d\theta$$

Bayesian Analysis: Summarizing the posterior

As the posterior holds all information of θ contained in data y, we will ultimately need to *summarize* the posterior

- Posterior probabilities: $P(a \leq \theta \leq b \mid y) = \int_a^b p(\theta \mid y) d\theta$
- Expectation: $\mathrm{E}[\theta \mid y] = \int \theta p(\theta \mid y) d\theta$
- Median: $\mathrm{Median}(\theta \mid y) = \left\{ m : \int_{-\infty}^{m} p(\theta \mid y) d\theta = \int_{m}^{\infty} p(\theta \mid y) d\theta = 1/2 \right\}$
- Mode: $\mathrm{Mode}(\theta \mid y) = \arg\max_\theta p(\theta \mid y)$
- Credible interval of probability p: endpoints θ_L and θ_U such that $P(\theta_L \leq \theta \leq \theta_U \mid y) = p$. The smallest such interval is known as the Highest Posterior Density (HPD) interval of size p

Bayesian Analysis: Prediction

- The objective is to predict a new value $\tilde{y}$ based on existing data y
- As before, all of our knowledge about $\tilde{y}$ will be contained in its posterior

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y) d\theta \\ &= \int p(\tilde{y} \mid \theta, y) p(\theta \mid y) d\theta \end{aligned}$$

- Note that this can be written as

$$p(\tilde{y} \mid y) = \mathrm{E}[p(\tilde{y} \mid \theta, y) \mid y]$$

where the expectation is with respect to the posterior $p(\theta \mid y)$

Bayesian Analysis: Model comparison

- Assume R possible models, M_i for $i = 1, \ldots, R$, each depending on its own set of parameters θ^i
- For each model, we can obtain

$$p(\theta^i \mid y, M_i) \propto p(y \mid \theta^i, M_i) p(\theta^i \mid M_i)$$

- Similarly, all knowledge about model M_i contained in the data is in

$$p(M_i \mid y) = \frac{p(y \mid M_i) p(M_i)}{p(y)}$$

- Note that $p(M_i)$ is a prior model probability, $p(y \mid M_i)$ is model i's marginal likelihood and $p(M_i \mid y)$ are *posterior model probabilities*
- We also have

$$p(y \mid M_i) = \int p(y \mid \theta^i, M_i) p(\theta^i \mid M_i) d\theta^i$$

Bayesian Analysis: Model comparison

- To compare models and to avoid computing $p(y)$, we usually use *posterior odd ratios* between models M_i and M_j

$$PO_{ij} = \frac{p(M_i \mid y)}{p(M_j \mid y)} = \frac{p(y \mid M_i)p(M_i)}{p(y \mid M_j)p(M_j)}$$

- These odds can be written as a product of what is known as the *Bayes factor* and the prior odds ratio between models i and j

$$BF_{ij} = \frac{p(y \mid M_i)}{p(y \mid M_j)}$$

- This Bayes factor can also be used to test hypotheses
- Example: M_1 imposes $\theta = \theta_0$ and model M_2 keeps θ unrestricted

Binomial likelihood

- Probability of event 1 in trial is θ
- Probability of event 2 in trial is $1 - \theta$
- Probability of several events in independent trials is e.g. $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\ldots$
- If there are n trials and we do not care about the order of events, then the probability that event 1 happens y times is

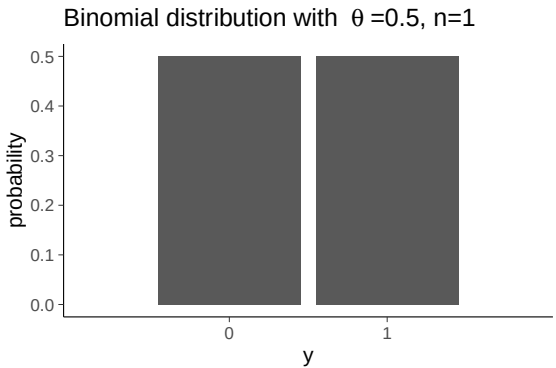
$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- This is precisely the likelihood when seen as a function of θ
- Frequentists would use $\hat{\theta} = \arg\max_\theta p(y \mid \theta, n) = \bar{y}$

