Lecture

Bayesian inference

Learning objectives

After this session students should be able to:

- List the fundamental components of Bayesian inference
- Describe Bayes' Theorem
- Describe what a conjugate model is and its features
- Describe the conjugate Binomial-Beta, Normal-Normal and Poisson-Gamma models
- Compare prior and posterior distributions

Outline

- Why Bayesian methods?
- Likelihood, prior and posterior: how are they connected?
- Bayes's theorem and example in diagnostic setting
- Inference on proportions: Binomial-Beta models
- Inference on continuous data: Normal-Normal models
- Inference on count data: Poisson-Gamma models

Why Bayesian methods?

- Bayesian methods have been widely applied in many areas:
  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....
- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

Example

A clinical trial is carried out to collect evidence about an unknown 'treatment effect'

Conventional analysis

- p-value for H_0: treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

Aim is to learn what this trial tells us about the treatment effect

Bayesian analysis

- Inference is based on probability statements summarising the posterior distribution of the treatment effect

Asks: 'how should this trial change our opinion about the treatment effect?'

Components of a Bayesian analysis

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the prior distribution)
- the support for different values of the treatment effect based *solely* on data from the trial (the likelihood),

and to combine these two sources to produce

- a final opinion about the treatment effect (the posterior distribution)

The final combination is done using Bayes theorem (and only simple rules of probability), which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution

One can view the Bayesian approach as a formalisation of the process of learning from experience

Bayesian inference: the posterior distribution

Posterior distribution forms basis for all inference — can be summarised to provide

- point and interval estimates of Quantities of Interest (QOI), e.g. treatment effect, small area estimates, . . .
- point and interval estimates of any function of the parameters
- probability that QOI (e.g. treatment effect) exceeds a critical threshold
- prediction of QOI in a new unit
- prior information for future experiments, trials, surveys, . . .
- inputs for decision making
- . . .

Bayes theorem and its link with Bayesian inference

Bayes' theorem

- Provable from probability axioms
- Let A and B be events, then

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- Similarly

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

- If A_i is a set of mutually exclusive and exhaustive events (*i.e.* $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}$$

Example: use of Bayes theorem in diagnostic testing

A new HIV test is claimed to have “95% sensitivity and 98% specificity”

In a population with an HIV prevalence of 1/1000, what is the chance that a patient testing positive actually has HIV?

- Let A be the event that patient is truly HIV positive, $\overline{A}$ be the event that they are truly HIV negative
- Let B be the event that they test positive
- We want $p(A|B)$
- “95% sensitivity” means that $p(B|A) = .95$
- “98% specificity” means that $p(B|\overline{A}) = .02$
- Now Bayes theorem says

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\overline{A})p(\overline{A})}$$

- Hence $p(A|B) = \frac{.95\times.001}{.95\times.001+.02\times.999} = .045$
- Thus over 95% of those testing positive will, in fact, not have HIV

Comments

- Our intuition is poor when processing probabilistic evidence
- The vital issue is *how should this test result change our belief that patient is HIV positive?*
- The disease prevalence can be thought of as a *'prior'* probability ($p = 0.001$)
- Observing a positive result causes us to modify this probability to $p = 0.045$. This is our *'posterior'* probability that patient is HIV positive.
- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established
- More controversial is the use of Bayes theorem in general statistical analyses, where *parameters* are the unknown quantities, and their prior distribution needs to be specified — this is Bayesian inference

Bayesian inference

Makes fundamental distinction between

- Observable quantities y, i.e. the data
- Unknown quantities θ
  - ▸ θ can be statistical parameters, missing data, mismeasured data …
  - → parameters are treated as random variables
  - → in the Bayesian framework, we make probability statements about model parameters
- ! in the Frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data

As with any statistical analysis, we start building a model which specifies $p(y \mid \theta)$

This is the likelihood, which relates all variables into a ‘full probability model’

Bayesian inference [continued]

