

Last Lecture

- Key difference between frequentist and Bayesian inference is in use and definition of probability
 - Frequentist: looks at behavior of events in a large-number of hypothetical repetitions
 - Bayesian: Probability is used to measure state of knowledge
- Bayes theorem is used to update knowledge

Bayesian Advantages

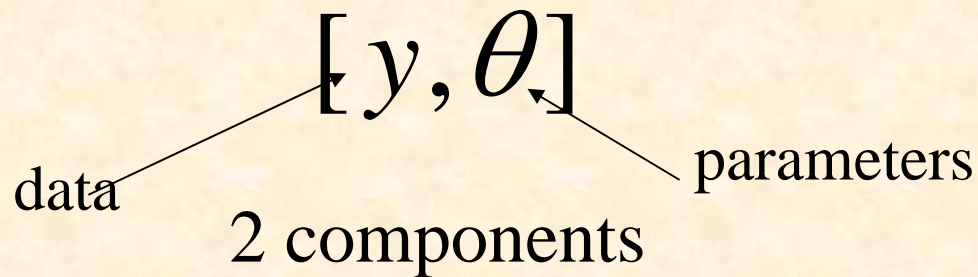
- All inference based on Bayes theorem
 - No worries about choice of estimator
- Richer model structure
 - Hierarchical modeling
- No reliance on asymptotics or tiresome optimality criteria
 - Works as advertised for all samples
- Focus on models not methods

Known and Unknown

- To a Bayesian quantities of interest are either known (observed) or unknown (unobserved)
- **Known:** data, covariates
- **Unknown:** parameters, predictions, missing data, true covariate values
- Model + Bayes theorem + data used to generate posterior distributions regarding unknowns.

Basic Bayes

All inference centers around the full (joint) probability model

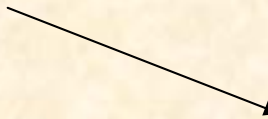


$$[y, \theta] = [\theta][y | \theta]$$


Prior distribution Sampling model (“likelihood”)

Posterior inference is from Bayes' Theorem

Knowledge about parameter
after data are collected


$$[\theta | y] = \frac{[\theta][y | \theta]}{\int [\theta][y | \theta] d\theta}$$

$$\propto [\theta][y | \theta]$$



Knowledge about parameter
before data are collected



Influence of data

The Posterior Distribution

- $$[\theta | y] = \frac{[y | \theta][\theta]}{\int [y | \theta][\theta] dy} \propto [y | \theta][\theta]$$
- In many simple examples, the posterior distribution can be found easily
 - Allows exact posterior inference
- In majority of cases, cannot be found exactly and must be approximated.

Binomial

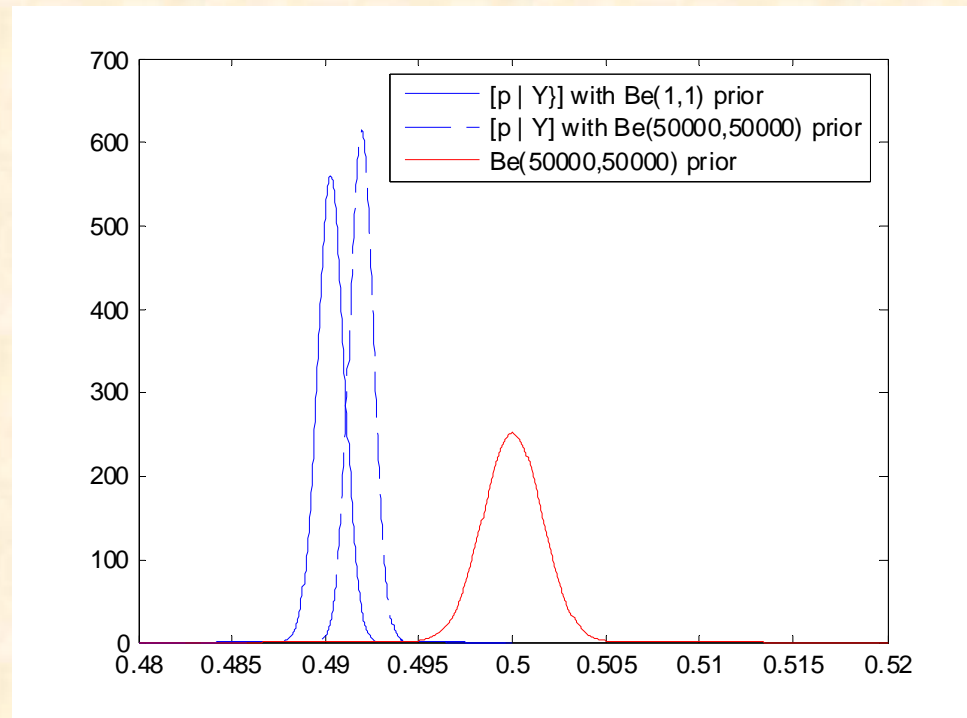
- Problem considered by Bayes (1763) and later by Laplace (1810)

