Introduction to Bayesian Data Analysis
Analytical Bayesian Analysis (Chapter 2 of book)

Prof. Dr. Shravan Vasishth
Professor, Linguistics
Cognitive Science / Linguistics, Uni Potsdam, Germany

Conditional probability and Bayes' rule

A and B are discrete events. Conditional probability is defined as follows:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ where } P(B) > 0 \tag{1}$$

This means that $P(A,B) = P(A|B)P(B)$.
Since $P(B,A) = P(A,B)$, we can write:

$$P(B,A) = P(B|A)P(A) = P(A|B)P(B) = P(A,B). \tag{2}$$

Rearranging terms:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3}$$

This is Bayes' rule.

Bayes' rule: PDFs

When looking at probability distributions, we will encounter the rule in the following form. y is a vector of (iid) data points.

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)} \tag{4}$$

Here, $f(\cdot)$ refers to a probability density function, not the probability of a single event.

- The parameter θ is now a random variable: $f(\theta)$! A radical move!
- $f(y)$ is the "normalizing constant" we saw earlier, which makes the left-hand side a probability distribution.

$$f(y) = \int f(y, \theta)\, d\theta = \int f(y \mid \theta) f(\theta)\, d\theta \tag{5}$$

Bayes' rule: The normalizing constant

If θ is a discrete random variable and the support is $\{\theta_1, \dots, \theta_n\}$, then

$$f(y) = \sum_{i=1}^{n} f(y \mid \theta_i) P(\theta = \theta_i) \tag{6}$$

This is called **integrating out a parameter**. In continuous space, this would be:

$$f(y) = \int f(y \mid \theta) f(\theta)\, d\theta \tag{7}$$

A simple example will help to clarify this!

Integrating out a parameter: Concrete example

Consider the discrete Binomial case; $n = 10$ trials and $k = 7$ successes. The likelihood function then is

$$p(k = 7, n = 10|\theta) = \binom{10}{7}\theta^7(1-\theta)^3 \tag{8}$$

(Note: I generally write $p(\cdot)$ for PMFs, and $f(\cdot)$ for PDFs)

- Suppose that there are three possible values of θ, call them $\theta_1 = 0.1$, $\theta_2 = 0.5$, and $\theta_3 = 0.9$.
- Each has probability $1/3$; so $p(\theta_1) = p(\theta_2) = p(\theta_3) = 1/3$

Integrating out a parameter: Concrete example

Here, we are "integrating" out (in discrete space!) the parameter θ to compute something called the **marginal likelihood**:

$$\begin{aligned} p(k=7,n=10) = & \binom{10}{7}\theta_1^7(1-\theta_1)^3 \times p(\theta_1) \\ & + \binom{10}{7}\theta_2^7(1-\theta_2)^3 \times p(\theta_2) \\ & + \binom{10}{7}\theta_3^7(1-\theta_3)^3 \times p(\theta_3) \end{aligned} \tag{9}$$

Integrating out a parameter: Concrete example

Writing the θ values and their probabilities, we get:

$$
\begin{aligned}
p(k=7, n=10) = & \binom{10}{7} 0.1^7 (1-0.1)^3 \times \frac{1}{3} \\
& + \binom{10}{7} 0.5^7 (1-0.5)^3 \times \frac{1}{3} \\
& + \binom{10}{7} 0.9^7 (1-0.9)^3 \times \frac{1}{3}
\end{aligned} \tag{10}
$$

Integrating out a parameter: Concrete example

$$
\begin{aligned}
p(k=7, n=10) =& \frac{1}{3}[\binom{10}{7} 0.1^7 (1-0.1)^3 \\
&+ \binom{10}{7} 0.5^7 (1-0.5)^3 \\
&+ \binom{10}{7} 0.9^7 (1-0.9)^3] \\
=& 0.0581973
\end{aligned} \quad (11)
$$

Integrating out a parameter: Concrete example

Using R:

```
sum(dbinom(7,size=10,prob=c(0.1,0.5,0.9)))/3

## [1] 0.05819729
```

Computing this marginal likelihood involves "integrating out a parameter"; it's a kind of weighted sum of the likelihood, weighted by the possible values of the parameter.

