Notes for Bayesian Statistics

Matt Cocci

January 24, 2013

Contents

1. Statistical Inference: From Classical to Bayesian

1.1. Statistical Inference

In conducting *Statistical Inference*, there are a few main, recurring goals:

1. *Point Estimation*: What is the "best" value for an unknown parameter?
2. *Variability Estimation*: Create an interval of likely values for some unknown parameter. Under the two main paradigms, this is called
   - Confidence Interval in classical statistics.
   - Posterior Interval in Bayesian statistics.
3. Figure out probabilities for specific events or test hypotheses.
4. Predict out of sample events.

1.2. Classical Statistics

Typically, you start by assuming a *statistical model* that links observed data to a set of unknown parameters. For example, you can assume $Y_i \sim N(\mu, \sigma^2)$, i.i.d.

Next, you want to estimate a vector of parameters, θ, which are assumed to be fixed, unknown values. Typically, you do this by the **likelihood principle**, where you say that the probability model is based on some density, $p(Y_i|\theta)$. Then, after observing data, you compute

$$\text{Likelihood} = \prod_{i=1}^{n} p(y_i|\theta)$$

Then from there, choose parameter values that make our data as likely as possible:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(\mathbf{y}|\theta)$$

For simple models, this is often possible to compute analytically, but for larger models, you'll need to resort to optimizing algorithms like Newton-Raphson.

However, given that there is some uncertainty in terms of which random sample (of many) that we observed, we would like to quantify the uncertainty in our measurements. To do so, we use the concept of a **sampling distribution**, which describes the distribution of values taken on by a parameter or statistic across *all possible* samples. This is the basis for hypothesis testing and confidence intervals. In very large samples (which classical statistics is optimized for), this is a very big deal.

However, there are some disadvantages which we will want to discuss.

1. Assumes parameters are *fixed constants*, disallowing prob. statements about θ.

2. The interpretation of Confidence Intervals is rather unwieldy. It's not a probability statement, and only says that in 95% of the many samples we could take, the CIs will contain μ.

3. Most of the variability results are *asymptotic*, implicitly requiring large samples.

1.3. From Classical to Bayesian Paradigm

The major feature of Bayesian Statistics is that unknown parameters, θ, do *not* have fixed values. Instead, they are random variables with their own distribution: $p(\theta)$. This allows us to make direct probability statements about θ, although it does complicate matters a bit.

As a consequence of how we now view θ, point estimates will factor much less heavily as we believe that θ has an entire distribution of values.

1.4. Summarizing the Bayesian Approach

The main goal of the Bayesian approach is to make probability statements about a parameter θ or a unobserved data $\tilde{\mathbf{y}}$ *conditional* on the observed value of $\mathbf{y}$.[1] Typically, we write these statements as $p(\theta|\mathbf{y})$ or $p(\tilde{\mathbf{y}}|\mathbf{y})$.

So to make those probability statements, we must specify a model that provides a *joint probability distribution* for θ and $\mathbf{y}$. However, we typically break this up as follows:

$$p(\theta, \mathbf{y}) = p(\theta)p(\mathbf{y}|\theta)$$

where $p(\theta)$ is a model for the parameter called the **prior** distribution, and we call $p(\mathbf{y}|\theta)$ the **likelihood function**, which we'll describe in more detail.[2]

The *likelihood function*, $p(\mathbf{y}|\theta)$ is viewed as a function of θ for fixed $\mathbf{y}$—i.e. after you observe the data. This is precisely the means through which one *updates* an initial prior distribution. So having observed data, you then *update* the model for θ using Bayes' Rule:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \cdot p(\theta)}{p(\mathbf{y})} \propto p(\mathbf{y}|\theta) \cdot p(\theta). \tag{1}$$

Note that we show the *unnormalized* posterior density after $\propto$ because $p(\mathbf{y})$ is simply a constant with respect to θ, so we will often drop it from our analysis since we can just normalize the numerator by a constant so that it integrates to 1.

[1] Note that $\mathbf{y}$ is a vector representing the observed data, $(y_1, \ldots, y_n)$.