$$[Y | p] = \binom{n}{p} p^Y (1-p)^{n-Y}$$

- Suppose $[p] = \text{Be}(\alpha, \beta)$
- $[p | Y = y] = \text{Be}(\alpha + y, n - y + \beta)$

Example

- Between 1745 and 1770 there were 241,945 girls and 251,527 boys born in Paris



Things to Note:

- The prior and the posterior are in the same family
- Calculation of the posterior easy
- Posterior exists in closed-form
- These three things are rarely true
- What if you prefer another prior that is not in the beta family?

$$[p] = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\text{logit}(p)-\psi)^2} \frac{1}{p(1-p)}$$

Something Else to Note (1)

- If $Y \sim \text{Be}(\alpha, \beta)$ then $E[Y] = \frac{\alpha}{\alpha + \beta}$

Prior (no data) Posterior data (no prior)

- $\frac{\alpha}{\alpha + \beta} \leq \frac{\alpha + y}{n + \alpha + \beta} \leq \frac{y}{n}$ or $\frac{\alpha}{\alpha + \beta} \geq \frac{\alpha + y}{n + \alpha + \beta} \geq \frac{y}{n}$

- Shrinkage – amount depends on how much data
 - Asymptotically data overwhelm prior
 - Asymptotically ML and Bayesian estimates coincide

Something Else to Note (2)

- If the prior is ‘flat’ (constant over all θ) then $[\theta | y] \propto [y | \theta]$
- E.g. $[p | y] \propto p^y(1-p)^{n-y}$
- Posterior is the likelihood scaled so that it integrates to 1.0
 - Asymptotically the posterior converges to the scaled likelihood

Something else to note (3)

- All the information in the data is contained in the likelihood
- Bayesian inference obeys the *likelihood principle*:

For a given sample of data any two probability models $p(y | \theta)$ that have the same likelihood function yield the same inference for θ

- Frequentist inference often violates the likelihood principle

More on the binomial- beta in a bit

Poisson

- $[Y | \lambda] = P(\lambda)$
- $[\lambda] = \text{Ga}(\alpha, \beta)$
- $[\lambda / Y = y]$ also a gamma distribution

Normal

- $[Y | \mu] = \text{N}(\mu, \tau)$ with τ known
- $[\mu] = \text{N}(\psi, \tau_\mu)$
- $[\mu / Y = y]$ also a normal distribution

Normal

- $[Y | \tau] = \text{N}(\mu, \tau)$ with μ known
- $[\tau] = \text{Ga}(\alpha, \beta)$
- $[\tau / Y = y] = \text{Ga}(\alpha + 1/2, \beta + (y - \mu)^2/2)$

Conjugacy

- Examples considered so far all easy
 - Posterior in same family as prior
- **Conjugate prior family:**
 - Family of prior probability distributions with the property that the posterior probability distribution also belongs to that family.
 - All members of the exponential family have conjugate priors
- What if the prior family is not conjugate?

Use of Conjugate Priors

- Easy to understand the results
- Able to express (usually) results in analytic form
- Simplify calculations
- Useful as building blocks for more complicated models, such as hierarchical models (considered later)

Exponential Family

$$[Y | \theta] = a(Y)b(\theta)e^{c(\theta)'d(Y)}$$

$$[Y_1, \dots, Y_n | \theta] \propto b(\theta)^n e^{c(\theta)'T(Y)}$$

$$[\theta] \propto b(\theta)^\alpha e^{c(\theta)'\beta}$$

$$[\theta | y] \propto b(\theta)^{\alpha+n} e^{c(\theta)'(\beta+T(y))}$$

Only class of distributions with natural conjugate priors

Example – Exponential Dist.