Bayes' rule: The normalizing constant

Without the normalizing constant, we have the relationship:

$$f(\theta \mid y) \propto f(y \mid \theta) f(\theta) \tag{12}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{13}$$

The next step: Computing the posterior analytically

Next, we will use Bayes' rule in a practical example.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The likelihood function (in this discrete case only!) will tell us Prob$(x \mid n, \theta)$ given some specific value for θ, here 0.5:

```
dbinom(x=46, size=100, prob=0.5)

## [1] 0.0579584
```

Note that we can ignore the normalizing constant $\binom{x}{n}$, and write:

$$f(x \mid n, \theta) \propto \theta^{46}(1-\theta)^{54} \tag{14}$$

So, to get the posterior distribution for θ, we just need to work out a prior distribution for θ, call it $f(\theta)$.

$$f(\theta \mid x) \propto f(x \mid n, \theta) f(\theta) \tag{15}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

- For the prior distribution of θ, we need a distribution that can represent our uncertainty about the probability θ of success.
- The beta distribution is commonly used as prior for proportions.
- We say that the beta distribution is conjugate to the binomial density; i.e., the two densities have similar functional forms.

The beta PDF (**using θ as a random variable here!**) is

$$f(\theta) = \begin{cases} \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

In R, we write $\theta \sim \texttt{beta}(\texttt{shape1} = a, \ \texttt{shape2} = b)$. The associated R function is $\texttt{dbeta}(\texttt{x}, \texttt{shape1}, \texttt{shape2})$.
The mean and variance are

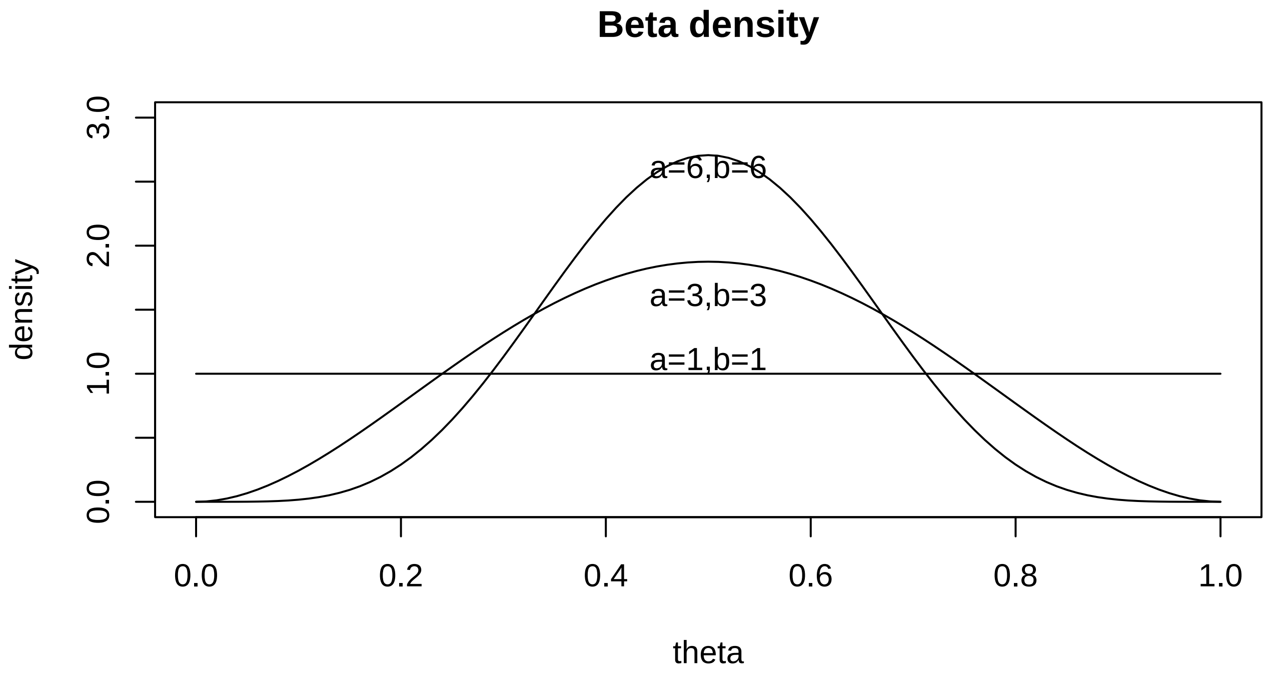
$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2\,(a+b+1)}. \tag{16}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