[2] The likelihood function assumes that name when θ varies and $\mathbf{y}$ is *fixed*, so that $p(\mathbf{y}|\theta)$ is a particular function of θ. As a result, we typically use this name *after* we observe data. However, when $\mathbf{y}$ is not fixed, we can also call it the *sampling distribution*—for obvious reasons—as Gelman does.

This value, $p(\mathbf{y})$ in the denominator, is often called the *prior predictive distribution* or the *marginal likelihood*. It's called the first as it represents how we would predict $\mathbf{y}$ before we observe any data. Specifically, before we see $\mathbf{y}$ and update θ to arrive at a posterior, we would average the likelihood of observing $\mathbf{y}$ *given* θ over all possible values of θ that we allow for in our prior. We can see this in the formulation of $\mathbf{y}$:

$$p(\mathbf{y}) = \int p(y, \theta)\ d\theta = \int p(y|\theta)p(\theta)\ d\theta \tag{2}$$

It's second moniker, the marginal likelihood, arises from this same averaging out of θ to get a marginal distribution.

Finally, we return to the updated result, $p(\theta|\mathbf{y})$, called the **posterior**, which we can state even more intuitively:

$$(\text{Posterior}) = \frac{(\text{Likelihood})(\text{Prior})}{(\text{Marginal Distribution})}$$

Note, if we have a lot of data, they will dominate the prior and wash out many of the effects of a potentially bad prior. (But in that case, why not just use classical statistics in the first place?)

1.5. Prediction

Often, we will want to use our probability model to make statements about out of sample observations, once we have observed some data and updated our model. For that reason, we look to the **posterior predictive distribution**:

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}, \theta|y)\ d\theta \tag{3}$$

$$= \int p(\tilde{\mathbf{y}}|\theta, y)p(\theta|\mathbf{y})\ d\theta \tag{4}$$

$$= \int p(\tilde{\mathbf{y}}|\theta)p(\theta|\mathbf{y})\ d\theta \tag{5}$$

Intuitively, Equation 3 says that our prediction of $\tilde{\mathbf{y}}$, which depends upon θ, should *average* over all possible values of θ.

Equation 4 tells us how to do that: consider the likelihood of observing new data $\tilde{\mathbf{y}}$ *given* some assumed θ, but weight that likelihood by the probability that θ actually does equal that assumed θ. What should the weight be? Well, the posterior probability seems sensible, since it represents the best guess of θ's distribution given previously observed data.

Finally, we can jump from Equation 4 to 5 because $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are conditionally independent given θ.

1.6. Likelihood and Odds Ratios

We call the ratio of the posterior density, $p(\theta|\mathbf{y})$ evaluated at points θ_1 and θ_2 the *posterior odds* for θ_1 relative to θ_2. Note that the posterior odds are equal to the prior odds multiplied by the likelihood ratio, $p(\mathbf{y}|\theta_1)/p(\mathbf{y}|\theta_2)$:

$$\frac{p(\theta_1|\mathbf{y})}{p(\theta_2|\mathbf{y})} = \frac{p(\theta_1)p(\mathbf{y}|\theta_1)}{p(\theta_2)p(\mathbf{y}|\theta_2)}$$

1.7. Advantages and Disadvantages of the Bayesian Approach

Let's start with the nice results:

1. Variability of θ summarized by the posterior distribution.
2. We can make direct probability statements about θ now.
3. Hypothesis testing is less important. For example, instead of testing a hypothesis as to whether the treatment effect of a drug is nonzero, we ask what is the *probability* that the effect is greater than zero. Whole different ballgame.
4. We are *not* relying on asymptotic results to come to our conclusions.

Now let's be fair and go over some of the disadvantages:

1. We need to assume or create a prior distribution, leaving us to specify two distributions (the likelihood and the prior) as opposed to just one (the likelihood).
2. Having to specify a prior can be tough if there is really little data or you just have no clue a priori what might be a good specification for the prior.