$$[Y_1, \dots, Y_n | \theta] = \theta^n e^{-n\bar{y}\theta}$$

$$[\theta] \propto \theta^{\alpha-1} e^{-\theta\beta}$$

$$[\theta | y_1, \dots, y_n] \propto \theta^{\alpha+n-1} e^{-\theta(\beta+n\bar{y})}$$

Posterior Summaries

- $[\theta | y]$ contains all the current information about θ
- Graphical displays useful – contour plots for >1 parameter
- Posterior inference based on simple intuitive summaries
 - Easy to make inference about complicated transformations

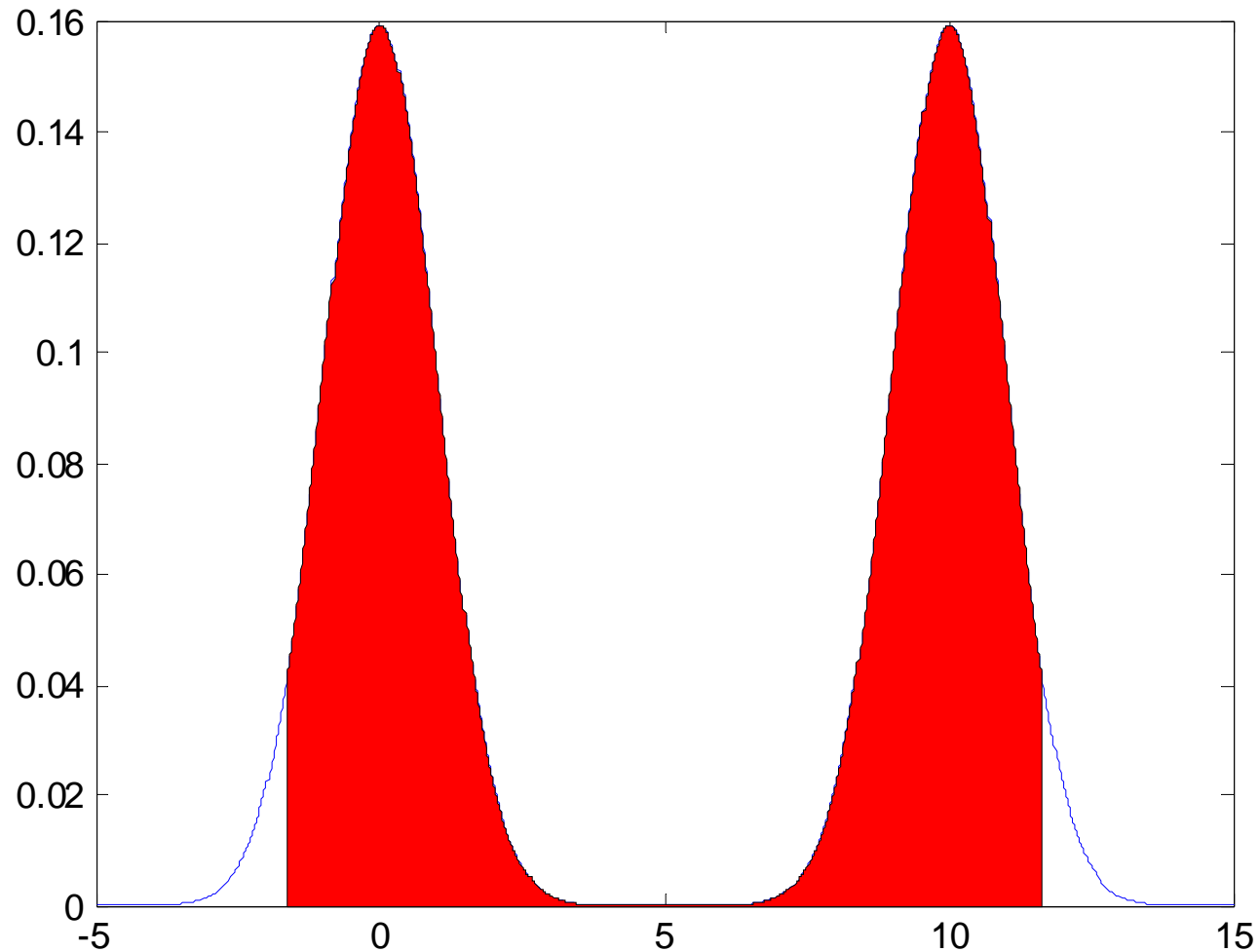
Posterior Summaries

- Often want to summarize the posterior with simple statistics
 - Mean, median, mode, SD, quantiles
- If the posterior is available analytically, summaries often available in closed-form
- If not available analytically, draw samples from the posterior distribution