- The beta distribution's parameters a and b can be interpreted as (our beliefs about) prior successes and failures.
- Once we choose values for a and b, we can plot the beta PDF.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

Here, we show the beta PDF for three sets of values of a,b.



Example 1: Binomial Likelihood, Beta prior, Beta posterior

What does $\theta \sim Beta(a, b)$ mean in practical terms?

- If we don't have much prior information, we could use a=b=1; this gives us a uniform prior; this is sometimes called an **uninformative prior**.
- If we have a lot of prior knowledge and/or a strong belief that θ has a particular range of values, we can use a larger a,b to reflect our greater certainty about the parameter.
- Notice that the larger our parameters a and b, the narrower the spread of the distribution; this makes sense because a larger sample size (a greater number of successes a, and a greater number of failures b) will lead to more precise estimates.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

Just for the sake of argument, let's take four different beta priors, each reflecting increasing certainty.

1. Beta(a=2,b=2)
2. Beta(a=3,b=3)
3. Beta(a=6,b=6)
4. Beta(a=21,b=21)

Each reflects a belief that $\theta = 0.5$, with varying degrees of (un)certainty. Now we just need to plug in the likelihood and the prior.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

$$f(\theta \mid x) \propto f(x \mid \theta) f(\theta) \tag{17}$$

The four corresponding posterior distributions would be:

$$f(\theta \mid x) \propto [\theta^{46}(1-\theta)^{54}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{48-1}(1-\theta)^{56-1} \tag{18}$$

$$f(\theta \mid x) \propto [\theta^{46}(1-\theta)^{54}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{49-1}(1-\theta)^{57-1} \tag{19}$$

$$f(\theta \mid x) \propto [\theta^{46}(1-\theta)^{54}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{52-1}(1-\theta)^{60-1} \tag{20}$$

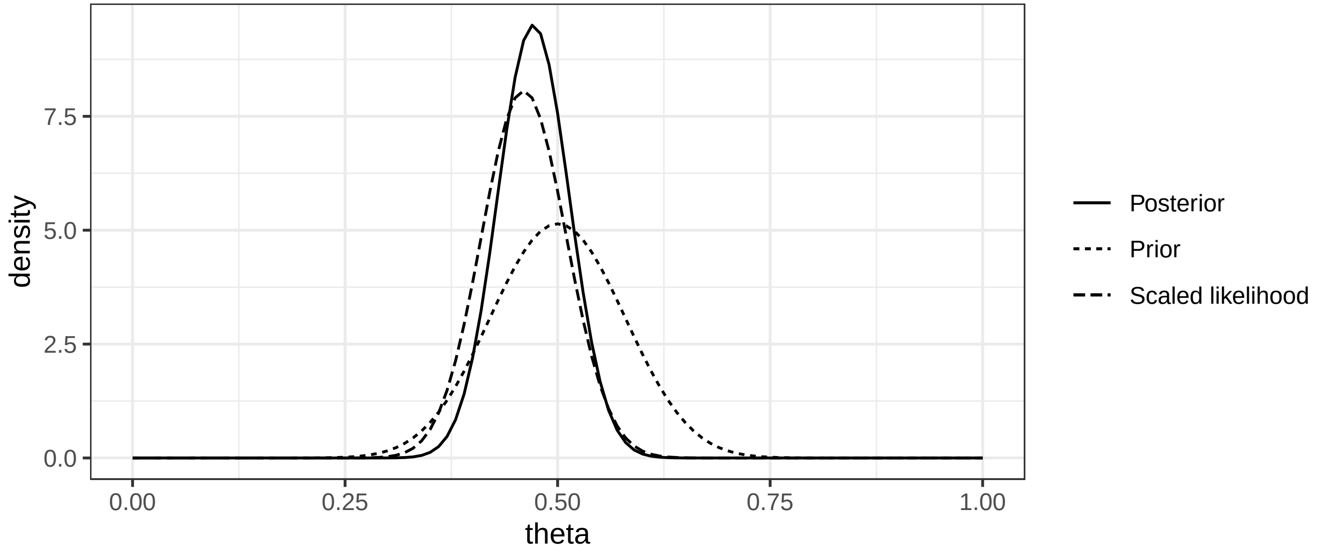
$$f(\theta \mid x) \propto [\theta^{46}(1-\theta)^{54}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{67-1}(1-\theta)^{75-1} \tag{21}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