2. Prior Distributions

Prior selection is one of the most important topics in Bayesian inference and much thought and care must be placed into choosing one. This section hopes the topic in a general framework.

2.1. Conjugate Priors

The formal definition of conjugacy is as follows: If $\mathcal{S}$ is a class of sampling distributions, or likelihoods, $p(\mathbf{y}|\theta)$, and $\mathcal{P}$ is a class of prior distributions, $p(\theta)$ for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{S}$ if

$$p(\theta|\mathbf{y}) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{S} \text{ and } p(\cdot) \in \mathcal{P}.$$

In words, a prior is conjugate if the posterior is of the same form as the prior. Common examples include

$\theta \sim$	$\mathbf{y}\|\theta \sim$
Beta(a,b)	Binomial
Gamma	Poisson
Normal	Normal, unknown mean
Gamma	Normal, unknown variance

Conjugate prior distributions often have nice interpretations, making it easy to understand the posterior results, and making it look like we're adding "prior counts" to get from the data to the posterior.

2.2. Other Classes of Priors

In addition to *conjugate priors*, there are other principles or forms of prior distributions we might adopt. Let's discuss some important ones along with a few general concepts.

Improper Prior In some instances, you can even have an *improper prior* that generates a proper posterior nonetheless. Examples include the Beta(0,0) distribution as a prior, which doesn't integrate to 1, but effectively adds zero "prior counts" to the posterior, which will integrate to 1 with one observation.

Non-Informative Priors Suppose we have virtually no idea about what the prior should look like. Then we might want to construct a prior that is *non-informative*, in the sense that it has as little influence upon the posterior as possible.

Flat Priors One way to have a non-informative prior involves setting $p(\theta) \propto 1$, a constant. This is called a *flat prior* with the consequence that the posterior is proportional to the likelihood:

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta) \cdot 1 \\ &\propto p(\mathbf{y}|\theta) \end{aligned}$$

But this approach does lead to a few problems.

1. Namely, if the range is not bounded, then a flat prior will be an improper prior, integrating to ∞ (ex. Normal distribution).
2. Also, this is a highly risky, as this approach may not lead to a proper posterior.
3. Finally, "flatness" may not be invariant to transformation. Specifically, if $p(\theta) \propto 1$ so that it is flat on the natural scale of θ, it does not necessarily follow that it will be flat relative to transformations of θ. This leads us to *Jeffrey's Invariance Principle* and *Jeffrey's Prior*.[3]

Jeffrey's Prior *Jeffrey's Invariance Principle* states that any rule for determining a non-informative prior for θ should yield the same results if applied to transformations of theta. Luckily, there does in fact exist such a prior, and it's called *Jeffrey's Prior* and is formulated as follows

$$p(\theta) \propto \sqrt{J(\theta)}, \qquad J(\theta) = -E_{\mathbf{y}}\left[\frac{d^2(\ln p(\mathbf{y}|\theta))}{d\theta^2} \mid \theta\right]$$

where $J(\theta)$ is commonly referred to as "Expected Fisher Information." Note that the subscript $\mathbf{y}$ is placed on the expectation to indicate explicitly that this expectation is taken relative to the $\mathbf{y}$, not θ.

Once we have $p(\theta)$ from above, it's an added bonus if we can write Jeffrey's Prior as a *conjugate* prior as well.

Neutral Prior The goal of a *neutral prior* is to craft a prior so that

$$P(\theta > \hat{\theta}_{\text{MLE}}|\mathbf{y}) \approx 1/2$$

which tries to ensure that the posterior distribution is cented at the MLE.

---

[3] As an example, let's consider the transformation $\phi = \ln\left(\frac{\theta}{1-\theta}\right)$, the so-called "log odds." If we make our prior flat with respect to ϕ flat, so that $p(\phi) \propto 1$, then

$$p(\theta) = p(\phi = g^{-1}(\theta))\left|\frac{d\phi}{d\theta}\right| = 1 \cdot \left|\frac{d\phi}{d\theta}\right| = \frac{1}{\theta(1-\theta)}$$

3. Single Parameter Models

In this section we will take a look at various probability distributions and how we can apply the Bayesian approach to such problems.

