# BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

# 3 Constructing Bayesian models

In the last lecture, we saw how to perform parameter estimation for repeated measurements with a Gaussian likelihood and prior that goes like the inverse of the standard deviation of the Gaussian. Most of last lecture was then finding ways to summarize the posterior. We saw, and this is generally true, that we need only to specify the likelihood and prior to build the statistical model. In this lecture, we will discuss ways to build a statistical model. We will do this using two examples, learning general principles as we work through them.

## 3.1 Example 1: Mitotic spindle size

Matt Good and coworkers (Good, et al., *Science*, **342**, 856–860, 2013) developed a microfluidic device where they could create droplets of cytoplasm extracted from *Xenopus* eggs and embryos (see Fig. 3). A remarkable property about *Xenopus* extract is that mitotic spindles spontaneously form; the extracted cytoplasm has all the ingredients to form them. This makes it an excellent model system for studying spindles. With their device, Good and his colleagues were able to study how the size of the cell affects the dimensions of the mitotic spindle; a simple, yet beautiful, question. The experiment is conceptually simple; they made the droplets and then measured their dimensions and the dimensions of the spindles using microscope images.

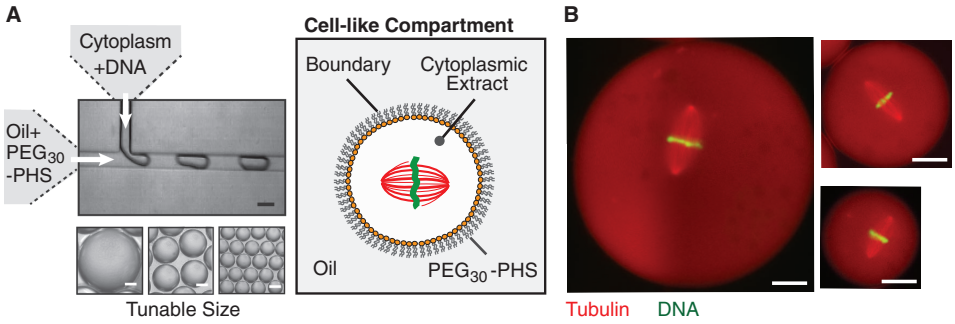


Figure 3: Schematic of spindle size experiment. Scale bars are 20 µm. Taken from Fig. 1 of Good, et al., *Science*, **342**, 856–860, 2013.

The question the authors were after was about how the spindle size scaled with the diameter of the droplet. The data they acquired are shown in Fig. 4.

### 3.1.1 The cartoon model

Recall in lecture 1 that we went through the process of developing a statistical model, starting with a cartoon model, mathematizing it, and then making a statistical model

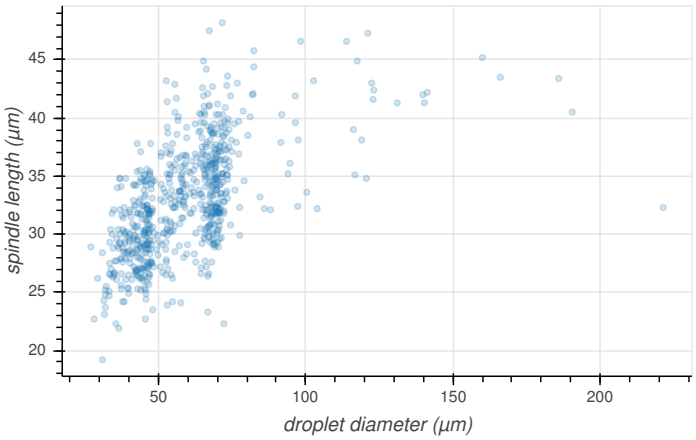


Figure 4: Spindle length versus droplet diameter.

describing how the data might vary from the mathematical model due to naturally occurring variability and that in the experiments. Good and coworkers hypothesized that the length of spindles is regulated by the total amount of tubulin available to make them. Specifically, the three key principles of their "cartoon" model are:

1. The total amount of tubulin in the droplet or cell is conserved.
2. The total length of polymerized microtubules is a function of the total tubulin concentration after assembly of the spindle. This results from the balances of microtubule polymerization rate with catastrophe frequencies.
3. The density of tubulin in the spindle is independent of droplet or cell volume.