- We can now visualize each of these triplets of priors, likelihoods and posteriors. I use Beta(21,21) as a prior here.
- Note that I normalize the likelihood because this allows me to visualize all three (prior, likelihood, posterior) in the same plot on the same scale.

```
x <- 46
n <- 100
## Prior specification:
a <- 21
b <- 21
binom_lh <- function(theta) {
dbinom(x=x, size =n, prob = theta)
}
## normalizing constant:
K <- 1/integrate(f = binom_lh, lower = 0, upper = 1)$value
binom_scaled_lh <- function(theta) K * binom_lh(theta)
```

Example 1: Binomial Likelihood, Beta prior, Beta posterior



Summary

- We saw how we can derive the posterior distribution given data.
- The posterior belongs to the same family of functions as the prior---this is called the conjugate case.
- Everything else we do from this point on is to derive the posterior given a likelihood and a prior for the parameters in the likelihood:

$$f(\theta \mid x) \propto f(x \mid \theta) f(\theta) \quad (22)$$

  θ can be a single parameter, or a vector of parameters.

Next: another example of a conjugate analysis (Poisson-Gamma).

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Suppose we are modeling the total number of regressions (leftward eye movements) per word in an eyetracking study (data from Vasishth et al., 2011):

```
dat<-read.table(file="data/TRCexample.txt",header=TRUE)
head(dat,n=3)
```

```
##      subject condition item variable value
## 8425       1         d   16      TRC     0
## 8426       1         d   16      TRC     0
## 8427       1         d   16      TRC     4
```

> Source:
> Shravan Vasishth, Katja Suckow, Richard L. Lewis, and Sabine Kern. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures. Language and Cognitive Processes, 25:533--567, 2011

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

```
summary(dat$value)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   1.000   1.165   2.000  21.000    2158
```

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

- The number of times x that regressions occurred from a word can be modeled by a Poisson distribution:
- The Poisson distribution (discrete) has one parameter (the rate):

$$f(x \mid \lambda) = \frac{\exp(-\lambda)\lambda^x}{x!} \tag{23}$$

- The rate (the mean no. of regressions per word) $\lambda > 0$ is unknown
- $x \geq 0$ (a vector): the observed numbers of regressions per word are independent given λ

