Chapter 2: Introduction to Bayesian Data Analysis

Shravan Vasishth (vasishth.github.io)

June 2025

Contents

Textbook

Introduction to Bayesian Data Analysis for Cognitive Science

Nicenboim, Schad, Vasishth

- Online version: https://bruno.nicenboim.me/bayescogsci/
- Source code: https://github.com/bnicenboim/bayescogsci
- Physical book: here

Be sure to read the textbook's chapter 2 in addition to watching this lecture.

Bayes' rule

Bayes' rule: When A and B are observable discrete events (such as "it has been raining" or "the streets are wet"), we can state the rule as follows:

Bayes' rule in terms of discrete events:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

Bayes' rule in terms of probability distributions:

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\Theta}) \times p(\boldsymbol{\Theta})}{p(\boldsymbol{y})} \tag{2}$$

The above statement can be rewritten in words as follows:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}} \tag{3}$$

The terms here have the following meaning. We elaborate on each point with an example below.

- The *Posterior*, $p(\boldsymbol{\Theta}|\boldsymbol{y})$, is the probability distribution of the parameter(s) conditional on the data.
- The *Likelihood*, $p(\boldsymbol{y}|\boldsymbol{\Theta})$ is as described in chapter 1: it is the PMF (discrete case) or the PDF (continuous case) expressed as a function of $\boldsymbol{\Theta}$.
- The *Prior*, $p(\boldsymbol{\Theta})$, is the initial probability distribution of the parameter(s), before seeing the data.
- The *Marginal Likelihood*, $p(\boldsymbol{y})$, was introduced in chapter 1 and standardizes the posterior distribution to ensure that the area under the curve of the distribution sums to 1, that is, it ensures that the posterior is a valid probability distribution.

Two examples will clarify all these terms.

An analytical example (Beta-Binomial)

Consider the following experiment. Subjects are shown the following sentence just once. The number of trials is therefore 100.

"It's raining. I'm going to take the . . . "

- Assume that we have 100 subjects
- 80 subjects complete the sentence with "umbrella"
- The estimated cloze probability or predictability (given the preceding context)

would be $\hat{\theta} = \frac{80}{100} = 0.8$.

The variability of the estimated probability under repeated sampling (100 experiments), given different sample sizes (number of trials, or size). First, consider a 100 experiments with 10 independent trials each.

```
estimated_means <- rbinom(100,
                          size = 10,
                          prob = 0.8) / 10
sd(estimated_means)
```

```
## [1] 0.127426
```

Compare with a much larger number of trials (100):

```
estimated_means <- rbinom(100,
                          size = 100,
                          prob = 0.8) / 100
sd(estimated_means)
```

```
## [1] 0.03568592
```

The repeated runs of the (simulated) experiment are giving us different point estimates of θ, and the variability in the estimate of θ under repeated sampling depends on the sample size.

What if we treat θ as a random variable? Suppose that $\theta \sim Uniform(0, 1)$.

Now, if we were to run our simulated experiments again and again, there would be *two* sources of variability in the estimate of the parameter: the data as well as the uncertainty associated with θ.

```r
theta <- runif(100, min = 0, max = 1)
estimated_means <- rbinom(100,
                          size = 10,
                          prob = theta) / 10
sd(estimated_means)
```

```
## [1] 0.2879745
```

The higher standard deviation is now coming from the uncertainty associated with the θ parameter.

Suppose $\theta \sim Uniform(0.3, 0.8)$. Now, the variability in the estimate is much smaller:

```r
theta <- runif(100, min = 0.3, max = 0.8)
estimated_means <- rbinom(n=100,
                          size = 10,
                          prob = theta) / 10
sd(estimated_means)
```

```
## [1] 0.2148502
```

In other words, the greater the prior uncertainty associated with the parameter θ, the greater the variability in the estimated means from the data.

- Frequentists: the true, unknown value of θ is some point value.
- Bayesians: θ is a random variable with a probability density/mass function associated with it. This PDF is called a prior distribution, and represents our prior belief or prior knowledge about possible values of this parameter.
- Once we obtain data, these data serve to modify our prior belief about this distribu-

tion; this updated probability density function of the parameter is called the posterior distribution.