Binomial likelihood

- Observation model (function of y, discrete)

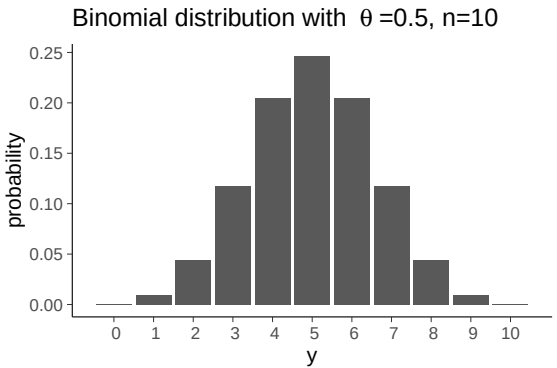
$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Binomial likelihood

- Observation model (function of y, discrete)

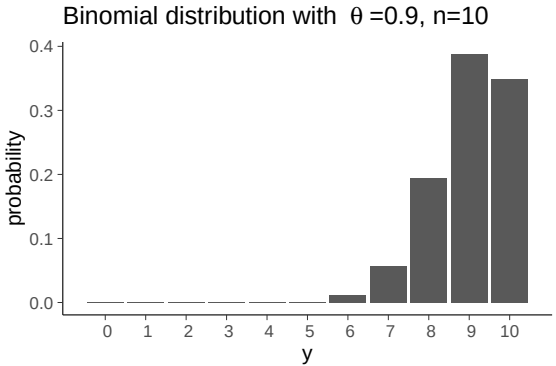
$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Binomial likelihood

- Observation model (function of y, discrete)

$$p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Bayesian analysis of Binomial model

- Posterior with Bayes rule (function of θ, continuous)

$$p(\theta \mid y, n, M) = \frac{p(y \mid \theta, n, M)p(\theta \mid n, M)}{p(y \mid n, M)}$$

where $p(y \mid n, M) = \int p(y \mid \theta, n, M)p(\theta \mid n, M)d\theta$

Bayesian analysis of Binomial model

- Posterior with Bayes rule (function of θ, continuous)

$$p(\theta \mid y, n, M) = \frac{p(y \mid \theta, n, M)p(\theta \mid n, M)}{p(y \mid n, M)}$$

  where $p(y \mid n, M) = \int p(y \mid \theta, n, M)p(\theta \mid n, M)d\theta$

- Start with uniform prior

$$p(\theta \mid n, M) = p(\theta \mid M) = 1, \text{ when } 0 \leq \theta \leq 1$$

Bayesian analysis of Binomial model

- Posterior with Bayes rule (function of θ, continuous)

$$p(\theta \mid y, n, M) = \frac{p(y \mid \theta, n, M)p(\theta \mid n, M)}{p(y \mid n, M)}$$

  where $p(y \mid n, M) = \int p(y \mid \theta, n, M)p(\theta \mid n, M)d\theta$

- Start with uniform prior

$$p(\theta \mid n, M) = p(\theta \mid M) = 1, \text{ when } 0 \leq \theta \leq 1$$

- Then

$$p(\theta \mid y, n, M) = \frac{p(y \mid \theta, n, M)}{p(y \mid n, M)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta}$$

$$= \frac{1}{Z}\theta^y(1-\theta)^{n-y}$$

Bayesian analysis of Binomial model

- Normalization term Z (constant given y)

$$Z = \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

- Normalisation term has *Beta* function form
  - when integrated over $(0, 1)$ the result can be represented with Gamma functions
  - with integers $\Gamma(n) = (n-1)!$
  - for large integers even this is challenging and usually $\log \Gamma(\cdot)$ is computed instead of $\Gamma(\cdot)$

Bayesian analysis of Binomial model

- Posterior is

$$p(\theta \mid y, n, M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y}(1-\theta)^{n-y},$$