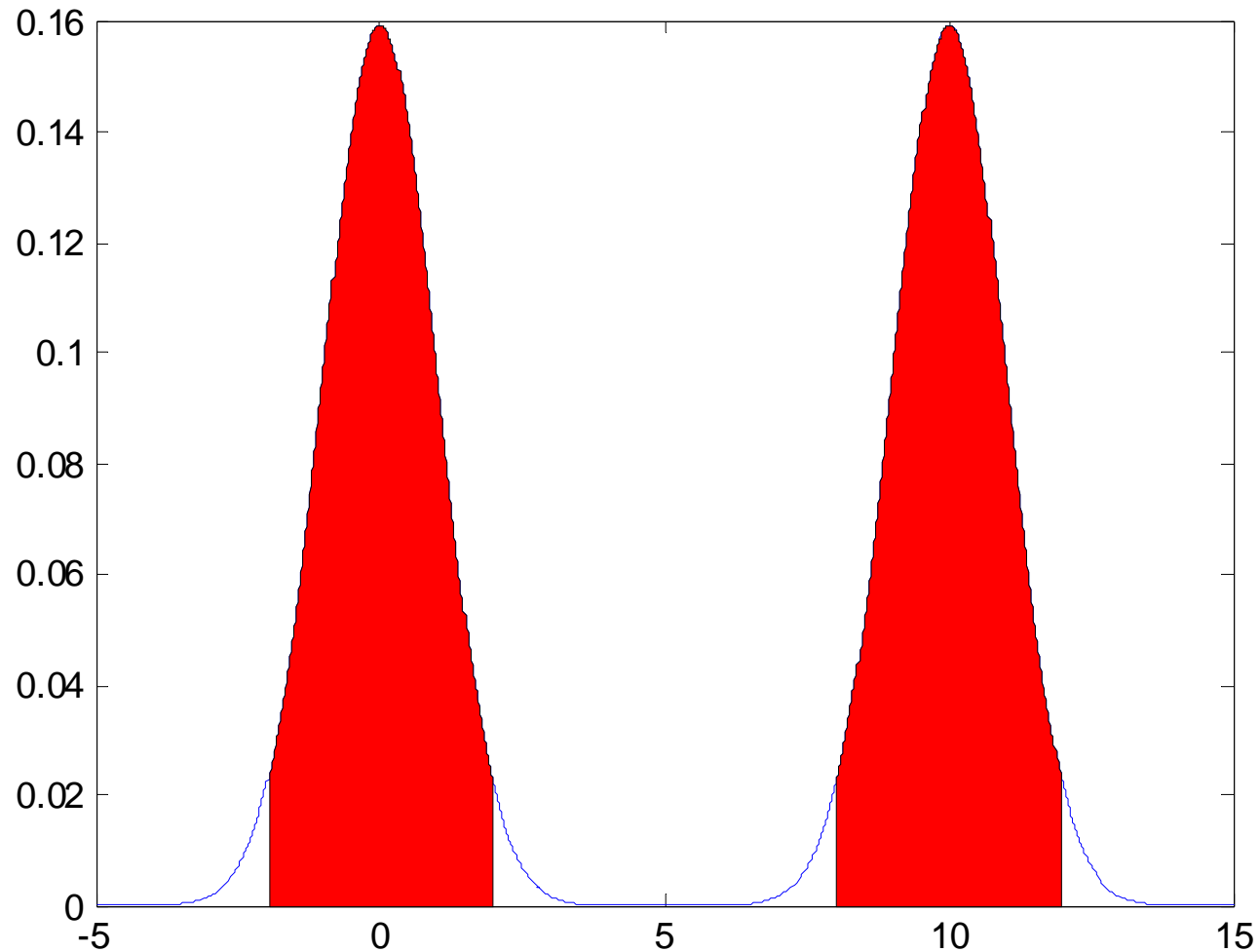
Posterior Intervals

- Intervals of the form $\{ \theta: Pr(\theta \in (\theta_1, \theta_2)) = 1-\alpha \}$ useful
 - We will refer to this as a $100(1-\alpha)\%$ credible interval (“Bayesian confidence interval”)
- A $100(1-\alpha)\%$ *central posterior interval* given by the pair of points below and above which lie exactly $100(1-\alpha/2)\%$ of posterior probability
 - Most common
- Can also compute shortest (continuous) central or highest posterior density (HPD)