Choosing a likelihood

Under the assumptions we have set up above, the responses follow a binomial distribution, and so the PMF can be written as follows.

$$p(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{4}$$

- k indicates the number of times "umbrella" is given as an answer
- n is the number of trials.

If we collect 100 data points and it turns out that $k = 80$, these data are now a fixed quantity. The only variable now in the PMF above is θ:

$$p(k=80|n=100,\theta) = \binom{100}{80}\theta^{80}(1-\theta)^{20} \tag{5}$$

The above function is a now a continuous function of the value θ, which has possible values ranging from 0 to 1. Compare this to the PMF of the binomial, which treats θ as a fixed value and defines a discrete distribution over the $n+1$ possible discrete values k that we can observe (the possible number of successes).

We can now rewrite the PMF as a likelihood function:

$$\mathcal{L}(\theta|k=80,n=100) = \binom{100}{80}\theta^{80}(1-\theta)^{20} \quad (6)$$

We now find out, using Bayes' rule, the posterior distribution of θ given our data: $p(\theta|n,k)$. Missing: a prior distribution over the parameter θ, which expresses our prior uncertainty about plausible values of θ.

Choosing a prior for θ

The parameter θ is a random variable that has a PDF whose range lies within [0,1], the range over which θ can vary (this is because θ represents a probability).

The beta distribution, which is a PDF for a continuous random variable, is commonly used as prior for parameters representing probabilities. One reason for this choice is that its PDF ranges over the interval $[0,1]$. The other reason for this choice is that it makes the Bayes' rule calculation remarkably easy.

The beta distribution has the following PDF.

$$p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \qquad (7)$$

The term $B(a,b)$ expands to $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta$, and is the normalizing constant; it ensures that the area under the curve sums to one.

The beta distribution's parameters a and b can be interpreted as expressing our prior beliefs about the probability of success; a represents

the number of "successes", in our case, answers that are "umbrella" and b the number of failures, the answers that are not "umbrella." The figure below shows the different beta distribution shapes given different values of a and b.

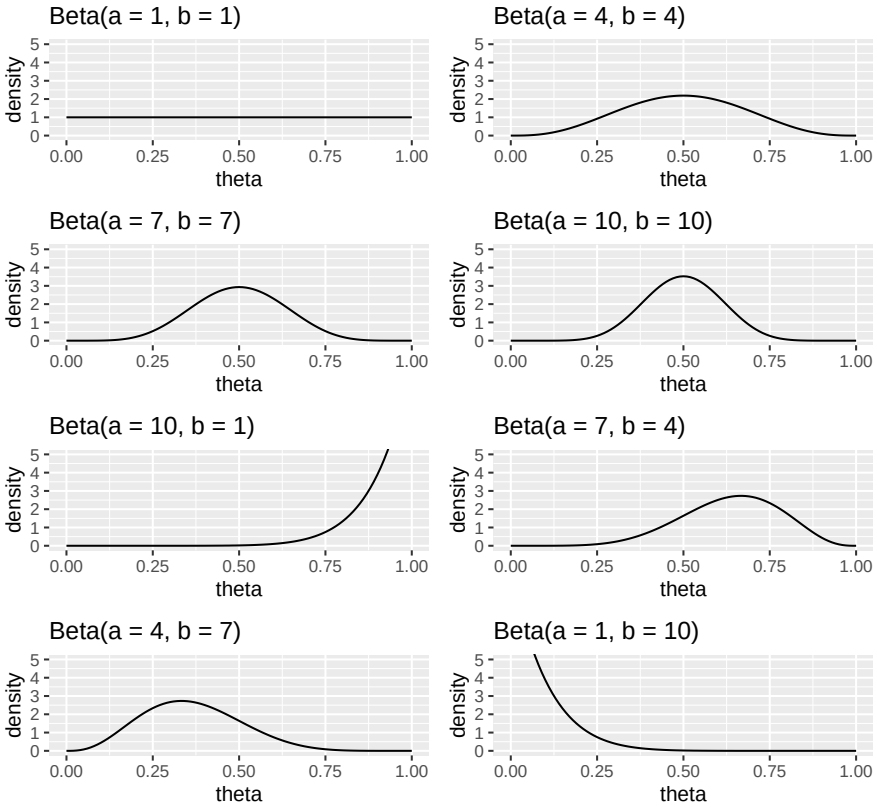


Figure 1: Examples of beta distributions with different parameters.