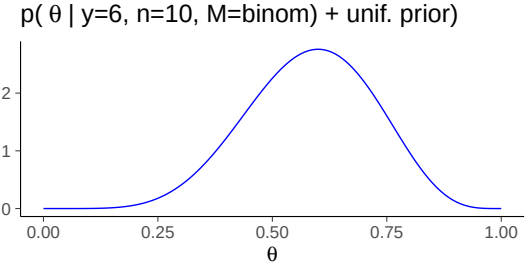
Bayesian analysis of Binomial model

- Posterior is

$$p(\theta \mid y, n, M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \theta^y (1-\theta)^{n-y},$$

which is called Beta distribution

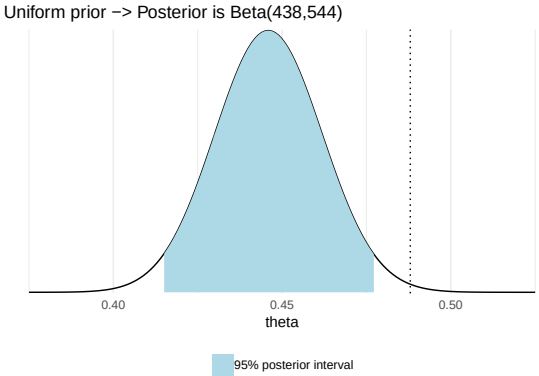
$$\theta \mid y, n \sim \text{Beta}(y+1, n-y+1), \quad \text{where} \quad \text{E}[\theta \mid y, n] = \frac{y+1}{n+2}$$



Placenta previa example

- Example given in BDA3, p. 37. Want to study probability of baby being born a girl when pregnant mother suffers from placenta previa
- 437 girls and 543 boys sampled, observed proportion of girls is 0.446
- In the general population, the probability of girl births is around 0.485
- The posterior mean with this data is 0.445. Are these quantities statistically different?

Placenta previa example



Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid \theta, y, n, M)$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid y, n, M) = \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid y, n, M) = \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta$$
$$= \int_0^1 \theta p(\theta \mid y, n, M) d\theta$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$
\begin{aligned}
p(\tilde{y} = 1 \mid y, n, M) &= \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta \\
&= \int_0^1 \theta p(\theta \mid y, n, M) d\theta = \mathrm{E}[\theta \mid y]
\end{aligned}
$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$\begin{aligned} p(\tilde{y} = 1 \mid y, n, M) &= \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta \\ &= \int_0^1 \theta p(\theta \mid y, n, M) d\theta = \mathrm{E}[\theta \mid y] \end{aligned}$$

- With uniform prior

$$\mathrm{E}[\theta \mid y] = \frac{y+1}{n+2}$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid y, n, M) = \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta$$
$$= \int_0^1 \theta p(\theta \mid y, n, M) d\theta = \mathrm{E}[\theta \mid y]$$

- With uniform prior

$$\mathrm{E}[\theta \mid y] = \frac{y+1}{n+2}$$

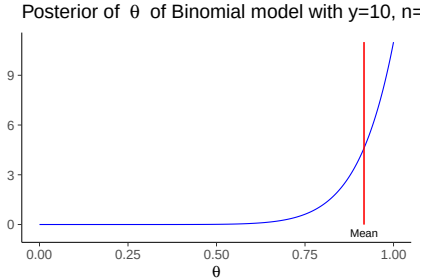
- Extreme cases

$$p(\tilde{y} = 1 \mid y = 0, n, M) = \frac{1}{n+2} \qquad p(\tilde{y} = 1 \mid y = n, n, M) = \frac{n+1}{n+2}$$

Benefits of integration: Comparison with MLE

Example: $n = 10, y = 10$

- With $y = 10$, the MLE will be $\hat{\theta} = 1$
- It would also imply $P(\tilde{y} = 1) = 1$
- Bayesian predictions take into account all the possible values of θ and weight the different possible predictions based on the posterior (regularization, stabilization)



Predictive distribution

- Posterior predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid y, n, M) = \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid y, n, M) d\theta$$

- Prior predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 \mid M) = \int_0^1 p(\tilde{y} = 1 \mid \theta, y, n, M) p(\theta \mid M) d\theta$$