From a Bayesian point of view

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data

  → need to specify a prior distribution $p(\theta)$

- y is known so we should condition on it

  → use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(\theta)\, p(y \mid \theta)$$

This is the posterior distribution

The prior distribution $p(\theta)$, expresses our uncertainty about θ before seeing the data

The posterior distribution $p(\theta \mid y)$, expresses our uncertainty about θ after seeing the data

Example: Inference on proportions

- Recall example from Lecture 2 where we consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpreted this as a distribution with mean = 0.4, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties
- Suppose we now treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses

Identifying the different model components

Likelihood (distribution of the data):

- Assuming patients are independent, with common unknown response rate θ, leads to a binomial likelihood

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

θ needs to be given a continuous prior distribution:

- Recall from Lecture 2 that the Beta distribution is used in these cases

$$\theta \sim \text{Beta}(a, b)$$
$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

Combining prior and likelihood

Combining the Binomial likelihood and the Beta prior gives the following posterior distribution

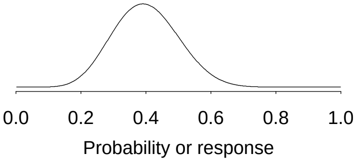
$$\begin{aligned} p(\theta \mid y, n) &\propto p(y \mid \theta, n)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{n-y+b-1} \end{aligned}$$

The posterior is still a Beta distribution (with different parameters):

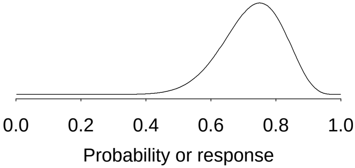
$$p(\theta \mid y, n) \propto \text{Beta}(y+a,\ n-y+b)$$

When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood

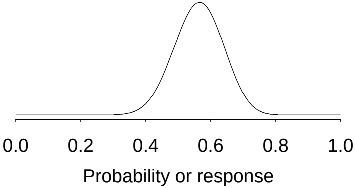
Prior, likelihood and posterior for Drug example



Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6



Likelihood arising from a Binomial observation of 15 successes out of 20 cases



Parameters of the Beta distribution are updated to (a+15, b+20-15) = (24.2, 18.8): mean 24.2/(24.2+18.8) = 0.56

Posterior inference using simulation methods

- In the Drug example, we have calculated the posterior distribution in closed form
  - this is possible because we are using conjugate priors
- No need to explicitly calculate posterior if using simulation methods
- In OpenBUGS just specify prior and likelihood separately
- OpenBUGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood
  - posterior doesn't need to be closed form
  - but can recognise conjugacy when it exists, in which case OpenBUGS will sample directly from the closed form posterior

OpenBUGS code for drug model

- Drug model in OpenBUGS syntax:

```
model {
 theta ~ dbeta(9.2, 13.8) # prior distribution
 y     ~ dbin(theta, 20)  # sampling distribution
 y <- 15                  # data
}
```

- Note that the only difference between this code and the code for implementing the drug model for predicting *y* seen in Lecture 2 is that we now specify the observed value of *y* from our study

- OpenBUGS automatically knows that if any variables have observed values (i.e. data) then these need to be conditioned on, and that posterior inference (rather than forward sampling) is required

Posterior Predictions for Drug example

Now suppose we would consider continuing a development program if the drug managed to achieve at least a further 25 successes out of m=40 future trials. How do we include this into the model?

$$\begin{array}{rcll}
\theta & \sim & \mathrm{Beta}[a, b] & \text{prior distribution} \\
y & \sim & \mathrm{Binomial}[\theta, n] & \text{sampling distribution} \\
y_{\mathrm{pred}} & \sim & \mathrm{Binomial}[\theta, m] & \text{predictive distribution} \\
P_{\mathrm{crit}} & = & P(y_{\mathrm{pred}} \geq m_{\mathrm{crit}}) & \text{Probability of exceeding critical threshold}
\end{array}$$