As in the binomial and normal distributions that we saw in chapter 1, one can analytically derive the formulas for the expectation and variance of the beta distribution. These are:

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \mathrm{Var}(X) = \frac{a \times b}{(a+b)^2(a+b+1)} \tag{8}$$

As an example, choosing $a = 4$ and $b = 4$ would mean that the answer "umbrella" is as likely as a different answer, but we are relatively unsure

about this. We could express our uncertainty by computing the region over which we are 95% certain that the value of the parameter lies; this is the *95% credible interval*. For this, we would use the `qbeta()` function in R; the parameters a and b are called `shape1` and `shape2` in R.

```
qbeta(c(0.025, 0.975),
      shape1 = 4,
      shape2 = 4)
```

```
## [1] 0.1840516 0.8159484
```

If we were to choose $a = 10$ and $b = 10$, we would still be assuming that a priori the answer "umbrella" is just as likely as some other answer, but now our prior uncertainty about this mean is lower, as the 95% credible interval computed below shows.

```
qbeta(c(0.025, 0.975),
      shape1 = 10,
      shape2 = 10)
```

```
## [1] 0.2886432 0.7113568
```

In the figure above, we can also see the difference in uncertainty in these two examples graphically.

Which prior should we choose? See the online chapter on priors.

For the moment, just for illustration, we choose the values $a = 4$ and $b = 4$ for the beta prior. Then, our prior for θ is the following beta PDF:

$$p(\theta) = \frac{1}{B(4,4)}\theta^3(1-\theta)^3 \tag{9}$$

Having chosen a likelihood, and having defined a prior on θ, we are ready to carry out our first Bayesian analysis to derive a posterior distribution for θ.

Using Bayes' rule to compute the posterior $p(\theta|n,k)$

Having specified the likelihood and the prior, we will now use Bayes' rule to calculate $p(\theta|n,k)$. Using Bayes' rule simply involves replacing the likelihood and the prior we defined above into the equation we saw earlier:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}} \tag{10}$$

Replace the terms for likelihood and prior into this equation:

$$p(\theta|n = 100, k = 80) \tag{11}$$

The above expands to:

$$\frac{\left[\binom{100}{80}\theta^{80} \times (1-\theta)^{20}\right] \times \left[\frac{1}{B(4,4)} \times \theta^3(1-\theta)^3\right]}{p(k = 80)}$$

where $p(k = 80)$ is $\int_0^1 p(k = 80|n = 100, \theta)p(\theta)\,\mathrm{d}\theta$. This term will be a constant once the number of successes k is known. In fact, once k is known, there are several constant values in the above equation; they are constants because none of them depend on the parameter of interest, θ. We can collect all of these together:

$$\left[\frac{\binom{100}{80}}{B(4,4) \times p(k=80)}\right] [\theta^{80}(1-\theta)^{20} \times \theta^{3}(1-\theta)^{3}]$$

The first term that is in square brackets, $\frac{\binom{100}{80}}{B(4,4)\times p(k=80)}$, is all the constants collected together, and is the normalizing constant we have seen before; it makes the posterior distribution $p(\theta|n=100,k=80)$ sum to one.

Since it is a constant, we can ignore it for now and focus on the two other terms in the equation. Because we are ignoring the constant, we will now say that the posterior is proportional to the right-hand side.

$$p(\theta|n=100,k=80) \propto [\theta^{80}(1-\theta)^{20}\times\theta^{3}(1-\theta)^{3}] \tag{12}$$

A common way of writing the above equation is:

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

Resolving the right-hand side now simply involves adding up the exponents! In this example, computing the posterior really does boil down to this simple addition operation on the exponents.

$$p(\theta|n=100,k=80) \propto [\theta^{80+3}(1-\theta)^{20+3}] = \theta^{83}(1-\theta)^{23} \tag{14}$$

The expression on the right-hand side corresponds to a beta distribution with parameters $a = 84$, and $b = 24$.