### 3.1.2 The mathematical model

From these principles, we need to derive a mathematical model that will provide us with testable predictions. The derivation follows, and you may read it if you are interested. Since our main focus here is building a statistical model, you can skip ahead to equation (3.14), where we define a mathematical expression relating the spindle length, $l$ to the droplet diameter, $d$, which depends on two parameters, $\gamma$ and $\phi$.

Principle 1 above (conservation of tubulin) implies

$$T_0 V_0 = T_1(V_0 - V_s) + T_s V_s, \tag{3.1}$$

where $V_0$ is the volume of the droplet or cell, $V_s$ is the volume of the spindle, $T_0$ is the total tubulin concentration (polymerized or not), $T_1$ is the tubulin concentration in the cytoplasm after the the spindle has formed, and $T_s$ is the concentration of tubulin in the spindle. If we assume the spindle does not take up much of the total volume

of the droplet or cell ($V_0 \gg V_s$, which is the case as we will see when we look at the data), we have

$$T_1 \approx T_0 - \frac{V_s}{V_0} T_s. \tag{3.2}$$

The amount of tubulin in the spindle can we written in terms of the total length of polymerized microtubules, $L_{MT}$ as

$$T_s V_s = \alpha L_{MT}, \tag{3.3}$$

where $\alpha$ is the tubulin concentration per unit microtubule length. (We will see that it is unimportant, but from the known geometry of microtubules, $\alpha \approx 2.7$ nmol/μm.)

We now formalize assumption 2 into a mathematical expression. Microtubule length should grow with increasing $T_1$. There should also be a minimal threshold $T_{min}$ where polymerization stops. We therefore approximate the total microtubule length as a linear function,

$$L_{MT} \approx \begin{cases} 0 & T_1 \leq T_{min} \\ \beta(T_1 - T_{min}) & T_1 > T_{min}. \end{cases} \tag{3.4}$$

Because spindles form in *Xenopus* extract, $T_0 > T_{min}$, so there exists a $T_1$ with $T_{min} < T_1 < T_0$. Thus, going forward, we are assured that $T_1 > T_{min}$. So, we have

$$V_s \approx \alpha \beta \frac{T_1 - T_{min}}{T_s}. \tag{3.5}$$

With insertion of our expression for $T_1$, this becomes

$$V_s \approx \alpha \beta \left( \frac{T_0 - T_{min}}{T_s} - \frac{V_s}{V_0} \right). \tag{3.6}$$

Solving for $V_s$, we have

$$V_s \approx \frac{\alpha \beta}{1 + \alpha \beta / V_0} \frac{T_0 - T_{min}}{T_s} = \frac{V_0}{1 + V_0 / \alpha \beta} \frac{T_0 - T_{min}}{T_s}. \tag{3.7}$$

We approximate the shape of the spindle as a prolate spheroid with major axis length $l$ and minor axis length $w$, giving

$$V_s = \frac{\pi}{6} l w^2 = \frac{\pi}{6} k^2 l^3, \tag{3.8}$$

where $k \equiv w/l$ is the aspect ratio of the spindle. We can now write an expression for the spindle length as

$$l \approx \left( \frac{6}{\pi k^2} \frac{T_0 - T_{min}}{T_s} \frac{V_0}{1 + V_0 / \alpha \beta} \right)^{\frac{1}{3}}. \tag{3.9}$$

For small droplets, with $V_0 \ll \alpha\beta$, this becomes

$$l \approx \left( \frac{6}{\pi k^2} \frac{T_0 - T_{\min}}{T_s} V_0 \right)^{\frac{1}{3}} = \left( \frac{T_0 - T_{\min}}{k^2 T_s} \right)^{\frac{1}{3}} d, \tag{3.10}$$

where $d$ is the diameter of the spherical droplet or cell. So, we expect the spindle size to increase linearly with the droplet diameter for small droplets.