Central posterior interval- Sometimes it stinks



Better: Highest probability density



Some single parameter models

- Binomial
- Normal known variance
- Poisson

Binomial

y successes in n independent, Bernoulli trials

Data model

$$[y | p] = \binom{n}{y} p^y (1 - p)^{n-y}$$

Take for now a prior distribution on p as uniform on the interval $(0,1)$.

$$[p] = 1 : 0 < p < 1$$

$$[p] = 0; \text{ otherwise}$$

Posterior distribution

$$[p | y] = \frac{[y | p][p]}{[y]} = \frac{\binom{n}{y} p^y (1-p)^{n-y}}{[y]}$$

note that both $\binom{n}{y}$ and $[y]$ are constants (do not vary with p) so

$$[p | y] \propto p^y (1-p)^{n-y}$$

This can be recognized as the kernel of a beta distribution with parameters

$$\alpha = y + 1, \beta = n - y + 1$$

$$[p | y] = cp^{\alpha-1}(1-p)^{\beta-1}$$

where

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

is the constant of integration assuring that

$$\int_0^1 [p | y] dp = 1$$

More generally

Take as a prior

$$[p] \sim \text{Beta}(\alpha, \beta)$$

$$[p] \propto p^{\alpha-1} (1-p)^{\beta-1}$$

Then for the binomial sampling model

$$\begin{aligned} [p | y] &\propto p^y (1-p)^{n-y} p^{\alpha-1} (1-p)^{\beta-1} \\ &= p^{y+\alpha-1} (1-p)^{n-y+\beta-1} \end{aligned}$$

This is the kernel of a Beta distribution with parameters

$$\alpha + y \quad \beta + n - y$$

One interpretation

- Previous experiment with
 - $\alpha - 1$ successes $\beta - 1$ failures $\alpha + \beta - 2$ trials
- Uniform is equivalent to 0 trials (no information)
 - Beta (1,1)

Inference

- Get directly from the Beta distributions (means, variances, quantiles, etc.)
- In particular:

$$E(p) = \frac{\alpha}{\alpha + \beta} \quad \text{var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

$$E(p | y) = \frac{\alpha + y}{\alpha + \beta + n} \quad \text{var}(p | y) = \frac{E(p | y)[1 - E(p | y)]}{\alpha + \beta + n + 1}$$

Data only:

$$E(p) = \frac{y}{n} \quad \text{var}(p) = \frac{E(p)[1 - E(p)]}{n}$$

Binomial example

- In 50 previous nest searches for the same species, 14 were parasitized
- In our current sample of 10 nests, 6 were parasitized

Binomial example

$$[p] \sim \text{Beta}(14 + 1, 50 - 14 + 1) = \text{Beta}(15, 37)$$

$$[p | y] \propto p^{15+6-1} (1-p)^{37+10-6-1}$$

$$[p | y] \sim \text{Beta}(21, 41)$$

Binomial example

Prior mean and variance

$$E(p) = \frac{\alpha}{\alpha + \beta} = \frac{15}{15 + 37} = 0.288$$

$$\text{var}(p) = \frac{15(37)}{(15 + 37)^2 (15 + 37 + 1)} = 0.003873$$

Posterior mean and variance

$$E(p | y) = \frac{\alpha + y}{\alpha + \beta + n} = \frac{15 + 6}{15 + 37 + 10} = 0.33871$$

$$\text{var}(p | y) = \frac{E(p | y)[1 - E(p | y)]}{\alpha + \beta + n + 1} = 0.003555$$

Binomial example

Just the data

$$E(p) = \frac{y}{n} = 0.6$$

$$\text{var}(p) = \frac{E(p)[1 - E(p)]}{n} = 0.024$$

So prior has big effect.

Bigger sample, same prior

- 1000 nests, 589 parasitized

$$[y, \theta] = [\theta][y | \theta]$$

$$\text{var}(p | y) = 0.000232$$

With no prior

$$E(p) = \frac{y}{n} = 0.589$$

$$\text{var}(p) = 0.000242$$

Now the data rules!

