Chapter 2.1

Conjugate priors

Selecting priors

- Selecting the prior is one of the most important steps in a Bayesian analysis
- There is no "right" way to select a prior
- The choices often depend on the objective of the study and the nature of the data
  1. Conjugate versus non-conjugate
  2. Informative versus uninformative
  3. Proper versus improper
  4. Subjective versus objective

Conjugate priors

- A prior is **conjugate** if the posterior is a member of the same parametric family
- We have seen that if the response is binomial and we use a beta prior, the posterior is also a beta
- This requires a pairing of the likelihood and prior
- There is a long list of conjugate priors `https://en.wikipedia.org/wiki/Conjugate_prior`
- The advantage of a conjugate prior is that the posterior is available in closed form
- This is a window into Bayes learning and the prior effect

Conjugate priors

- Here is an example of a non-conjugate prior
- Say $Y \sim \textit{Poisson}(\lambda)$ and $\lambda \sim \text{Beta}(a, b)$
- The posterior is

$$f(\lambda|Y) \propto \left\{\exp(-\lambda)\lambda^Y\right\}\left\{\lambda^{a-1}(1-\lambda)^{b-1}\right\}$$

- This is not a beta PDF, so the prior is not conjugate
- In fact, this is not a member of any known (to me at least) family of distributions
- For some likelihoods/parameters there is no known conjugate prior

Estimating a proportion using the beta/binomial model

- A fundamental task in statistics is to estimate a proportion using a series of trials:
  - What is the success probability of a new cancer treatment?
  - What proportion of voters support my candidate?
  - What proportion of the population has a rare gene?
- Let $\theta \in [0, 1]$ be the proportion we are trying to estimate (e.g., the success probability).
- We conduct n independent trials, each with success probability θ, and observe $Y \in \{0, ..., n\}$ successes.
- We would like obtain the posterior of θ, a 95% interval, and a test that θ equals some predetermined value θ_0.

Frequentist analysis

- The maximum likelihood estimate is the sample proportion

$$\hat{\theta} = Y/n$$

- For large Y and $n - Y$, the sampling distribution of $\hat{\theta}$ is approximately

$$\hat{\theta} \sim \text{Normal}\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

- The standard error (standard deviation of the sampling distribution) is approximated as

$$\text{SE}(\hat{\theta}) \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

- A 95% CI is then

$$\hat{\theta} \pm 2\text{SE}(\hat{\theta})$$

Bayesian analysis - Likelihood

- Since Y is the number of successes in n independent trials, each with success probability θ, its distribution is

$$Y|\theta \sim Binomial(n, \theta)$$

- PMF: $P(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

- Mean: $E(Y|\theta) = n\theta$

- Variance: $V(Y|\theta) = n\theta(1-\theta)$

Bayesian analysis - Prior

- The parameter θ is continuous and between 0 and 1, therefore a natural prior is

$$\theta \sim \text{Beta}(a, b)$$

- PDF: $f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

- Mean: $E(\theta) = \frac{a}{a+b}$

- Variance: $V(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$

Derivation of the posterior

- The posterior is $\theta|Y \sim \text{Beta}(a + Y, b + n - Y)$

- See "Beta-binomial" in the online derivations

Derivation of the posterior

- The likelihood is $f(Y|\theta) = \binom{n}{Y}\theta^Y(1-\theta)^{n-Y}$
- The prior is $\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$
- The posterior is

$$
\begin{aligned}
p(\theta|Y) &= \frac{f(Y|\theta)\pi(\theta)}{m(Y)} \\
&= \frac{\left[\binom{n}{Y}\theta^Y(1-\theta)^{n-Y}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right]}{m(Y)}
\end{aligned}
$$