This becomes evident if we rewrite the right-hand side such that it represents the kernel of a beta PDF. All that is missing is a normalizing constant.

$$\theta^{83}(1-\theta)^{23} = \theta^{84-1}(1-\theta)^{24-1} \tag{15}$$

Without a normalizing constant, the area under the curve will not sum to one. Let's check this:

```
PostFun <- function(theta) {
  theta^84 * (1 - theta)^24
}
(AUC <- integrate(PostFun,
                  lower = 0,
                  upper = 1)$value)
```

```
## [1] 1.424179e-26
```

So the area under the curve (AUC) is not 1—the posterior that we computed above is not a proper probability distribution. What we have just done above is to compute the following integral:

$$\int_0^1 \theta^{84}(1-\theta)^{24} \tag{16}$$

We can use this integral to figure out what the normalizing constant is. Basically, we want to know what the constant k is such that the area under the curve sums to 1:

$$k \int_0^1 \theta^{84}(1-\theta)^{24} = 1 \tag{17}$$

We know what $\int_0^1 \theta^{84}(1-\theta)^{24}$ is; we just computed that value (called `AUC` in the R code

above). So, the normalizing constant is:

$$k = \frac{1}{\int_0^1 \theta^{84}(1-\theta)^{24}} = \frac{1}{AUC} \tag{18}$$

So, all that is needed to make the kernel $\theta^{84}(1-\theta)^{24}$ into a proper probability distribution is to include a normalizing constant, which, according to the definition of the beta distribution (equation, would be $B(84, 24)$. This term is in fact the integral we computed above.

So, what we have is the distribution of θ given the data, expressed as a PDF:

$$p(\theta|n = 100, k = 80) = \frac{1}{B(84, 24)}\theta^{84-1}(1-\theta)^{24-1} \tag{19}$$

Now, this function will sum to one:

```
PostFun <- function(theta) {
  theta^84 * (1 - theta)^24 / AUC
}
integrate(PostFun, lower = 0, upper = 1)$value
```

```
## [1] 1
```

Summary of the procedure

To summarize, we:

- started with data $(n = 100, k = 80)$
- assumed a binomial likelihood
- multiplied the likelihood function with the prior probability density function $\theta \sim Beta(4, 4)$

- obtained the posterior $p(\theta|n,k) = Beta(84,24)$

The constants were ignored when carrying out the multiplication; we say that we computed the posterior *up to proportionality*.

Finally, we showed how, in this simple example, the posterior can be rescaled to become a probability distribution, by including a proportionality constant.

The above example is a case of a *conjugate* analysis: the posterior on the parameter has the same form (belongs to the same family of probability distributions) as the prior. The above combination of likelihood and prior is called the beta-binomial conjugate case. There are several other such combinations of Likelihoods and Priors that yield a posterior that has a PDF that belongs to the same family as the PDF on the prior; some examples will appear in the exercises.

Formally, conjugacy is defined as follows: Given the likelihood $p(y|\theta)$, if the prior $p(\theta)$ results in a posterior $p(\theta|y)$ that has the same form as $p(\theta)$, then we call $p(\theta)$ a conjugate prior.

For the beta-binomial conjugate case, we can derive a very general relationship between the likelihood, prior, and posterior. Given the binomial likelihood up to proportionality (ignoring the constant) $\theta^k(1-\theta)^{n-k}$, and given the prior, also up to proportionality, $\theta^{a-1}(1-\theta)^{b-1}$, their product will be:

$$\theta^k(1-\theta)^{n-k}\theta^{a-1}(1-\theta)^{b-1} = \theta^{a+k-1}(1-\theta)^{b+n-k-1} \tag{20}$$

Thus, given a $Binomial(n, k|\theta)$ likelihood, and a $Beta(a, b)$ prior on θ, the posterior will be $Beta(a + k, b + n - k)$.