Problem in Excel

- Calculate means, variances directly
- Use BETADIST(x,alpha, beta) to get cumulative distribution function F(x)
 - Differencing to get probability density function
 - $f(x+d) \sim F(x+d) - F(x)$
 - Compute for both prior and posterior and plot
- Quantiles from BETAINV(cum_prob,alpha,beta)
- [posteriors.xls](#)

Normal distribution

2 parameters: mean μ and variance σ^2

Sampling model for 1 observation

$$[y | \mu] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]$$

$$\propto \exp\left[-\frac{\tau}{2}(y - \mu)^2\right]$$

$$\tau = 1 / \sigma^2 = \textit{precision}$$

Conjugate prior= Normal

$$[\mu] \propto \exp\left[-\frac{\tau_0}{2}(\mu - \mu_0)^2\right]$$

Posterior

$$[\mu | y] \propto \exp\left[-\frac{\tau_0}{2}(\mu - \mu_0)^2\right] \exp\left[-\frac{\tau}{2}(y - \mu)^2\right]$$

$$= \exp\left[-\frac{\tau_1}{2}(\mu - \mu_1)^2\right]$$

Prior precision

Data precision

Kernel of a Normal

$$\mu_1 = \frac{\tau_0 \mu_0 + \tau y}{\tau_0 + \tau}$$

$$\tau_1 = \tau_0 + \tau$$

Normal- n observations

$$[\mu | \bar{y}] \propto \exp\left(-\frac{1}{2}\left[\tau_0(\mu - \mu_0)^2 + \tau \sum_{i=1}^n (y_i - \mu)^2\right]\right)$$

Kernel of a Normal distribution

$$\mu_1 = \frac{\tau_0 \mu_0 + \tau n \bar{y}}{\tau_0 + n \tau}$$

$$\tau_1 = \tau_0 + n \tau$$

Exercise (1)

- Prior knowledge suggests that duck weights are normally distributed with a mean of 1000g and SD of 100g
- You take a sample of 100 ducks and obtain a sample mean of 1055 g and SD of 50 g
 - Assume that SD is known
- Obtain prior and posterior mean, precision, 2.5%, 50%, and 97.5% quantiles. Plot prior and posterior.
 - Hint: $\tau = 1./SD^2$
 - Hint (2): Excel NORMDIST() and NORMINV() use mean and SD

Normal distribution

- Known mean, unknown variance
 - Posterior winds up being a chi-square when prior is inverse gamma
- Unknown mean, unknown variance-
 - Messier, but conjugate priors still exist

Poisson distribution

- Discrete, used for counting events, abundance, etc.

For a vector

$$y = (y_1, y_2, \dots, y_n)$$

$$[y | \lambda] = \prod_{i=1}^n \frac{1}{y_i!} \lambda^{y_i} e^{-\lambda} \propto \lambda^{n\bar{y}} e^{-n\lambda}$$

where

$$n\bar{y} = \sum_{i=1}^n y_i$$

Conjugate prior

$$[\lambda] \propto e^{-\beta\lambda} \lambda^{\alpha-1}$$

which is the kernel of a gamma distribution:

Gamma (α, β)

$$[\lambda] = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\lambda} \lambda^{\alpha-1}$$

$$E[\lambda] = \frac{\alpha}{\beta} \quad \text{var}[\lambda] = \frac{\alpha}{\beta^2}$$

Posterior

$$\begin{aligned} [\lambda | y] &\propto e^{-\beta\lambda} \lambda^{\alpha-1} \lambda^{n\bar{y}} e^{-n\lambda} \\ &= e^{-(\beta+n)\lambda} \lambda^{\alpha+n\bar{y}-1} \end{aligned}$$

Kernel of a Gamma $(\alpha + n\bar{y}, \beta + n)$

Exercise (2)

- Prior knowledge suggests that density of fish per reach is $E[\lambda]=20$ with $\text{var}[\lambda]=16$
- You take 10 electroshock samples and obtain an average count of 15 fish
- Assuming a gamma prior and Poisson sampling model
 - Obtain prior and posterior $E[N]$, $\text{var}[N]$, 2.5%, 50%, and 97.5% quantiles. Plot prior and posterior.
 - Hint(1): $\beta = E[\lambda] / \text{var}[\lambda]$ $\alpha = E[\lambda]\beta$
 - Hint (2): the beta in the gamma distribution in Excel is $1./\beta$
e.g.=GAMMAINV(.025,alpha,1/beta)
Gives the 0.025 quantile when alpha, beta are defined as above