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

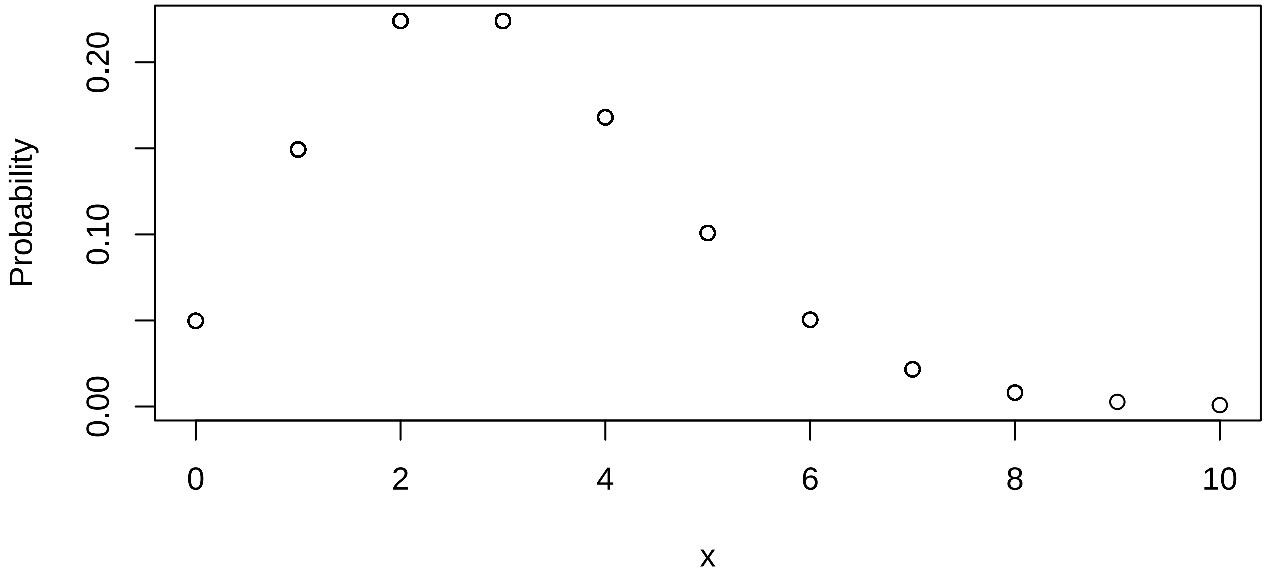
Simulated data (n=10, number of data points):

```
(x<-rpois(n=10,lambda=3))

##  [1] 2 2 4 0 3 5 3 6 3 4
```

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Visualization with $\lambda = 3$:



Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

- Suppose that prior research (or expert knowledge) suggests that the prior mean of λ is 3 and prior variance for λ is 1.5.
- The first step is to define a PDF for λ; this will reflect our prior belief, before seeing any new data.
- One good choice (but not the only possible choice!) is the gamma(a,b) distribution.

Note: I will talk about the choice of prior in Bayesian analyses later.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

The gamma PDF (continuous) for some variable x (parameters $a, b > 0$):

$$f(x \mid a, b) = \begin{cases} \frac{b^a \exp(-bx) x^{a-1}}{\Gamma(a)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{24}$$