In BUGS syntax:

```
# Model description
model {
  theta       ~ dbeta(a,b)              # prior distribution
  y           ~ dbin(theta,n)           # sampling distribution
  y.pred      ~ dbin(theta,m)           # predictive distribution
  P.crit   <- step(y.pred-mcrit+0.5) # =1 if y.pred >= mcrit,
                                         0 otherwise
}
```

Data files

Data can be written after the model description, or held in a separate .txt or .odc file

```
 list( a = 9.2,    # parameters of prior distribution
  b = 13.8,
  y = 15,     # number of successes
  n = 20,     # number of trials
  m = 40,     # future number of trials
  mcrit = 25) # critical value of future successes
```

Alternatively, in this simple example, we could have put all data and constants into model description:

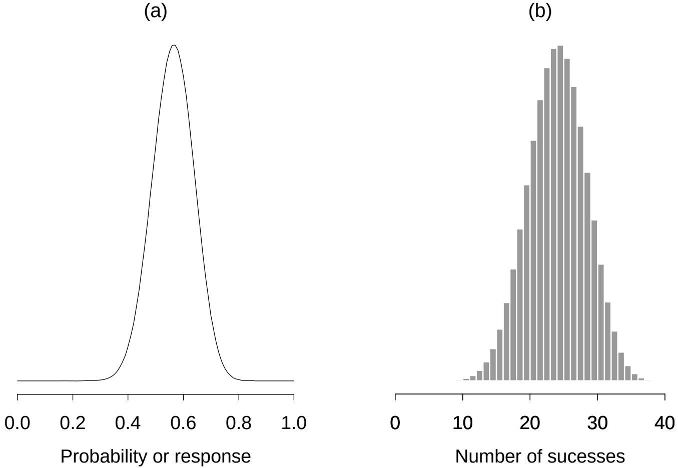
```
model{
  theta     ~ dbeta(9.2,13.8)      # prior distribution
  y         ~ dbin(theta,20)       # sampling distribution
  y.pred    ~ dbin(theta,40)       # predictive distribution
  P.crit   <- step(y.pred-24.5)    # =1 if y.pred >= mcrit,
                                      0 otherwise

  y        <- 15
}
```

Posterior and predictive distributions for Drug example

(a) Beta posterior distribution for θ after observing 15 successes in 20 trials

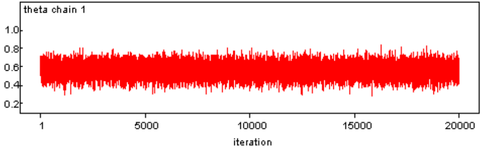
(b) Predictive Beta-Binomial distribution of the number of successes y_{pred} in the next 40 trials: mean 22.5 and standard deviation 4.3

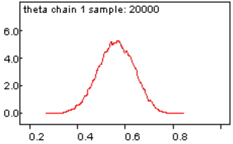


OpenBUGS output and exact answers

OpenBUGS

```
node          mean      sd        MC error    2.5%     median    97.5%    start   sample
theta        0.5633    0.07458   4.292E-4    0.4139   0.5647    0.7051   1001    30000
y.pred       22.52     4.278     0.02356     14.0     23.0      31.0     1001    30000
P.crit       0.3273    0.4692    0.002631    0.0      0.0       1.0      1001    30000
```





Exact answers from conjugate analysis

- θ: mean 0.563 and standard deviation 0.075
- y_{pred}: mean 22.51 and standard deviation 4.31
- Probability of at least 25: 0.329

Summary

So far

- We have introduced the two components which play a key role in Bayesian analysis (prior and likelihood) and we have seen how to combine them to obtain the posterior distribution through Bayes' theorem
- We have familiarised with conjugate models (when the prior and the posterior distribution come from the same family)
- We have learnt about the Binomial-Beta model and seen how this can be applied in OpenBUGS

BREAK

Outline

- Why Bayesian methods?
- Likelihood, prior and posterior: how are they connected?
- Bayes's theorem and example in diagnostic setting
- Inference on proportions: Binomial-Beta models
- Inference on continuous data: Normal-Normal models
- Inference on count data: Poisson-Gamma models