Negative Binomial

Marginal distribution of y (single observation)

$$[y] = \int [y, \lambda] d\lambda = \int [y | \lambda][\lambda] d\lambda$$

When prior in conjugate can find via relationship

$$[y] = \frac{[y | \lambda][\lambda]}{[\lambda | y]}$$

Under gamma-Poisson:

$$[y] = \frac{\text{Poisson}(y | \lambda) \text{Gamma}[\lambda | \alpha, \beta]}{\text{Gamma}(\lambda | \alpha + y, \beta + 1)} = \binom{\alpha + y - 1}{y} \left(\frac{\beta}{\beta + 1} \right)^\alpha \left(\frac{1}{\beta + 1} \right)^y$$

$$[y] = \text{NegBin}(\alpha, \beta)$$

Inference about Functions of Parameters

- Inference regarding functions of parameters straightforward in a Bayesian framework
- Suppose $Y \sim B(n, p)$ and based on $Y = y$ we wish to make inference about the log-odds:

$$\eta = \ln \left(\frac{p}{1-p} \right)$$

Transformation of Variables

- Given $[y]$ what is $[w]$ where $w = f(y)$?
 - If y is discrete and f invertible

$$[w] = [f^{-1}(w)]$$

- If y is discrete and f invertible

$$[w] = [f^{-1}(w)] | J |$$

where J is the Jacobian of $y = f^{-1}(w)$

Models with >1 parameter

- Virtually every problem in statistics has more than one unknown
 - Usually interested in a subset of parameters
 - Nuisance parameters are those needed in the model to ensure proper inference but that are of little interest
- Multiparameter problems highlight the advantages of Bayesian inference
- Require the *marginal* distribution of the parameters of interest.

Bayes theorem for >1 Parameter

- If Y depends on $\theta_1, \dots, \theta_k$ then:

$$[\theta_1, \dots, \theta_k | Y = y] \propto [Y | \theta_1, \dots, \theta_k][\theta_1, \dots, \theta_k]$$

- $[\theta_1, \dots, \theta_k | Y = y]$ is the *joint posterior* distribution

Marginal Distribution

- If a random vector Z can be partitioned into X and Y the distribution of X can be found by integrating (summing) over the components of Y :

$$[X] = \int [X, Y] dY$$

- $[X] = f_X(x)$ is the marginal distribution of X

Example

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{3x} & x = 1, 2, 3 \text{ and } 0 \leq y < x \\ 0 & \text{elsewhere} \end{cases}$$

Joint Posterior Inference

- Inference can be made using joint posteriors
 - Contour plots for pairs of variables
- Can also examine parameters one at a time using marginal posterior distributions

Example

- A random sample drawn from a $N(\mu, \tau_y)$ with both parameters unknown
 - $N(\psi, \kappa\tau_y)$ prior for μ
 - $\text{Ga}(\alpha, \beta)$ prior for τ_y
- (1) Inference about μ (nasty exercise)
- (2) Inference about τ

Prior Predictive Distribution

- Inference about an unknown observable is called *prediction*
- Before the data have been collected we can predict what a new value would be using the *prior predictive distribution* (also called the *marginal distribution*)

Prior predictive (marginal) distribution

$$[y] = \int [y, \lambda] d\lambda = \int [y | \lambda][\lambda] d\lambda$$

Posterior Predictive Distribution

- Inference about an unknown observable is called *prediction*
- After the data have been collected we can predict what a new value would be using the *posterior predictive distribution*:

$$\begin{aligned} [y^{new} | y] &= \int [y^{new}, \theta | y] d\theta \\ &= \int [y^{new} | \theta][\theta | y] d\theta \end{aligned}$$

Example

- After carrying out a binomial experiment with y successes in n trials we wish to predict the outcome on one more trial
 - Assume a $\text{Be}(\alpha, \beta)$ prior for p

$$\Pr(y^{\text{new}} = 1 \mid y) = \int_0^1 \Pr(y^{\text{new}} = 1)[p \mid y]dp$$

- $\Pr(y^{\text{new}} = 0) = 1 - \Pr(y^{\text{new}} = 1 \mid y)$ completes our distribution

Generally get this by simulation!