- This will be useful when thinking about how to select or justify priors

Justification for uniform prior

- Uniform prior on a probability between 0 and 1: $p(\theta \mid M) = 1$
- Good option if we think all values of θ are equally likely
- Another justification: we want the prior predictive distribution to be uniform

$$p(\tilde{y} \mid M) = \frac{1}{2},$$

or

$$p(y \mid n, M) = \frac{1}{n+1}, \quad y = 0, \ldots, n$$

Priors

- Conjugate prior (BDA3 p. 35)
- Noninformative prior (BDA3 p. 51)
- Proper and improper prior (BDA3 p. 52)
- Weakly informative prior (BDA3 p. 55)
- Informative prior (BDA3 p. 55)
- Prior sensitivity (BDA3 p. 38)

Conjugate prior

- Prior and posterior have the same form
  - available for exponential family distributions (plus for some irregular cases)
- Important for computational reasons
- Used sometimes in some models to allow partial analytic marginalization
- Used for data augmentation as it alleviates computational burden
- No large benefits against current computational techniques, e.g., with dynamic Hamiltonian Monte Carlo used in Stan

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$p(\theta \mid y, n, M) \propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta \mid y, n, M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta \mid y, n, M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

after normalization

$$p(\theta \mid y, n, M) = \text{Beta}(\theta \mid \alpha + y, \beta + n - y)$$

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta \mid y, n, M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

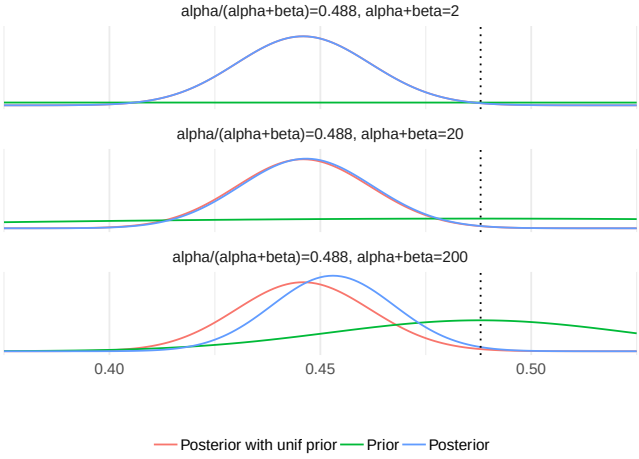
after normalization

$$p(\theta \mid y, n, M) = \text{Beta}(\theta \mid \alpha + y, \beta + n - y)$$

- $(\alpha - 1)$ and $(\beta - 1)$ can considered to be number of prior observations
- Uniform prior when $\alpha = 1$ and $\beta = 1$

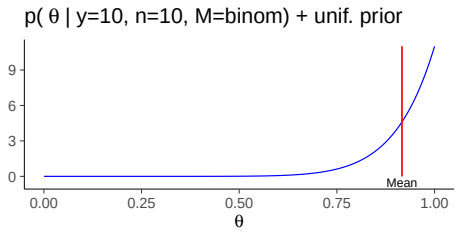
Placenta previa

- Beta prior centered on population average 0.485



Benefits of integration and prior

Example: $n = 10, y = 10$ - uniform vs Beta(2,2) prior



Beta prior for Binomial model

- Posterior

$$p(\theta \mid y, n, M) = \text{Beta}(\theta \mid \alpha + y, \beta + n - y)$$

- Posterior mean

$$\text{E}[\theta \mid y] = \frac{\alpha + y}{\alpha + \beta + n}$$

  - combination prior and likelihood information
  - when $n \to \infty$, $\text{E}[\theta \mid y] \to y/n = \bar{y} = \hat{\theta}_{\text{MLE}}$

Beta prior for Binomial model

- Posterior

$$p(\theta \mid y, n, M) = \text{Beta}(\theta \mid \alpha + y, \beta + n - y)$$

- Posterior mean

$$\text{E}[\theta \mid y] = \frac{\alpha + y}{\alpha + \beta + n}$$

  - combination prior and likelihood information
  - when $n \to \infty$, $\text{E}[\theta \mid y] \to y/n = \bar{y} = \hat{\theta}_{\text{MLE}}$