For large $V_0$, the spindle size becomes independent of droplet size;

$$l \approx \left( \frac{6\alpha\beta}{\pi k^2} \frac{T_0 - T_{\min}}{T_s} \right)^{\frac{1}{3}}. \tag{3.11}$$

We can define two parameters to describe the data,

$$\gamma = \left( \frac{T_0 - T_{\min}}{k^2 T_s} \right)^{\frac{1}{3}} \tag{3.12}$$

$$\phi = \left( \frac{6\alpha\beta}{\pi} \right)^{\frac{1}{3}}. \tag{3.13}$$

We assume that $\gamma$ and $\phi$ are the same for all data. We can rewrite the general model expression in terms of these parameters as

$$l(d; \gamma, \phi) \approx \frac{\gamma d}{(1 + (d/\phi)^3)^{\frac{1}{3}}}. \tag{3.14}$$

For small and large droplets, respectively, we have

$$l \approx \gamma d \quad \text{for } d/\phi \ll 1, \tag{3.15}$$

$$l \approx \gamma \phi \quad \text{for } d/\phi \gg 1. \tag{3.16}$$

Note that the expression for the linear regime gives bounds for $\gamma$. Obviously, $\gamma > 0$. Because $l \leq d$, lest the spindle not fit in the droplet, we also have $\gamma \leq 1$. The parameter $\phi$ is independent of the system geometry, so it only has the physical lower bound of $\phi > 0$.

### 3.1.3 A comment on the model parameters

We went through some algebraic manipulations to get our mathematical model in a form with two parameters. We want to try to identify *independent* parameters in your mathematical before doing regression analysis. In a trivial example, imagine someone proposed the following model to use in a regression on $(x, y)$ data:

$$y = ax + bx + c. \tag{3.17}$$

Obviously, it would be silly to have both $a$ and $b$ as regression parameters, and we should instead define a new parameter $d = a + b$ and use that as a regression parameter. In the case of spindle length, we had parameters $T_0$, $T_{\min}$, $T_s$, $k$, $\alpha$, and $\beta$, but, as we saw, we can only resolve two parameters, $\gamma$ and $\phi$. Furthermore, if we happen to be in the linear regime, $\phi$ does not enter the expressions, so we obviously cannot resolve it. Similarly, we can only determine $\phi$ if we are in the plateau regime.

### 3.1.4 The statistical model: The likelihood

We have a mathematical model, so now we are left to specify the likelihood and prior. We will start with the likelihood. The data are pairs of droplet diameters and spindle lengths. We denote one such pair as $(d_i, l_i)$, and the whole data set as $D = \mathbf{d}, \mathbf{l}$. The parameters are $\theta = \gamma, \phi$. So, the likelihood is $f(D \mid \theta) = f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \theta_s)$, where $\theta_s$ are the parameters associated with the statistical model (as opposed to $\gamma$ and $\phi$, which are associated with the mathematical model).

We need a probabilistic model about how the observe data might vary stochastically about the mathematical model. We can write

$$l_i = l(d_i; \gamma, \phi) + e_i, \tag{3.18}$$

where $e_i$ is how much the measured spindle length differs from the predicted length for the measured drop diameter. So, we are left to choose how $e_i$ is distributed.