Visualizing the prior, likelihood, and posterior

We established in the example above that the posterior is a beta distribution with parameters $a = 84$, and $b = 24$. We visualize the likelihood, prior, and the posterior side by side in the figure below.

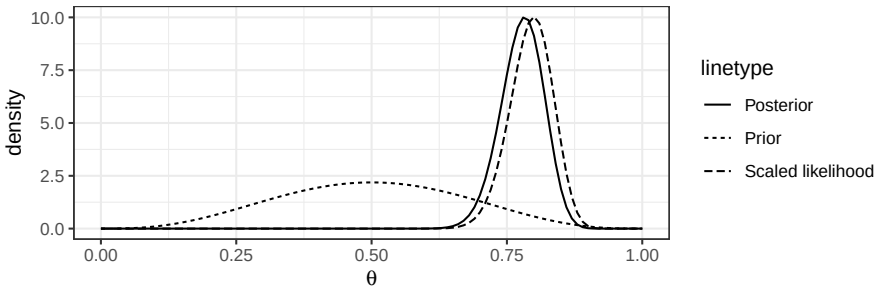


Figure 2: The (scaled) likelihood, prior, and posterior in the beta-binomial conjugate example. The likelihood is scaled to integrate to 1 to make it easier to compare to the prior and posterior distributions.

We can summarize the posterior distribution either graphically as we did above, or summarize it by computing the mean and the variance. The mean gives us an estimate of the cloze probability of producing "umbrella" in that sentence (given the model, i.e., given the likelihood and prior):

$$\mathrm{E}[\hat{\theta}] = \frac{84}{84 + 24} = 0.78 \tag{21}$$

$$\text{var}[\hat{\theta}] = \frac{84 \times 24}{(84+24)^2(84+24+1)} = 0.0016 \tag{22}$$

We could also display the 95% credible interval, the range over which we are 95% certain that θ lies, given the data and model.

```
qbeta(c(0.025, 0.975), shape1 = 84, shape2 = 24)
```

```
## [1] 0.6951283 0.8506904
```

Typically, we would summarize the results of a Bayesian analysis by displaying the posterior distribution of the parameter (or parameters) graphically, along with the above summary statistics: the mean, the standard deviation or variance, and the 95% credible interval. You will see many examples of such summaries later.

The posterior distribution is a compromise between the prior and the likelihood

Recall from the preceding sections that the a and b parameters in the beta distribution determine the shape of the prior distribution on the θ parameter. Just for the sake of illustration, let's take four different beta priors, which reflect increasing prior certainty about θ.

- $Beta(a = 2, b = 2)$
- $Beta(a = 3, b = 3)$
- $Beta(a = 6, b = 6)$
- $Beta(a = 21, b = 21)$

Each of these priors reflects a belief that $\theta = 0.5$, but with varying degrees of (un)certainty. Given

the general formula we developed above for the beta-binomial case, we just need to plug in the likelihood and the prior to get the posterior:

$$p(\theta|n,k) \propto p(k|n,\theta)p(\theta) \tag{23}$$

The four corresponding posterior distributions would be:

$$p(\theta \mid k,n) \propto [\theta^{80}(1-\theta)^{20}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{82-1}(1-\theta)^{22-1} \tag{24}$$

$$p(\theta \mid k,n) \propto [\theta^{80}(1-\theta)^{20}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{83-1}(1-\theta)^{23-1} \tag{25}$$

$$p(\theta \mid k,n) \propto [\theta^{80}(1-\theta)^{20}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{86-1}(1-\theta)^{26-1} \tag{26}$$

$$p(\theta \mid k,n) \propto [\theta^{80}(1-\theta)^{20}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{101-1}(1-\theta)^{41-1} \tag{27}$$

We can visualize each of these triplets of priors, likelihoods and posteriors; see the figure below.

Given some data and given a likelihood function, the tighter the prior, the greater the extent to which the posterior orients itself towards the prior. In general, we can say the following about the likelihood-prior-posterior relationship:

- The posterior distribution of a parameter is a compromise between the prior and the likelihood.