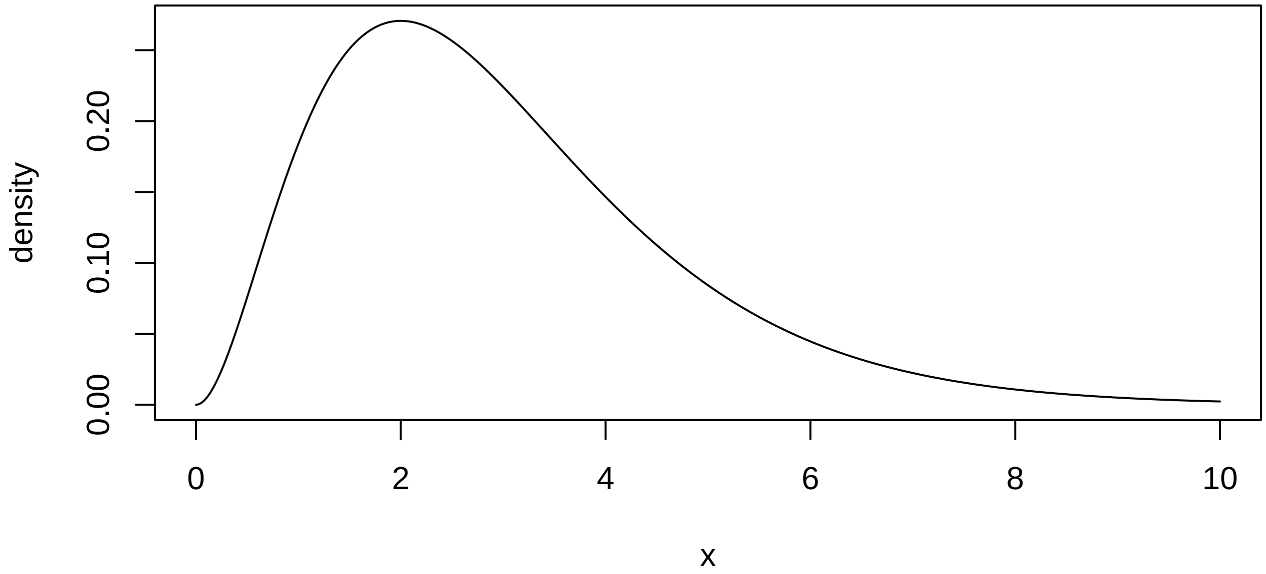
Here, $\Gamma(a) = (a-1)!$ for integer values of a. $\frac{b^a}{\Gamma(a)}$ is the normalizing constant.
In R, the a,b parameters are called shape and rate, respectively.
Simulated data from Gamma(a=3,b=1):

```
round(rgamma(n=10,shape=3,rate=1),2)

##  [1] 2.52 2.59 3.93 1.08 0.96 2.22 2.71 1.91 5.52 2.81
```

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Visualize the gamma PDF with a=3,b=1:



Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

In order to decide on the prior:

$$\lambda \sim Gamma(a, b)$$

we first need to figure out the parameters for a gamma density prior.

Key question: What should the parameters a,b be? We know that

- In a gamma PDF with parameters a, b, the mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$
- Suppose we know that the mean and variance of λ from prior research is 3 and 1.5
- Solve for a,b, which gives us the parameters we need for the gamma prior on λ.

Example 2: Poisson likelihood, gamma prior, gamma posterior

$$\frac{a}{b} = 3 \tag{25}$$

$$\frac{a}{b^2} = 1.5 \tag{26}$$

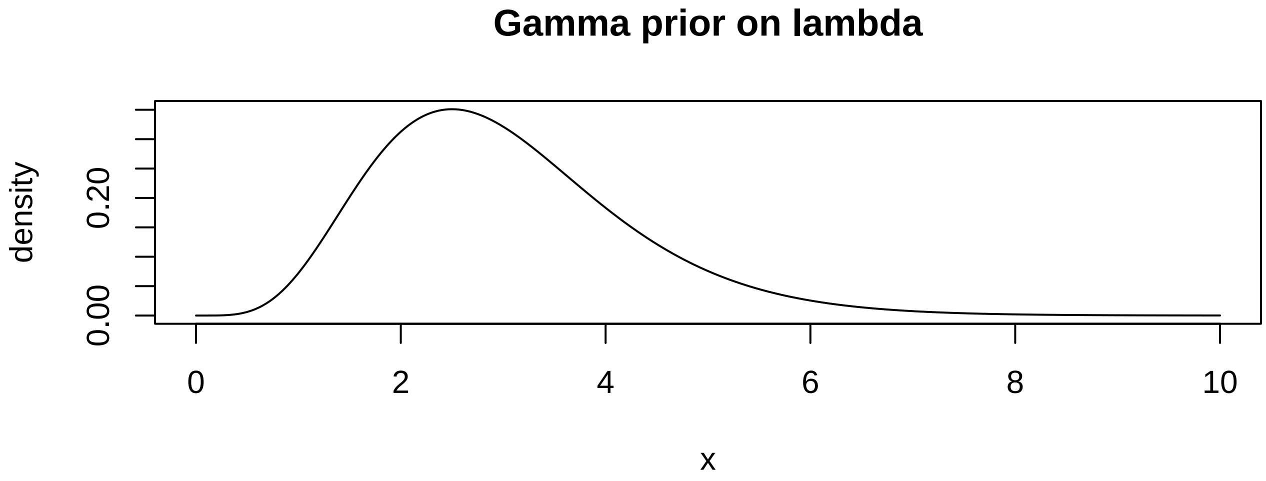
Just solve for a and b (exercise).