Derivation of the posterior

- Some housekeeping gives

$$\begin{aligned} p(\theta|Y) &= \left[\binom{n}{Y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{1}{m(Y)}\right]\theta^{Y+a-1}(1-\theta)^{n-Y+b-1} \\ &= C\theta^{A-1}(1-\theta)^{B-1} \end{aligned}$$

  where $A = Y + a$, $B = n - Y + b$ and C is the mess

- The terms that involve θ,

$$\theta^{A-1}(1-\theta)^{B-1},$$

  are the **kernel** of a Beta(A, B) distribution

- Therefore, $\theta|Y \sim \text{Beta}(Y + a, n - Y + b)$

Simplifying the derivations

- In the end, we are always going look at the terms that involve θ (the kernel) and find a matching distribution
- Therefore, the mess (C) will never be a factor
- Derivations simplify by absorbing all terms that do not include a θ into the normalizing constant
- For example, instead of

$$p(\theta|Y) = C\theta^{A-1}(1-\theta)^{B-1}$$

  we can write

$$p(\theta|Y) \propto \theta^{A-1}(1-\theta)^{B-1}$$

- "$\propto$" means "is proportional to"

Derivation of the posterior

- Here is a much simpler derivation

$$\begin{aligned} p(\theta|Y) &\propto f(Y|\theta)\pi(\theta) \\ &\propto \left[\theta^{Y}(1-\theta)^{n-Y}\right]\left[\theta^{a-1}(1-\theta)^{b-1}\right] \\ &\propto \theta^{A-1}(1-\theta)^{B-1} \end{aligned}$$

  where $A = Y + a$ and $B = n - Y + b$

- Therefore, $\theta|Y \sim \text{Beta}(Y + a, n - Y + b)$

- Note: $m(Y)$ was dropped in the first line, and thus is excluded from all these computations

Shrinkage

- The posterior mean is

$$\hat{\theta}_B = \mathrm{E}(\theta|Y) = \frac{Y+a}{n+a+b}$$

- The posterior mean is between the sample proportion Y/n and the prior mean $a/(a+b)$:

$$\hat{\theta}_B = w\frac{Y}{n} + (1-w)\frac{a}{a+b}$$

  where the weight on the sample proportion is $w = \frac{n}{n+a+b}$
- When (in terms of n, a and b) is the $\hat{\theta}_B$ close to Y/n?

- When is the $\hat{\theta}_B$ shrunk towards the prior mean $a/(a+b)$?

Selecting the prior

- The posterior is $\theta|Y \sim \text{Beta}(a + Y, b + n - Y)$
- Therefore, a and b can be interpreted as the "prior number of success and failures"
- This is useful for specifying the prior
- What prior to select if we have no information about θ before collecting data?
- What prior to select if historical data/expert opinion indicates that θ is likely between 0.6 and 0.8?

Related problem

- The success probability of independent trials is θ
- Y is the number of successes before we observe n failures
- Then $Y|\theta \sim \text{NegativeBinomial}(n, \theta)$ and

$$\text{Prob}(Y = y|\theta) = \binom{y+n+1}{y}\theta^y(1-\theta)^n$$

- Assume the prior $\theta \sim \text{Beta}(a, b)$ and find the posterior

Related problem

- The likelihood is $f(y|\theta) \propto \theta^y(1-\theta)^n$
- The prior is $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$
- Therefore, the posterior is

$$
\begin{aligned}
p(\theta|Y) &\propto [\theta^y(1-\theta)^n]\left[\theta^{a-1}(1-\theta)^{b-1}\right] \\
&= \theta^{A-1}(1-\theta)^{B-1}
\end{aligned}
$$

  where $A = y + a$ and $B = n + b$
- This is the kernel of the beta distribution, $\theta|Y \sim \text{Beta}(A, B)$