Conjugate Bayesian inference for continuous data: THM example

- Regional water companies in the UK are required to take routine measurements of trihalomethane (THM) concentrations in tap water samples for regulatory purposes
- Samples are tested throughout the year in each water supply zone
- Suppose we want to estimate the mean THM concentration in a particular water zone
- Two independent measurements, y_1 = 128$\frac{\mu g}{l}$ and y_2=132$\frac{\mu g}{l}$ are taken, and their mean, $\overline{y}$, is 130$\frac{\mu g}{l}$
- Suppose we know that the true standard deviation of THM measurements in this water zone is $\sigma = 5\frac{\mu g}{l}$ (from the known assay measurement error)
- What should we estimate the mean THM concentration to be in this water zone?

Denote the mean THM concentration for the zone by θ

Frequentist analysis

- A standard analysis would use the sample mean $\overline{y}$ = $130\frac{\mu g}{l}$ as an estimate of θ, with standard error $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{2}} = 3.5\frac{\mu g}{l}$
- A 95% confidence interval is $\overline{y} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$, i.e. 123.1 to 136.9 $\frac{\mu g}{l}$

Bayesian analysis

As we did previously with the Binomial-Beta model, try to think of the two components which we need to perform Bayesian inference:

- Likelihood (distribution of the data)
- Prior belief

Components of a Bayesian analysis

Likelihood:

$$y_i \sim \mathrm{N}(\theta, \sigma^2) \quad (i = 1, ..., n)$$

N.B. Here we assume that σ^2 is known

θ **is given a prior distribution**:

$$\theta \sim \mathrm{N}(\mu, \omega^2)$$

- It is convenient to write the prior variance as a function of the data variance, i.e. $\omega^2 = \frac{\sigma^2}{n_0}$
- We will see that n_0 $(= \frac{\sigma^2}{\omega^2})$ can be interpreted as an implicit prior sample size

OpenBUGS notation: `y ~ dnorm(theta, tau)` where `tau` is the inverse of the variance

Specifying values for the parameters of the prior

- Suppose historical data on THM levels in other zones supplied from the same source showed that the average of the zone-specific mean THM concentrations was 120 $\frac{\mu g}{l}$ with standard deviation 10 $\frac{\mu g}{l}$

- Suggests $\mathrm{N}(120, 10^2)$ prior for θ

- Expressing the prior standard deviation as a function of the sampling standard deviation, $\omega^2 = \frac{\sigma^2}{n_0}$, and solving for n_0 gives $n_0 = \frac{\sigma^2}{\omega^2} = \frac{5^2}{10^2} = 0.25$

- So our prior can be written as $\theta \sim \mathrm{N}(120, \frac{\sigma^2}{0.25})$

- As n_0 tends to 0, the prior variance becomes larger and the distribution becomes ‘flatter’

Combining prior and likelihood

Combining the Normal likelihood and the Normal prior

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

$$p(\theta) = \frac{1}{\sqrt{2\pi(\sigma^2/n_0)}} e^{-(\theta-\mu)^2 / 2(\sigma^2/n_0)}$$

$$p(\theta|\mathbf{y}) \propto \frac{1}{\sqrt{\sigma^{2n}(\sigma^2/n_0)}} e^{\frac{-(\theta-\mu)^2}{2(\sigma^2/n_0)} + \frac{-\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2}}$$

After some algebra the exponent can be rearranged to

$$-\frac{1}{2}\left[\frac{\left(\theta - \frac{n_0\mu + n\bar{y}}{n_0+n}\right)^2}{\frac{\sigma^2}{n_0+n}}\right]$$