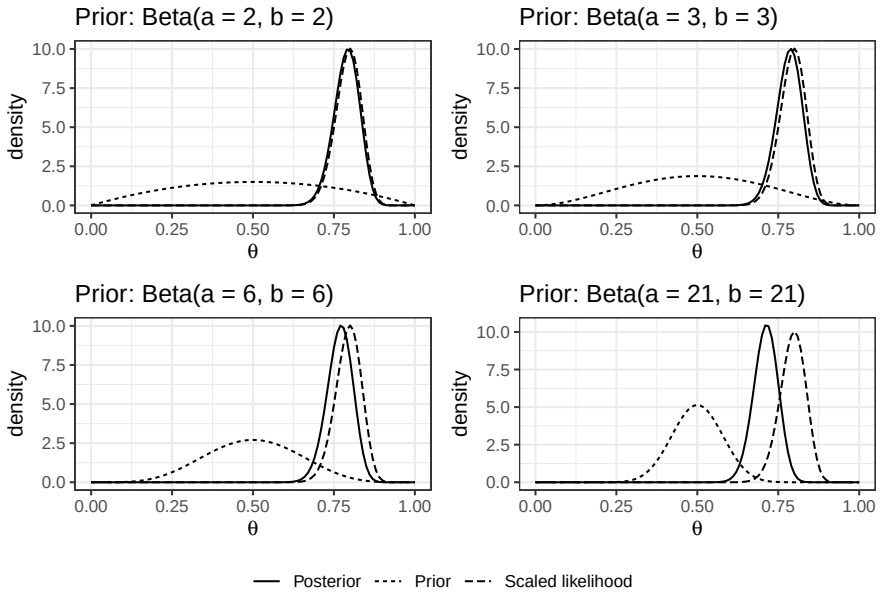


Figure 3: The (scaled) likelihood, prior, and posterior in the beta-binomial conjugate example, for different uncertainties in the prior. The likelihood is scaled to integrate to 1 to make its comparison easier.

- For a given set of data, the greater the certainty in the prior, the more heavily will the posterior be influenced by the prior mean.
- Conversely, for a given set of data, the greater the *un*certainty in the prior, the more heavily will the posterior be influenced by the likelihood.

Another important observation emerges if we increase the sample size (here, the number of trials) from 100 to, say, 1000000. Suppose we still get a sample mean of 0.8 here, so that $k = 800000$. Now, the posterior mean will be influenced almost entirely by the sample mean. This is because, in the general form for the posterior $Beta(a + k, b + n - k)$ that we computed above, the n and k become very large relative to the a, b values, and dominate in determining the posterior mean.

Whenever we do a Bayesian analysis, it is good practice to check whether the parameter you are interested in estimating is sensitive to the prior specification. Such an investigation is called a *sensitivity analysis*. Later in this book, we will see many examples of sensitivity analyses in realistic data-analysis settings.

Incremental knowledge gain using prior knowledge

Our posterior in the example with a $Beta(4, 4)$ prior was $Beta(84, 24)$.

We could now use this posterior as our prior for the next study. Suppose that we were to carry out a second experiment, again with 100 subjects, and this time 60 produced "umbrella." We could now use our new prior ($Beta(84, 24)$) to obtain an updated posterior. We have $a = 84, b = 24, n = 100, k = 60$. This gives us as posterior:

$$Beta(a+k, b+n-k) = Beta(84+60, 24+100-60) = Beta(144, 64)$$

Now, if we were to pool all our data that we have from the two experiments, then we would have as data $n = 200, k = 140$. Suppose that we keep our initial prior of $a = 4, b = 4$. Then, our posterior would be

$$Beta(4 + 140, 4 + 200 - 140) = Beta(144, 64)$$

This is exactly the same posterior that we got

when first analyzed the first 100 subjects' data, derived the posterior, and then used that posterior as a prior for the next 100 subjects' data.

This toy example illustrates an important point that has great practical importance for cognitive science:

One can incrementally gain information about a research question by using information from previous studies and deriving a posterior, and then use that posterior as a prior for the next study. This approach allows us to build on the information available from previous work.