- Posterior variance

$$\text{Var}[\theta \mid y] = \frac{\text{E}[\theta \mid y](1 - \text{E}[\theta \mid y])}{\alpha + \beta + n + 1}$$

  - decreases as n increases
  - when $n \to \infty$, $\text{Var}[\theta \mid y] \to 0$

Noninformative prior, proper and improper prior

- Vague, flat, diffuse of noninformative
  - try to "to let the data speak for themselves"
  - flat is not non-informative
  - flat can be stupid
  - making prior flat somewhere can make it non-flat somewhere else
- Proper prior has $\int p(\theta) = 1$
- Improper prior density doesn't have a finite integral
  - the posterior can still sometimes be proper

Example of informative prior

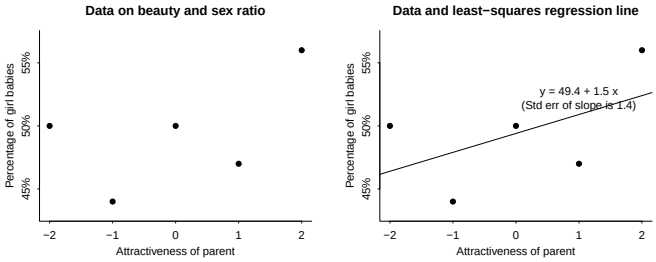
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate

Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)

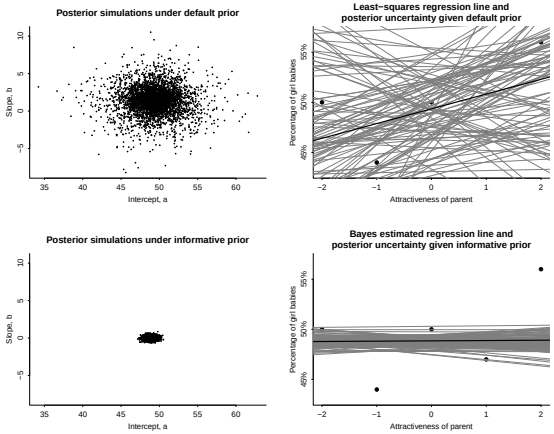
Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)



Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate

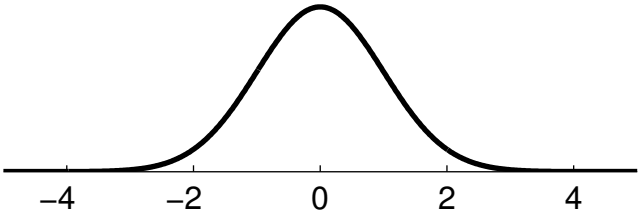


Normal / Gaussian

- Observations y real valued
- Mean θ and variance σ^2 (or deviation σ)
  (first assume σ^2 known)

$$p(y \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

$$y \sim \mathcal{N}(\theta, \sigma^2)$$



Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y \mid \theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y \mid \theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

Posterior $$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

Normal distribution - conjugate prior for θ

- Posterior (see ex 2.14a)

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta \mid y \sim \mathcal{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Normal distribution - conjugate prior for θ

- Posterior (see ex 2.14a)

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta \mid y \sim \mathcal{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- 1/variance = precision
- Posterior precision = prior precision + data precision
- Posterior mean is precision weighted mean

Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta \mid y) = \mathcal{N}(\theta \mid \mu_n, \tau_n^2)$$

$$\text{where} \quad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0

Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta \mid y) = \mathcal{N}(\theta \mid \mu_n, \tau_n^2)$$

$$\text{where} \quad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0
- If $\tau_0 \to \infty$ when n fixed
  or if $n \to \infty$ when τ_0 fixed

$$p(\theta \mid y) \approx \mathcal{N}(\theta \mid \bar{y}, \sigma^2/n)$$

Normal distribution - conjugate prior for θ

- Posterior predictive distribution

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta$$

$$p(\tilde{y} \mid y) \propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right) \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right) d\theta$$

$$\tilde{y} \mid y \sim \mathcal{N}(\mu_1, \sigma^2 + \tau_1^2)$$

- Predictive variance = observation model variance σ^2 + posterior variance τ_1^2