Result: $a = 6, b = 2$.

Example 2: Poisson likelihood, gamma prior, gamma posterior

The prior on λ is:

$$\lambda \sim Gamma(a = 6, b = 2) \tag{27}$$



Example 2: Poisson likelihood, gamma prior, gamma posterior

Cross-check using Monte Carlo simulations that the mean and variance are as they should be:

```
lambda<-rgamma(10000,shape=6,rate=2)

round(mean(lambda),1)

## [1] 3

round(var(lambda),1)

## [1] 1.5
```

Example 2: Poisson likelihood, gamma prior, gamma posterior

Given that

$$\text{Posterior} \propto \text{Likelihood Prior} \tag{28}$$

and given that the PDF we assume for the data is Poisson (n **independent** data points $\mathbf{x}$):

$$\mathbf{x} =< x_1, \dots, x_n >$$

$$\begin{aligned} f(\mathbf{x} \mid \lambda) =& \frac{\exp(-\lambda)\lambda^{x_1}}{x_1!} \times \dots \times \frac{\exp(-\lambda)\lambda^{x_n}}{x_n!} \\ =& \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \\ =& \frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{i=1}^{n} x_i!} \end{aligned} \tag{29}$$

Example 2: Poisson likelihood, gamma prior, gamma posterior

Computing the posterior is surprisingly easy now:

$$\text{Posterior} = \underset{\substack{\uparrow \\ Likelihood}}{\left[\frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}!}\right]} \underset{\substack{\uparrow \\ Prior}}{\left[\frac{\mathbf{b^a}\lambda^{a-1}\exp(-b\lambda)}{\mathbf{\Gamma(a)}}\right]} \tag{30}$$

The terms $x!, \Gamma(a), b^a$ do not involve λ and make up the normalizing constants; we can drop these.

This gives us the posterior **up to proportionality**:

$$\begin{aligned}\text{Posterior} &\propto \exp(-n\lambda)\lambda^{\sum_i^n x_i}\lambda^{a-1}\exp(-b\lambda) \\ &= \lambda^{a-1+\sum_i^n x_i}\exp(-\lambda(b+n))\end{aligned} \tag{31}$$

Example 2: Poisson likelihood, gamma prior, gamma posterior

$$\begin{aligned} \text{Posterior} &\propto \exp(-n\lambda)\lambda^{\sum_i^n x_i}\lambda^{a-1}\exp(-b\lambda) \\ &= \lambda^{a-1+\sum_i^n x_i}\exp(-\lambda(b+n)) \end{aligned} \tag{32}$$

- First, note that the gamma distribution in general is $Gamma(a,b) \propto \lambda^{a-1}\exp(-\lambda b)$.
- So it's enough to state the above as a gamma distribution with some updated parameters a*, b*.

If we equate $a^* - 1 = a - 1 + \sum_i^n x_i$ and $b^* = b + n$, we can rewrite the above as:

$$\lambda^{a^*-1}\exp(-\lambda b^*) \tag{33}$$

Example 2: Poisson likelihood, gamma prior, gamma posterior

- This means that $a^* = a + \sum_i^n x_i$ and $b^* = b + n$.
- We can find a constant k such that the above is a proper probability density function, i.e.:

$$k \int_0^\infty \lambda^{a^*-1} \exp(-\lambda b^*) = 1 \qquad (34)$$

- Thus, the posterior has the form of a gamma distribution with parameters $a^* = a + \sum_i^n x_i, b^* = b + n$. Hence the Gamma distribution is a conjugate prior for the Poisson.