A second analytical example (Poisson-Gamma)

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

```
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
```

- 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):

$$f(x \mid \lambda) = \frac{\exp(-\lambda)\lambda^x}{x!} \tag{28}$$

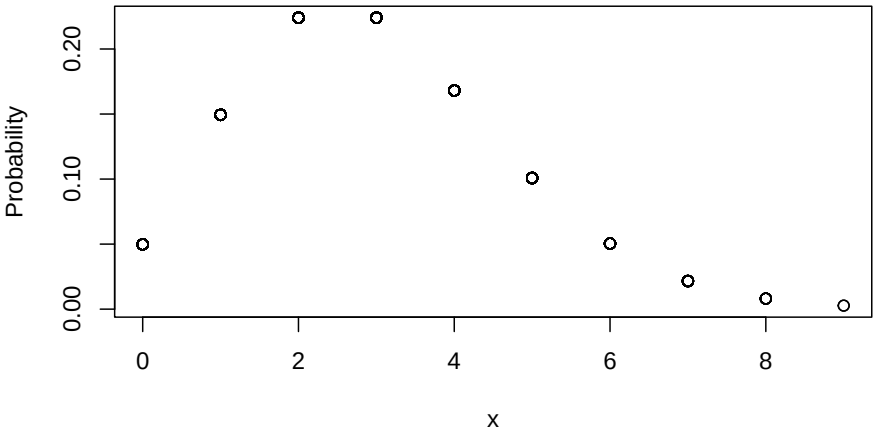
- The rate (the mean no. of regressions per word) $\lambda > 0$ is unknown - $x \geq 0$ (a vector): the observed numbers of regressions per word are assumed to be independent given λ (**Note: this assumption is incorrect here, as we have repeated measures from each subject; however, we ignore this detail for now**).

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

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

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

Visualization with $\lambda = 3$:



- 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.

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

$$f(x \mid a, b) = \begin{cases} \frac{b^a \exp(-bx)x^{a-1}}{\Gamma(a)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{29}$$

Here, $\Gamma(a) = (a-1)!$ for integer values of a. $\frac{b^a}{\Gamma(a)}$ is the normalizing constant.

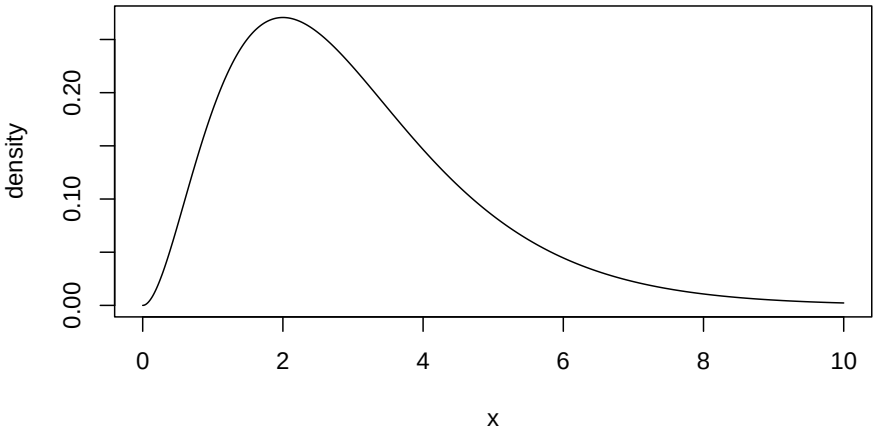
In R, the a,b parameters are called shape and rate, respectively.