Because many processes come together to make a spindle, and then to measure its length, it is reasonable to assume that $e_i$ is Gaussian distributed. The mean of this Gaussian should be zero, since on average, the model should fit the data. One way to write this is

$$e_i \sim \text{Norm}(0, \sigma_i). \tag{3.19}$$

This reads as, "The error $e_i$ is Normally distributed with mean zero and standard deviation $\sigma_i$." This notation is commonly used to make a sentence like the one I just quoted more concise. We could also write out the full PDF.

$$f(e_i \mid \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \, \mathrm{e}^{-e_i^2/2\sigma_i^2}. \tag{3.20}$$

Thus, for a single data point, we have

$$f(d_i, l_i \mid \gamma, \phi, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \, \exp\left[-\frac{(l_i - l(d_i; \gamma, \phi))^2}{2\sigma_i^2}\right]. \tag{3.21}$$

This could be equivalently written as

$$l_i \sim \text{Norm}(l(d_i; \gamma, \phi), \sigma_i). \tag{3.22}$$

Now, if each measurement is independent, the likelihoods of each data point multiply, giving

$$f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \{\sigma\}) = \frac{1}{(2\pi)^{n/2} \prod_i \sigma_i} \exp\left[ -\sum_i \frac{(l_i - l(d_i; \gamma, \phi))^2}{2\sigma_i^2} \right], \tag{3.23}$$

where $n$ is the number of observations of $d_i, l_i$ pairs we have and $\{\sigma\}$ represents the $\sigma_i$ values. If these are all equal, we have a single $\sigma$, which gives a likelihood of

$$f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_i (l_i - l(d_i; \gamma, \phi))^2 \right]. \tag{3.24}$$

This can equivalently be written as

$$l_i \sim \text{Norm}(l(d_i; \gamma, \phi), \sigma) \quad \forall i. \tag{3.25}$$

We thus have our likelihood. We have assumed that each measurement is independent of the others and that the variation from the model is **homoscedastic**, which means that the magnitude of the error of measured data from the model is the same for all data points (as opposed to **heteroscedastic**).

### 3.1.5 Choice of prior

We are now left to the choice of the prior. Before we embark on the journey of choosing the prior, I quote Efron and Hastie from their book, *Computer Age Statistical Inference*.

> For 200 years, however, two impediments stood between Bayesian theory's philosophical attraction and its practical application.
>
> 1 In the absence of relevant past experience, the choice of a prior distribution introduces an unwanted subjective element into scientific inference.
> 2 Bayes' rule looks simple enough, but carrying out the numerical calculation of a posterior distribution often involves intricate higher-dimensional integrals.

We will deal with the second impediment in coming weeks when we use **Markov chain Monte Carlo** to handle the intricate integrals. Our goal now is to come up with a prior distribution that avoids subjectivity. As Efron and Hastie called this process an impediment, we proceed with trepidation.

The prior encodes our knowledge about the parameters of the statistical model. In this case, we have three parameters, $\gamma$ and $\phi$, which entered through the physical model, and $\sigma$ which entered through our modeling of the variability inherent in the system and in measurement. So, we need to specify $g(\gamma, \phi, \sigma)$.

**Independence of priors.** Our first step on the journey to specifying $g(\gamma, \phi, \sigma)$ is to note that these parameters should be independent of each other. The parameter $\gamma$ depends only on the aspect ratio of spindles, and the total concentration of tublin in the cell, the concentration of tubulin in the cytoplasm, and the critical concentration of tubulin where microtubule growth arrests. The parameter $\phi$ depends on the concentration of tubulin in a single microtubule (known from the geometry of microtubules) and a constant of proportionality between microtubule length and cytoplasmic tubulin concentration. Because they depend on distinct, independent physical quantities, the parameters $\gamma$ and $\phi$ are independent of each other. Similarly, the parameter $\sigma$ describes how much the spindle length differs from the prediction. It is a bit harder to state that this is independent of $\gamma$ and $\phi$. However, doing so is a less egregious approximation, perhaps, than assuming homoscedasticity in the first place. So, we will proceed assuming all three parameters are independent, so

$$g(\gamma, \phi, \sigma) = g(\gamma)\, g(\phi)\, g(\sigma). \tag{3.26}$$

**Uninformative priors.** If we want to reduce subjectivity in our prior, we want to remain as ignorant as possible about the parameters before we see the data. However, we are not *completely* ignorant. For example, we know for sure that $0 \leq \gamma \leq 1$ based on physical arguments stated at the end of section 3.1.2. This should also be encoded in our prior, such that $g(\gamma) = 0$ for all negative $\gamma$ and for all $\gamma > 1$.