3.1. Binomial Data with Conjugate Beta Prior

3.1.1. Likelihood

Consider the random variable $\mathbf{y}$, which is the number of successes in n trials $y_1, \ldots, y_n$, so that it is binomially distributed with parameters n, θ. This gives the sampling model (or likelihood)

$$p(\mathbf{y} = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

3.1.2. Conjugate Prior and Special Cases

For our prior distribution, let's suppose that θ has a beta distribution

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

If we get specific and set $a = b = 1$ so that our prior is the uniform distribution, we will obtain what's called the "Wilson Estimate" in our posterior which has many nice properites. In effect, his prior makes each point in the entire range of θ, $[0, 1]$, equally likely.

Also, we could set $a = b = 0$ to get an improper, non-informative prior which nonetheless gives a proper posterior.

3.1.3. Posterior Distribution

As a result, we get the posterior distribution

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(Y|\theta)p(\theta) \\ &\propto \binom{n}{y}\theta^y(1-\theta)^{n-y} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{y+a-1}(1-\theta)^{n-y+b-1} \end{aligned}$$

As a result, the posterior has a beta distribution, $\theta|\mathbf{y} \sim \text{Beta}(y+a,\ n-y+b)$. From there, it's easy to compute expectaitons and variances given the properties of the beta distribution. For examply,

$$E[\theta|\mathbf{y}] = \frac{y+a}{n+a+b},$$

which lends a nice intuitive interpreation. Namely, a acts as prior counts and $a+b$ acts as prior total trials. Here, it's very easy to see how our prior influences the analysis.

3.1.4. Posterior Inference

Say we want to compare the parameter for two different populations, a and b. Assuming the two populations are independent, we can compute

$$\begin{aligned} P(\theta_a, \theta_b | \mathbf{y}_a, \mathbf{y}_b) &\propto P(\mathbf{y}_a, \mathbf{y}_b | \theta_a, \theta_b) P(\theta_a, \theta_b) \\ &\propto P(\mathbf{y}_a, \mathbf{y}_b | \theta_a, \theta_b) P(\theta_a) P(\theta_b) \\ &\propto P(\mathbf{y}_a | \theta_a) P(\theta_a) P(\mathbf{y}_b | \theta_b) P(\theta_b) \end{aligned}$$

To compare the two, we compute

$$P(\theta_a > \theta_b) \int \int_{\theta_a > \theta_b} P(\mathbf{y}_a, \mathbf{y}_b | \theta_a, \theta_b) \, d\theta_a \, d\theta_b$$

which is easy to do with simulation:

1. Draw N values from $\theta_a | \mathbf{y}_a$.
2. Draw N values from $\theta_b | \mathbf{y}_b$.
3. Pair up values just generated and tabulate the proportion of $\theta_a > \theta_b$.

And if we want to do prediction, we will need to compute

$$P(\tilde{\mathbf{y}}_a | \mathbf{y}_a) = \int p(\tilde{\mathbf{y}}_a | \theta_a) p(\theta_a | \mathbf{y}_a) \, d\theta_a$$

which is to say, we find the probability of new results $\tilde{\mathbf{y}}_a$ given the observed data, which boils down to finding the probability of new results given parameter θ which is conditioned on the observed data. Again, this is something we can do by simulation.

3.2. Poisson Data

3.2.1. Likelihood

We'll model the likelihood for a vector of n iid RV's as a Poisson Random variable,

$$p(\mathbf{y}|\theta) \propto \theta^{\Sigma y_i} e^{-n\theta}$$

3.2.2. Conjugate Prior and Special Cases

Naturally, we would like a *conjugate prior*, and the Gamma distribution happens to be the perfect candidate. So $\theta \sim \text{Gamma}(\alpha, \beta)$ and

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}.$$

Next, we could consider *Jeffrey's Prior*, which—after running through the straightforward calculations—would give us

$$p(\theta) \propto \theta^{-1/2}$$
$$\Rightarrow \theta \sim \text{Gamma}(1/2, 0)$$

which is improper, but will give a proper posterior nonetheless.