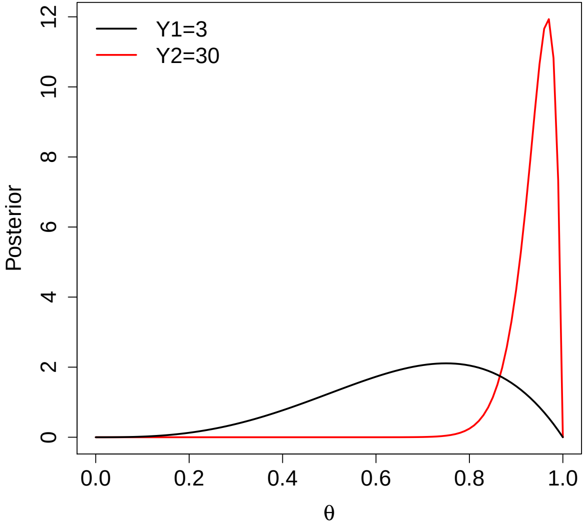
Smoking example

- Two smokers have just quit
- Say subject *i* has probability θ_i of abstaining each day
- The number of days until relapse for two patients is 3 and 30 days
- Can we conclude the patients have different probabilities of relapse?
- What is probability that their next attempts will exceed 30 days?

Smoking example

- The likelihood is $Y_i \sim \text{NegativeBinomial}(1, \theta_i)$
- Assume uniform priors $\theta_i \sim \text{Beta}(1, 1)$
- The posteriors are $\theta_i | Y_i \sim \text{Beta}(Y_i + 1, 2)$
- The posterior are plotted on the next slide
- The following slide uses Monte Carlo sampling to address the two motivating questions

Smoking example



Smoking example

```
> S      <- 1000000
> theta1 <- rbeta(S,3+1,2)
> theta2 <- rbeta(S,30+1,2)
> mean(theta2>theta1)
[1] 0.957222
>
> samp1 <- rnbinom(S,1,prob=1-theta1)
> samp2 <- rnbinom(S,1,prob=1-theta2)
> quantile(samp1,c(0.05,0.5,0.95))
5% 50% 95%
0   1  15
> quantile(samp2,c(0.05,0.5,0.95))
5% 50% 95%
0  13 109
> mean(samp1>30); mean(samp2>30)
[1] 0.015781
[1] 0.254129
```

Estimating a rate using the Poisson/gamma model

- Estimating a rate has many applications:
  - Number of virus attacks per day on a computer network
  - Number of Ebola cases per day
  - Number of diseased trees per square mile in a forest
- Let $\lambda > 0$ be the rate we are trying to estimate
- We make observations over a period (or region) of length (or area) N and observe $Y \in \{0, 1, 2, ...\}$ events
- The expected number of events is $N\lambda$ so that λ is the expected number of events per time unit
- MLE: $\hat{\lambda} = Y/N$ is the sample rate
- We would like obtain the posterior of λ

Bayesian analysis - Likelihood

- Since Y is a count with mean $N\lambda$, a natural model is

$$Y|\lambda \sim Poisson(N\lambda)$$

- PMF: $P(Y = y|\lambda) = \frac{\exp(-N\lambda)(N\lambda)^y}{y!}$

- Mean: $E(Y|\lambda) = N\lambda$

- Variance: $V(Y|\lambda) = N\lambda$

Bayesian analysis - Prior

- The parameter λ is continuous and positive, therefore a natural prior is
$$\lambda \sim \text{Gamma}(a, b)$$
- PDF: $f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$
- Mean: $\text{E}(\lambda) = \frac{a}{b}$
- Variance: $\text{V}(\lambda) = \frac{a}{b^2}$

Derivation of the posterior

- The likelihood is $\frac{\exp(-N\lambda)(N\lambda)^y}{y!} \propto \exp(-N\lambda)\lambda^y$
- The prior is proportional to $\exp(-b\lambda)\lambda^{a-1}$
- Therefore, the posterior is

$$\begin{aligned} p(\lambda|Y) &\propto [\exp(-N\lambda)\lambda^y]\left[\lambda^{a-1}\exp(-b\lambda)\right] \\ &= \lambda^{A-1}\exp(-B\lambda) \end{aligned}$$

  where $A = y + a$ and $B = N + b$

- The posterior is $\lambda|Y \sim \text{Gamma}(a + Y, b + N)$
- See “Poisson-gamma” in the online derivations