If we want to avoid subjectivity, we might say, then, that any value of $\gamma$ on the interval from zero to one is equally likely as any other. In this case, we have

$$g(\gamma) = \begin{cases} 1 & 0 \leq \gamma \leq 1, \\ 0 & \text{otherwise}, \end{cases} \tag{3.27}$$

or

$$\gamma \sim \text{Uniform}(0, 1), \tag{3.28}$$

which says that the prior distribution of $\gamma$ is Uniform on the interval $[0, 1]$[8]. This notion of assigning equal probability to all possibilities is often referred to as **Laplace's**

[8]Strictly speaking, we should have that $\gamma > 0$, not $\gamma \geq 0$, since a zero value for $\gamma$ under this model would mean that spindles always have zero length.

**Principle of Insufficient Reason**.

Unlike $\gamma$, $\phi$ has no upper bound. Yes, it must be positive, but the upper bound is not apparent. Remember, though, that the prior contains all information we know before the experiment. For example, we know that the mitotic spindle in a one-cell *Xenopus* embryo do not span the entire cell, and that the cell is about 2 mm across (huge!). So, the absolute maximum we could expect for $\gamma\,\phi$ is 2 mm. So there is a reasonable maximum we could choose.

So, we could choose a Uniform prior for $\phi$ on the interval of, say zero to ten mm. But is this really uninformative? John Venn and Ronald Fisher, famous attackers of a Bayesian approach, would say no. They could argue that we could equally well have chosen to to parametrize the model in terms of $\xi = \phi^{-3}$ instead so that the theoretical expression for spindle length is

$$l(d;\gamma,\phi) = \frac{\gamma d}{(1+\xi d^3)^{\frac{1}{3}}}. \tag{3.29}$$

If we chose a Uniform prior on $\phi$, then the prior on $\xi$ is no longer Uniform. Recall the **change of variables formula** from multivariate calculus.

$$g(\xi) = \left|\frac{\mathrm{d}\phi}{\mathrm{d}\xi}\right| g(\phi). \tag{3.30}$$

Taking $g(\phi)$ to be a constant (which it is for a Uniform distribution), we perform the change of variables, to get

$$g(\xi) = \left|-\frac{\xi^{-\frac{4}{3}}}{3}\right| g(\phi) = \text{constant}\cdot\,\xi^{-\frac{4}{3}}, \tag{3.31}$$

which is no longer flat. So perhaps in cases like this, a Uniform prior is not actually uninformative; we are biasing toward a certain parametrization. We desire **transformation invariance**, meaning that the prior should be the same functional dependence on $\phi$ if we transform the parameter in a certain way.

Generically, this means that if we have a set of parameters $\theta$ that are transformed into a new set of parameters $\zeta$, we should choose $g(\theta)$ such that

$$\left|\frac{\partial(\zeta_1,\zeta_2,\ldots)}{\partial(\theta_1,\theta_2,\ldots)}\right| g(\zeta(\theta)) \tag{3.32}$$

has the same functional form as $g(\theta)$, up to a multiplicative constant. The first factor in this equation denotes the Jacobian, which is the absolute value of the determinant of the Jacobi matrix.

So, in the present example let's say we want our prior to be invariant if we transform $\phi$ to a new variable $\xi$ such that $\xi = \phi^a$. That is, we want

$$g(\xi(\phi)) = \left|\frac{\mathrm{d}\phi}{\mathrm{d}\xi}\right| g(\phi) = a\phi^{a-1} g(\phi) \tag{3.33}$$

to have the same $\phi$ dependence as $g(\phi)$. If we pick $g(\phi) = c/\phi$, where $c$ is a constant, we see that this is indeed the case.

$$g(\xi(\phi)) = \frac{ac}{\phi}, \tag{3.34}$$

which has the same $\phi$-dependence.