Simulated data from Gamma(a=3,b=1):

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

```
##  [1] 2.24 3.87 4.67 1.31 1.26 3.99 2.57 3.21 3.09 3.25
```

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



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.

$$\frac{a}{b} = 3 \tag{30}$$

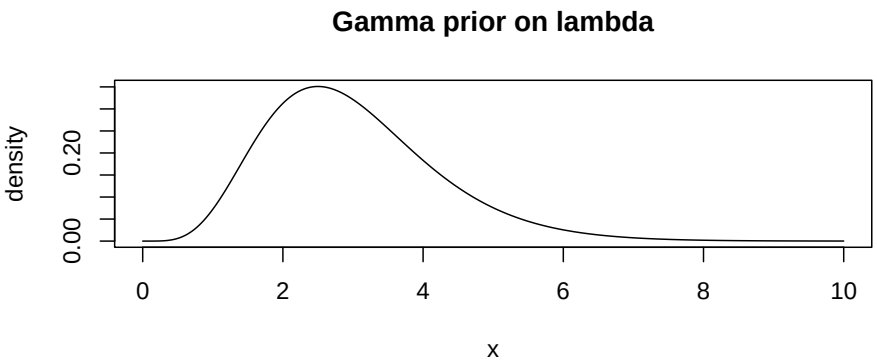
$$\frac{a}{b^2} = 1.5 \tag{31}$$

Just solve for a and b (exercise).

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

The prior on λ is:

$$\lambda \sim Gamma(a = 6, b = 2) \tag{32}$$



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
```

Given that

$$\text{Posterior} \propto \text{Likelihood Prior} \tag{33}$$

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

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

$$\begin{aligned} f(\mathbf{x} \mid \lambda) =& \frac{\exp(-\lambda)\lambda^{x_1}}{x_1!} \times \cdots \times \frac{\exp(-\lambda)\lambda^{x_n}}{x_n!} \\ =& \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \\ =& \frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!} \end{aligned} \tag{34}$$

Computing the posterior is surprisingly easy now:

$$\text{Posterior} = \left[\frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}!}\right] \left[\frac{\mathbf{b^a}\lambda^{a-1}\exp(-b\lambda)}{\mathbf{\Gamma(a)}}\right] \tag{35}$$

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**:

$$\begin{aligned}\text{Posterior} \propto & \exp(-n\lambda)\lambda^{\sum_i^n x_i}\lambda^{a-1}\exp(-b\lambda) \\ = & \lambda^{a-1+\sum_i^n x_i}\exp(-\lambda(b+n))\end{aligned} \tag{36}$$

$$\begin{aligned}\text{Posterior} \propto & \exp(-n\lambda)\lambda^{\sum_i^n x_i}\lambda^{a-1}\exp(-b\lambda) \\ = & \lambda^{a-1+\sum_i^n x_i}\exp(-\lambda(b+n))\end{aligned} \tag{37}$$

- 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:

$$\lambda^{a^*-1}\exp(-\lambda b^*) \tag{38}$$

- 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.:

$$k\int_0^\infty \lambda^{a^*-1}\exp(-\lambda b^*) = 1 \tag{39}$$

- 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.

Concrete example given data

- Suppose the regressive eye movements from one subject on five 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

$$\begin{aligned}Gamma(a^* = a + \sum_i^n x_i, b^* = b + n) = & Gamma(6 + 16, 2 + 5) \\ = & Gamma(22, 7)\end{aligned} \tag{40}$$

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

The posterior's mean is a weighted mean of the MLE and the prior mean

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:

$$\frac{a*}{b*} = \frac{a + \sum x_i}{n + b} \tag{41}$$

This can be rewritten as

$$\frac{a*}{b*} = \frac{a + n\bar{x}}{n + b} \tag{42}$$

Dividing both the numerator and denominator by b:

$$\frac{a*}{b*} = \frac{(a + n\bar{x})/b}{(n + b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \tag{43}$$

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

$$\frac{a/b + n\bar{x}/b}{1 + n/b} = \frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \tag{44}$$

We can rewrite this as:

$$\frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} = \frac{m \times 1}{1 + \frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \tag{45}$$

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

$$m\frac{w_1}{w_1 + w_2} + \bar{x}\frac{w_2}{w_1 + w_2} \tag{46}$$

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.

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:

$$\frac{a*}{b*^2} = \frac{(a + n\bar{x})}{(n + b)^2} \tag{47}$$

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.
- 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

Youtube lecture: https://youtu.be/ymuLt6LBTwY