Shrinkage

- The posterior mean is

$$\hat{\lambda}_B = \mathrm{E}(\lambda|Y) = \frac{Y+a}{N+b}$$

- The posterior mean is between the sample rate Y/n and the prior mean a/b:

$$\hat{\theta}_B = w\frac{Y}{n} + (1-w)\frac{a}{b}$$

where the weight on the sample rate is $w = \frac{n}{n+b}$

- When (in terms of N, a and b) is the $\hat{\lambda}_B$ close to Y/n?

- When is the $\hat{\lambda}_B$ shrunk towards the prior mean a/b?

Selecting the prior

- The posterior is $\lambda|Y \sim \text{Gamma}(a + Y, b + N)$
- Therefore, a and b can be interpreted as the "prior number of events and observation time"
- This is useful for specifying the prior
- What prior to select if we have no information about θ before collecting data?
- What prior to select if historical data/expert opinion indicates that λ is likely between 0.6 and 0.8?

Posterior with two observations

- Derive the posterior if $Y_1 \sim \text{Poisson}(N_1\lambda)$; $Y_2 \sim \text{Poisson}(N_2\lambda)$; and $\lambda \sim \text{Gamma}(a, b)$
- Derive the posterior if $Y_i, ..., Y_m \sim \text{Poisson}(N\lambda)$ and $\lambda \sim \text{Gamma}(a, b)$
- We will work these problem in lab this week
- See “Poisson-gamma” in the online derivations

AB testing example

- A tech company runs their regular user interface for $N_1 = 8$ hours and gets $Y_1 = 4721$ clicks

- The next day they launch a new user interface for $N_2 = 8$ hours and get $Y_2 = 5209$ clicks

- Assuming uninformative conjugate priors, determine if the new user interface has a higher click rate

AB testing example

- Period 1: the likelihood is $Y_1|\lambda_1 \sim \text{Poisson}(N_1\lambda_1)$
- The conjugate prior is $\lambda_1 \sim \text{Gamma}(a, b)$
- The posterior is $\lambda_1|Y_1 \sim \text{Gamma}(Y_1 + a, N_1 + b)$
- Period 2: $\lambda_2|Y_2 \sim \text{Gamma}(Y_2 + a, N_2 + b)$

Monte Carlo approximation

```
> S <- 100000
> a <- b <- 0.1
> N1 <- N2 <- 8
> Y1 <- 4721
> Y2 <- 5209
>
> # MC samples
> lambda1 <- rgamma(S,Y1+a,N1+b)
> lambda2 <- rgamma(S,Y2+a,N2+b)
>
> # Prob(new interface is better|data)
> mean(lambda2>lambda1)
[1] 1
> # The new interface almost surely works!
```

Gaussian models

- The final distribution we'll discuss is the Gaussian (normal) distribution, $Y \sim \text{Normal}(\mu, \sigma^2)$
  - Domain: $Y \in (-\infty, \infty)$
  - PDF: $f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$
  - Mean: $\text{E}(Y) = \mu$
  - Variance: $\text{V}(Y) = \sigma^2$
- In this section, we will discuss:
  - Estimating the mean assuming the variance is known.
  - Estimating the variance assuming the mean is known.

Estimating a normal mean - Likelihood

- We assume the data consist of n independent and identically distributed observations $Y_1, ..., Y_n$.