This kind of prior, which is uninformative maintaining transformational invariance like we have just described, is a case of a **Jeffreys prior** (discussed very briefly at the end of this lecture). In fact, in portions of the literature, including in Sivia's book, such a prior, $g(\theta) \propto 1/\theta$, is just called "a Jeffreys prior." *For the purposes of this course, this is what we mean when we refer to a Jeffreys prior.*

It makes sense, then, to also parametrize $\sigma$ with a Jeffreys prior, since we could also have chosen to parametrize the likelihood with $\tau = \sigma^{-1}$.

**Proper and improper priors.** Our prior for $\gamma$, being Uniform on the interval from zero to one, is **proper**, in the sense that it is properly normalized. If we did not have bounds on it, we would call it an **improper prior**, since it cannot be normalized. The same is true for the Jeffreys prior. If we do not define bounds for a prior of the form $g(\theta) \propto 1/\theta$, it cannot be normalized, since

$$\int_b^\infty \frac{\mathrm{d}\theta}{\theta} \tag{3.35}$$

diverges for any positive b, as does

$$\int_0^b \frac{\mathrm{d}\theta}{\theta}. \tag{3.36}$$

Usually, this is not a problem for the problem of parameter estimation, that is computing $g(\theta \mid D)$. This is because for extreme values of the parameters $\theta$, the likelihood typically is vanishingly small. Recall, the posterior is

$$g(\theta \mid D) = \frac{f(D \mid \theta)\, g(\theta)}{\int \mathrm{d}\theta\, f(D \mid \theta)\, g(\theta)}. \tag{3.37}$$

Since $f(D \mid \theta)$ typically is tiny for extreme parameter values, it overwhelms the finite $g(\theta)$ in the numerator, and in the integral in the denominator. Furthemore, any normalization constants for $g(\theta)$ cancel out with those appearing in the denominator while computing the posterior.

While this is convenient for the parameter estimation problem, as we will see in later lectures, we do need to *exactly* compute the evidence,

$$\int \mathrm{d}\theta\, f(D \mid \theta)\, g(\theta), \tag{3.38}$$

when doing model selection. So, we should bound and normalize the priors with reasonable bounds. We can write our prior for our example problem with spindle lengths as

$$g(\gamma, \phi, \sigma) = g(\gamma)\, g(\phi)\, g(\sigma), \tag{3.39}$$

with

$$g(\gamma) = \begin{cases} 1 & 0 \leq \gamma \leq 1, \\ 0 & \text{otherwise}, \end{cases} \tag{3.40}$$

$$g(\phi) = \begin{cases} \frac{1}{\phi\, \ln(\phi_{\max}/\phi_{\min})} & \phi_{\min} \leq \phi \leq \phi_{\max}, \\ 0 & \text{otherwise}, \end{cases} \tag{3.41}$$

$$g(\sigma) = \begin{cases} \frac{1}{\sigma\, \ln(\sigma_{\max}/\sigma_{\min})} & \sigma_{\min} \leq \sigma \leq \sigma_{\max}, \\ 0 & \text{otherwise}. \end{cases} \tag{3.42}$$

Alternatively, we could write this as

$$\gamma \sim \text{Uniform}(0, 1), \tag{3.43}$$

$$\phi \sim \text{Jeffreys}(\phi_{\min}, \phi_{\max}), \tag{3.44}$$

$$\sigma \sim \text{Jeffreys}(\sigma_{\min}, \sigma_{\max}). \tag{3.45}$$

### 3.1.6 Choosing bounds

We saw that we could choose the bounds on $\gamma$ with physical arguments. We would like to make similar choices for bounds for $\phi$ and $\sigma$. We already made a physical argument based on the size of *Xenopus* embryos that the maximal $\phi$ cannot be more than a few millimeters. Its lower bound cannot be zero because this would mean that the spindle length would always be zero. We might instead choose a lower bound to be something like 10 nanometers, about the size of a microtubule nucleus.