Combining prior and likelihood

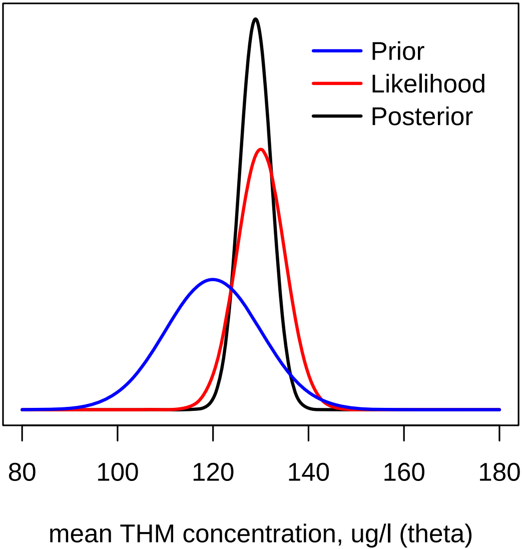
Combining the Normal likelihood and the Normal prior gives the following posterior distribution

$$\begin{aligned}\theta \mid \mathbf{y} &\sim \mathrm{N}\left(\frac{n_0\mu + n\overline{y}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)\\ &\sim \mathrm{N}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \frac{5^2}{0.25 + 2}\right)\\ &= \mathrm{N}(128.9, \quad 3.33^2)\end{aligned}$$

This gives 95% posterior interval for θ of 122.4 – 135.4 $\frac{\mu g}{l}$

- Posterior mean $\frac{(n_0\mu + n\overline{y})}{(n_0+n)}$ is a weighted average of the prior mean μ and parameter estimate $\overline{y}$, weighted by their precisions (relative 'sample sizes') $\rightsquigarrow$ compromise between the two
- Posterior variance is based on an implicit sample size equivalent to the sum of the prior 'sample size' n_0 and the sample size of the data n
- As $n \to \infty$, $p(\theta \mid \mathbf{y}) \to \mathrm{N}(\overline{y}, \frac{\sigma^2}{n})$ which does not depend on the prior

Prior, likelihood and posterior for THM example



Bayesian inference using count data

Suppose we have an independent sample of counts $y_1, \ldots, y_n$ which can be assumed to follow a Poisson distribution with unknown mean μ:

$$p(\mathbf{y} \mid \mu) = \prod_i \frac{\mu^{y_i} e^{-\mu}}{y_i!}$$

The conjugate prior for the mean of a Poisson distribution is a Gamma distribution:

$$p(\mu) = \text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$$

The Gamma distribution

Flexible distribution for positive quantities

If $\mu \sim \mathrm{Gamma}[a, b]$

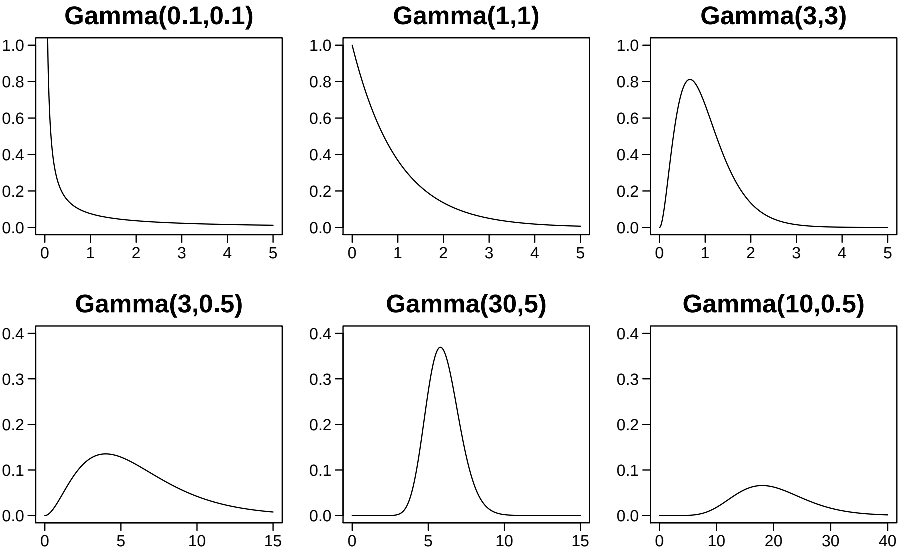
$$\begin{aligned} p(\mu \mid a, b) &= \frac{b^a}{\Gamma(a)} \mu^{a-1}\, e^{-b\mu}, \quad \mu \in (0, \infty) \\ \mathrm{E}(\mu \mid a, b) &= \frac{a}{b} \\ \mathrm{V}(\mu \mid a, b) &= \frac{a}{b^2} \end{aligned}$$