Next, if we want to try a *neutral prior*, then the neutral prior is

$$\theta \sim \text{Gamma}(1/3, 0)$$

3.2.3. Posterior Distribution

This will give us a posterior distribution,

$$p(\theta|\mathbf{y}) \propto \theta^{\Sigma y_i + \alpha - 1} e^{-(n+\beta)\theta}$$
$$\Rightarrow \theta|\mathbf{y} \sim \text{Gamma}(\Sigma\ y_i + \alpha\ ,\ n + \beta)$$

3.2.4. Extension: Rate, Exposure Modification

It may be convenient to extend the Poisson model for data points $y_1, \ldots, y_n$ to the form

$$y_i \sim \text{Poisson}(x_i\theta)$$

where θ is the *rate* and x_i *exposure* of the ith unit. Typically, the values x_i are known constants, while θ is some unknown parameter. If we set out our model to accomodate this extension, we get

$$\theta \sim \text{Gamma}(\alpha, \beta)$$
$$p(y|\theta) \propto \theta^{\Sigma y_i} e^{-(\Sigma x_i)\theta}$$
$$\theta|y \sim \text{Gamma}\left(\alpha + \Sigma y_i,\ \beta + \Sigma x_i\right)$$

As an example, suppose θ is the Poisson rate of plane crashes, which is scaled by the x_i representing miles flown in a given year.

3.3. Normal Data (Known Variance)

3.3.1. Likelihood

Let's assume for our likelihood that the random variables, y_i follow a normal distribution where

$$p(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \qquad p(\mathbf{y}|\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \Pi\, e^{-\frac{1}{2\sigma^2}\Sigma(y_i-\mu)^2}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\Sigma(y_i-\mu)^2\right\}$$

and σ is known and fixed.

3.3.2. Conjugate Prior and Resulting Posterior

Next, in choosing our prior for the parameter μ, we will assume that $\mu \sim N(\mu_0, \tau^2)$, and, again, σ is constant. This implies a posterior distribution[4]

$$\mu \sim N(\mu_0, \tau^2) \quad \Rightarrow \quad \mu|\mathbf{y} \sim N\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\,,\ \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

One way to view this is as the prior mean adjusted towards the observed data, $\mathbf{y}$:

$$\mu_{\text{pos}} = \alpha + (\bar{y} - \alpha)\frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}$$

Now, if we want a particularly *uninformative prior*, we can let $\tau \to \infty$, which will give us the Central Limit Theorem. But note, our prior will be improper.

3.3.3. Jeffrey's Prior

Next, if we want *Jeffrey's prior*, we can do the computeations and get

$$p(\mu) \propto 1$$

or the flat prior over the entire real line, which gets us to the Central Limit Theorem-type posterior

$$\mu|\mathbf{y} \sim N(\bar{y}, \sigma^2/n)$$

[4] Clearly, as $n \to \infty$, we get the same result as the Central Limit Theorem.

3.3.4. Posterior Predictive Distribution

To get the *predictive distribution*, we can integrate

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int_{-\infty}^{\infty} p(\tilde{\mathbf{y}}|\mu)p(\mu|\mathbf{y})\ d\mu \tag{6}$$

Now that's a bit of a tough integration, so we'll often compute that by simulation. But that doesn't stop us from summarizing the posterior predictive distribution. Since we know that $p(\tilde{\mathbf{y}}|\mu)$ is normal, and since $p(\mu|\mathbf{y})$ is normal, their product will be too. So we need only find the appropriate mean and variance to say all we need to say.