Choosing bounds on $\sigma$ can be a bit more challenging, because it is describing variability in the experiment. We might choose an upper bound close to the maximal size of a spindle, since we would not get variation bigger than that. So, one millimeter is plenty big for an upper bound. For the lower bound, we might again choose 10 nanometers, as this is about the size of four or five tubulin diameters, which should be the smallest fluctuation we could imagine seeing.

### 3.1.7 Computing the posterior

Our specification of the posterior is now complete. We have specified the likelihood and prior. The evidence can be calculated by integrating the product of the likelihood and prior over all parameter values. Actually computing, plotting, and summarizing the posterior is a separate challenge. Specifically, it is impediment number 2 laid out by Efron and Hastie. This is the subject of the next few weeks of the course.

## 3.2 Example 2: Worm reversals

In Homework 3.3, we consider reversals upon exposure to blue light of *C. elegans* that have a Channelrhodopsin in a specific neuron. There is some probability $p$ of reversal. Say we do $n$ trials and observe $r$ reversals. The likelihood is Binomially distributed according to the story of the Binomial distribution. So, Bayes theorem reads

$$g(p \mid n, r) = \frac{f(r \mid p, n)\, g(p)}{f(r \mid n)}, \tag{3.46}$$

where

$$f(r \mid p, n) = \frac{n!}{(n-r)!r!}\, p^r (1-p)^{n-r}, \tag{3.47}$$

which we could alternatively write as

$$r \mid p, n \sim \text{Binom}(n, p), \tag{3.48}$$

(Note that I wrote $g(p)$ instead of $g(p \mid n)$ because they are equal; $n$ has no bearing on $p$.)

As we consider our choice of prior, $g(p)$, Think back to the first lecture when we talked about Bayes's theorem as a model for learning. The idea there was that we know something before (*a priori*) acquiring data, and then we update our knowledge after (*a posteriori*). So, we come in with the prior and out with the posterior after acquiring data. It might make sense, then, that the prior and the posterior distributions are the same. That is to say they are the same distribution, but with different parameters. The parameters get updated going from the prior to the posterior. When this is the case, the prior is said to be **conjugate** to the likelihood. This makes sense: the likelihood determines the relationship between the prior and the posterior, so it should determine the functional form of the prior/posterior such that they are the same.

### 3.2.1 Conjugate priors

What functional form can we choose for the prior $g(p)$ such that the posterior $g(p \mid n, r, I)$ has the same functional form? This requires some serious mathematical work,

but the answer is the Beta distribution. The Beta distribution is parametrized by two positive parameters, $a$ and $b$,

$$g(p \mid a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \tag{3.49}$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{3.50}$$

is the Beta function. The distribution is defined on the interval $0 \leq p \leq 1$. Importantly, or $a = b = 1$, we get a Uniform distribution. The Uniform distribution on the interval from zero to one is therefore a special case of the Beta distribution.

Now, if we insert a Beta distribution for the posterior and prior, we have

$$g(p \mid n, r, a, b) = \frac{f(r \mid p, n)\, g(p \mid a, b)}{f(r \mid n)} \tag{3.51}$$

$$= \frac{1}{f(r \mid n)} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \tag{3.52}$$

$$= \frac{1}{f(r \mid n)\, B(a,b)} \frac{n!}{(n-r)!r!} p^{r+a-1}(1-p)^{n-r+b-1}. \tag{3.53}$$

In looking at this expression, the only bit that depends on $p$ is $p^{r+a-1}(1-p)^{n-r+b-1}$, which is exactly the $p$-dependence of a Beta distribution with parameters $r+a$ and $n-r+b$. Because the posterior must be normalized, the posterior must be a Beta distribution and

$$\frac{1}{f(r \mid n)\, B(a,b)} \frac{n!}{(n-r)!r!} = \frac{1}{B(r+a, n-r+b)}. \tag{3.54}$$

We have just normalized the posterior without doing any nasty integrals! So, the posterior is

$$g(p \mid n, r, a, b) = \frac{p^{r+a-1}(1-p)^{n-r+b-1}}{B(r+a, n-r+b)}, \tag{3.55}$$

or,

$$p \mid n, r, a, b \sim \text{Beta}(r+a, n-r+b). \tag{3.56}$$

So, we can see that conjugacy is useful. For a given likelihood, if we know its conjugate prior, we can just immediately write down the posterior in a clear form. The Wikipedia page on conjugate priors has a useful table of likelihood-conjugate pairs.