OpenBUGS notation: `mu ~ dgamma(a,b)`

The Gamma distribution [continued]

- Gamma[1,b] distribution is exponential with mean $\frac{1}{b}$
- Gamma[$\frac{v}{2}, \frac{1}{2}$] is a Chi-squared χ^2_v distribution on v degrees of freedom
- $\mu \sim$ Gamma[0,0] means that $p(\mu) \propto \frac{1}{\mu}$, or that $\log \mu \sim$ Uniform
- Also used as conjugate prior distribution for inverse variances (precisions) of Normal distributions
- Can also be used as sampling distribution for skewed positive valued quantities (alternative to log normal likelihood)

Shape of the Gamma distribution



Combining likelihood and prior

This implies the following posterior

$$\begin{aligned}
p(\mu \mid \mathbf{y}) &\propto p(\mu)\, p(\mathbf{y} \mid \mu) \\
&= \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu} \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{y_i}}{y_i!} \\
&\propto \mu^{a+n\overline{y}-1} e^{-(b+n)\mu} \\
&= \mathrm{Gamma}(a+n\overline{y},\, b+n)
\end{aligned}$$

The posterior is another (different) Gamma distribution

$$E(\mu \mid \mathbf{y}) = \frac{a+n\overline{y}}{b+n} = \overline{y}\left(\frac{n}{n+b}\right) + \frac{a}{b}\left(1 - \frac{n}{n+b}\right)$$

So posterior mean is a compromise between the prior mean $\frac{a}{b}$ and the sample mean ($\overline{y}$)

Example: Estimation of disease risk in a single area

Often interested in estimating the rate or relative risk rather than the mean for Poisson data:

- Suppose we observe $y = 5$ cases of leukaemia in one region, with age-sex-standardised expected number of cases E = 2.8
- Assume Poisson likelihood for y with mean $\mu = \lambda \times E$, where λ is the unknown relative risk:

$$p(y \mid \lambda, E) = \frac{(\lambda E)^y e^{-\lambda E}}{y!}$$

- Assume Gamma(a, b) prior for the relative risk λ:

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

- Posterior for λ is then

$$\begin{aligned} p(\lambda \mid y, E) &\propto \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \frac{(\lambda E)^y e^{-\lambda E}}{y!} \\ &\propto \lambda^{a+y-1} e^{-(b+E)\lambda} = \text{Gamma}(a+y, b+E) \end{aligned}$$

Class exercise: Comparing priors

1. Suppose we wish to express vague prior information about λ
    - A Gamma(0.1, 0.1) distribution represents a prior for the relative risk λ
    - What are the prior mean and variance?
    - What are the parameters of the posterior distribution? What is the posterior mean?
2. Alternatively, we may have strong prior information in the form of a Gamma(48,40)
    - What are the prior mean and variance?
    - What are the parameters of the posterior distribution? What is the posterior mean?

Compare the two priors and the two posteriors. What can you conclude?

Summary

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the sample mean
- the posterior standard deviation is less than each of the prior standard deviation and the standard error

> *'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule' (Senn, 1997)*

As $n \to \infty$,

- the posterior mean $\to$ the sample mean
- the posterior standard deviation $\to$ the standard error
- the posterior does not depend on the prior

Summary [continued]

- When the posterior is in the same family as the prior then we have what is known as conjugacy
- This has the advantage that prior parameters can usually be interpreted as a prior sample
- Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive
- Computations for non-conjugate priors are harder, but possible using MCMC (see Lecture 4)