- Each is Gaussian,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  where σ is known

- The likelihood is then

$$\prod_{i=1}^{n} f(y_i|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

Bayesian analysis - Prior

- The parameter μ is continuous over the entire real line, therefore a natural prior is

$$\mu \sim \text{Normal}(\theta, \tau^2)$$

- The prior mean θ is the best guess before we observe data

- The math is slightly more interpretable if we set $\tau^2 = \frac{\sigma^2}{m}$

- As we'll see, the prior variance via $m > 0$ controls the strength of the prior

Derivation of the posterior

- Then the posterior is ($w = n/(n+m)$)

$$\mu|Y_1, ..., Y_n \sim \text{Normal}\left(w\bar{Y} + (1-w)\theta, \frac{\sigma^2}{n+m}\right)$$

- See “normal-normal” in the online derivations

Shrinkage

- The posterior mean is

$$\hat{\mu}_B = \mathrm{E}(\mu | Y_1, ..., Y_n) = w\bar{Y} + (1 - w)\theta$$

  where $w = n/(n + m)$

- Therefore, if m is small then $\hat{\mu}_B \approx \bar{Y}$, and if m is large $\hat{\mu}_B \approx \theta$

- If no prior information is available, take m to be small and thus the prior is uninformative

- Small m gives large prior variance (relative to σ)

Shrinkage

- The posterior variance is

$$V(\mu | Y_1, ..., Y_n) = \frac{\sigma^2}{n + m}$$

- The sampling variance of $\bar{Y}$ is $\frac{\sigma^2}{n}$

- Therefore, we can loosely interpret m as the "prior number of observations"

Blood alcohol level analysis

- You are a defense attorney
- Your client is pulled over and given a breathalyzer test
- The $n = 2$ samples are $Y_1 = 0.082$ and $Y_2 = 0.084$
- The machine's error has SD 0.005 (not really)
- What prior should we choose?
- Use the online GUI to explore the posterior
  `https://shiny.stat.ncsu.edu/bjreich/BAC/`
- Is your client likely guilty of having BAC > 0.080?

Estimating a normal variance - Likelihood

- We assume the data consist of n independent and identically distributed observations $Y_1, ..., Y_n$.

- Each is Gaussian,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

  where μ is known

- The likelihood is then

$$\prod_{i=1}^{n} f(y_i|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

Bayesian analysis - Prior

- The parameter σ^2 is continuous over $(0, \infty)$, therefore a natural prior is $\sigma^2 \sim \text{Gamma}(a, b)$
- However, the math is easier if we pick a gamma prior for the inverse variance (precision) $1/\sigma^2$
- If $1/\sigma^2 \sim \text{Gamma}(a, b)$ then $\sigma^2 \sim \text{InverseGamma}(a, b)$
- This is the definition of the inverse gamma distribution
- The inverse gamma prior for σ^2 is PDF

$$f(\sigma^2) = \frac{b^a (\sigma^2)^{-a-1} \exp(-b/\sigma^2)}{\Gamma(a)}$$

Derivation of the posterior

- The posterior is

$$\sigma^2 | Y_1, ..., Y_n \sim \text{InverseGamma}\left(n/2 + a, SSE/2 + b\right)$$

  where $SSE = \sum_{i=1}^{n}(Y_i - \mu)^2$

- See “normal-inverse-gamma” in the online derivations

Shrinkage

- The mean of an InverseGamma(a, b) distribution only exists if $a > 1$
- The prior mean (if it exists) is $b/(a-1)$
- The posterior mean is

$$\frac{SSE + b}{n + 2a - 2}$$

- It is common to take a and b to be small to give an uninformative prior
- Then the posterior mean approximates the sample variance $SSE/(n-1)$

Conjugate prior for a normal precision

- The precision is the inverse variance, $\tau = 1/\sigma^2$
- If Y_i have mean μ and precision τ, the likelihood is proportional to

$$\prod_{i=1}^{n} f(y_i|\mu) \propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

- If $\tau \sim \text{Gamma}(a, b)$, then

$$\tau|Y \sim \text{Gamma}(n/2 + a, SSE/2 + b)$$

- This is the exact same analysis as the inverse gamma prior for the variance