Note though that a closed form conjugate does not always exist for a given likelihood, especially for complicated models, and when they do exist, they may be very difficult to find. This does limit their utility. Further, there is no reason why a posterior and prior should have the same functional form; all analysis is completely valid without conjugacy. Sivia has stinging words about using conjugate priors: "While we might expect our initial understanding of the object of interest to have a bearing on the experiment we conduct, it seems strange that the choice of the prior pdf should have to wait for, and depend in detail upon, the likelihood function."

## 3.3 The impediment is not resolved

We tried to be as objective as possible in choosing our priors. We intentionally tried to be uninformative, and took into account transformation invariance. This has flaws, since it is mathematically impossible to come up with a prior that can be invariant to *all* transformations. There are other strategies for choosing uninformative priors. Among them are

- Using what is generically called a Jeffreys prior by computing the Fisher information from the likelihood.
- Using the principle of maximum entropy. Entropy can be thought of as a formal metric of ignorance, which we wish to maximized when being objective. Sivia talks about this in Chapter 5.

Both of these methods are outside the scope of this course, but they are important to consider when choosing priors.

Now consider Sivia's comment I just quoted. And now consider the title of a recent paper by Gelman, Simpson, and Betancourt, "The prior can generally only be understood in the context of the likelihood." Some of the section headings in that paper are also gold, like "Uniform priors are not a panacea and can do unbounded damage."

So, obviously there is disagreement about constructing priors. One the one hand, we want to be as objective as possible in predicting priors. That said, we almost always do know *something* about parameter values *a priori*. We really *should* encode that in the prior. Furthemore, a probability density function is just a pdf. It only becomes a prior when it is connected to a likelihood. So we may need to this with some degree of pragmatism.

It is very hard to be truly informative in more complicated models, such as the very powerful hierarchical models we will work with later in the class. Furthermore, when using flat priors, the other impediment comes back in. Flat priors can really wreak havoc on Markov chain Monte Carlo (MCMC) samplers in hierarchical mod-

els, thereby exacerbating the difficulty in computing the posterior. (If you cannot compute it, what good is it?)

It's probably no coincidence that Gelman and coworkers are lead developers on one of the major MCMC packages, called Stan. In fact, the Stan wiki has some guidelines about choice of priors, which run quite contradictory to what we have just discussed here. Specifically, as of October 11, 2017, there is this: "Some principles we don't like: invariance, Jeffreys, entropy." The Stan developers are obviously going to be more pragmatic in their views, since they are in the business of actually computing posteriors. They tend to favor **weakly informative priors**; things like broad Gaussians. Their reasons, again quoting the Wiki,

> - Weakly informative prior should contain enough information to regularize: the idea is that the prior rules out unreasonable parameter values but is not so strong as to rule out values that might make sense
> - Weakly informative rather than fully informative: the idea is that the loss in precision by making the prior a bit too weak (compared to the true population distribution of parameters or the current expert state of knowledge) is less serious than the gain in robustness by including parts of parameter space that might be relevant.

In the end, my view is that you want to encode all of the information you confidently have about parameters, and not more, into the prior. For example, *before seeing the data*, if you think that the variability in measured spindle length should be about a 10 microns, you could choose a weakly informative prior, like a Gaussian with mean of one micron and standard deviation of 3 microns, and you would probably be fine. As you can see in homework 3.4, the choice of prior often has very little effect on the end result of your inference.