Example 2: Poisson likelihood, gamma prior, gamma posterior

Concrete example given data

- Suppose the regressive eye movements from one subject on n=5 words is: $2, 4, 3, 6, 1$.
- The prior we chose was Gamma(a=6,b=2).
- $\sum_i^n x_i = 16$ and sample size $n = 5$.

It follows that the posterior is

$$\begin{aligned} Gamma(a^* = a + \sum_i^n x_i, b^* = b + n) =& Gamma(6 + 16, 2 + 5) \\ =& Gamma(22, 7) \end{aligned} \quad (35)$$

Example 2: Poisson likelihood, gamma prior, gamma posterior

- The mean of the posterior is $\frac{a*}{b*} = \frac{22}{7} = 3.14$
- The variance is $\frac{a*}{b*^2} = \frac{22}{7^2} = 0.45$

Stepping back, and summary

- We saw two examples of conjugate analyses: the binomial-beta and the Poisson-gamma.
- In each example, we derived the posterior given a likelihood and a prior.

Next: the posterior's mean is a weighted mean of the MLE and the prior mean.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

We can express the posterior mean as a weighted sum of the prior mean and the maximum likelihood estimate of λ.
The posterior mean is:

$$\frac{a*}{b*} = \frac{a + \sum x_i}{n + b} \tag{36}$$

This can be rewritten as

$$\frac{a*}{b*} = \frac{a + n\bar{x}}{n + b} \tag{37}$$

Dividing both the numerator and denominator by b:

$$\frac{a*}{b*} = \frac{(a + n\bar{x})/b}{(n + b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \tag{38}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Since a/b is the mean m of the prior, we can rewrite this as:

$$\frac{a/b + n\bar{x}/b}{1+n/b} = \frac{m + \frac{n}{b}\bar{x}}{1+\frac{n}{b}} \tag{39}$$

We can rewrite this as:

$$\frac{m + \frac{n}{b}\bar{x}}{1+\frac{n}{b}} = \frac{m \times 1}{1+\frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1+\frac{n}{b}} \tag{40}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This is a weighted average: setting $w_1 = 1$ and $w_2 = \frac{n}{b}$, we can write the above as:

$$m\frac{w_1}{w_1 + w_2} + \bar{x}\frac{w_2}{w_1 + w_2} \tag{41}$$

A n approaches infinity, the weight on the prior mean m will tend towards 0, making the posterior mean approach the maximum likelihood estimate of the sample.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

In general, in a Bayesian analysis, as sample size increases, the likelihood will dominate in determining the posterior mean.
Regarding variance, since the variance of the posterior is:

$$\frac{a*}{b*^2} = \frac{(a + n\bar{x})}{(n + b)^2} \tag{42}$$

as n approaches infinity, the posterior variance will approach zero: more data will reduce variance (uncertainty).

Stepping back

- We saw two examples where we can do the computations to derive the posterior using simple algebra.
- There are several other such simple cases.
- **A big insight**: the posterior mean is a compromise between the prior mean and the sample mean.
- When data are sparse, the prior will dominate in determining the posterior mean.
- When a lot of data are available, the MLE will dominate in determining the posterior mean.
- Given sparse data, informative priors based on expert knowledge, existing data, or meta-analysis will play an important role.

The next steps: Realistic data analysis

- In realistic data analysis settings, we can't use these simple conjugate analyses
- For such cases, we need to use MCMC (Markov chain Monte Carlo) sampling techniques so that we can sample from the posterior distributions of the parameters.

Some sampling approaches are:

- Gibbs sampling using inversion sampling
- Metropolis-Hastings
- Hamiltonian Monte Carlo

See this book for a good overview:

Lambert, B. (2018). A student's guide to Bayesian statistics. Sage.

The next steps: Realistic data analysis

Next topic: Sampling algorithms.