1. Since the normal is conjugate to itself, we know that this posterior predictive distribution *will* be normal.

2. So we can compute the expectation

$$\begin{aligned} E[\tilde{\mathbf{y}}|\mathbf{y}] &= E[E(\tilde{\mathbf{y}}|\mu)|\mathbf{y}] = E[\mu|\mathbf{y}] \\ &= \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \end{aligned}$$

3. Next, we can compute the variance

$$\begin{aligned} Var[\tilde{\mathbf{y}}|\mathbf{y}] &= E[Var(\tilde{\mathbf{y}}|\mu)|\mathbf{y}] + Var[E(\tilde{\mathbf{y}}|\mu)|\mathbf{y}] \\ &= E[\sigma^2|\mathbf{y}] + Var(\mu|\mathbf{y}) \\ &= \sigma^2 + \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \end{aligned}$$

We now have two alternatives. First, we can approximate the integral in Expression 6 through simulation. Or, alternatively, we can use conjugacy along with the mean and variance which we just derived to sample straight from the normal distribution defined

$$\tilde{y}|y \sim N\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},\ \sigma^2 + \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

3.4. Normal Data (Known Mean, Unknown Variance)

Given a vector of length n, denoted y, of random draws from a normal distribution with unknown variance and known mean μ, we have a likelihood of

$$p(y|\sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$$

The *conjugate prior* is an inverse Gamma distribution,

$$\sigma^2 \sim \text{InvGamma}(\alpha,\beta) \quad \Leftrightarrow \quad \frac{1}{\sigma^2} \sim \text{Gamma}(\alpha,\beta)$$

$$p(\sigma^2) = \frac{\beta}{\Gamma(\alpha)}\,(\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}$$

where the distribution function comes froNon-Informative Priorm a simple transformation of variables.

4. Multiparameter Models

4.1. Univariate Normal Model

Here we begin with one of the most useful models, full stop. We will extend this in many directions after the initial specification considered here.

4.1.1. Likelihood

In this model, θ is a vector of two unknown parameters, $\theta = (\mu, \sigma)$. The likelihood is

$$\begin{aligned} p(\mathbf{y}|\mu,\sigma) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\} \\ &\propto (\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\} \end{aligned}$$

4.1.2. Conjugate Prior and Resulting Posterior

We break the prior density up into two components:

$$\mu|\sigma^2 \sim N\left(\mu_0,\ \frac{\sigma^2}{\kappa_0}\right), \qquad \sigma^2 \sim \text{InvGamma}\left(\frac{\nu_0}{2},\ \frac{\nu_0\sigma_0^2}{2}\right)$$

Note that μ and σ^2 are *a priori* dependent. A high variance term, σ^2, will also induce a high-variance prior distribution for μ. But you need this dependence for conjugacy.

Joint Posterior Distribution: Now if we use this prior with the likelihood, the resulting non-standard joint posterior distribution simplifies to

$$p(\mu,\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\left(\frac{n+\nu_0+1}{2}+1\right)}\ \exp\left\{-\frac{1}{2\sigma^2}\left(\sum(y_i-\mu)^2+\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2\right)\right\}$$

Marginal Posterior Densities: Often, we will be interested in the marginal densities. So let's find the marginal posterior density of σ^2. With a lot of algebra, a nice identity,[5], and completing the square, we get:

$$\begin{aligned} p(\sigma^2|\mathbf{y}) &= \int p(\mu,\sigma^2|\mathbf{y})\,d\mu \\ &= (\sigma^2)^{-\left(\frac{n+\nu_0}{2}+1\right)}\ \exp\left\{-\frac{1}{2\sigma^2}\left(\left[\sum(y_i-\bar{y})^2\right]+\nu_0\sigma_0^2+\frac{nk_0}{n+k_0}(\bar{y}-\mu_0)^2\right)\right\} \\ \sigma^2|\mathbf{y} &\sim \text{InvGamma}\left(\frac{n+\nu_0}{2},\ \frac{1}{2}\left(\left[\sum(y_i-\bar{y})^2\right]+\nu_0\sigma_0^2+\frac{nk_0}{n+k_0}(\bar{y}-\mu_0)^2\right)\right) \end{aligned} \quad (7)$$

This has a nice clear interpretation, where the second parameter is a combination of the sum of squares, $\nu_0\sigma^2$, and the discrepancy between the data mean and the prior mean.

[5]Specifically, we'll use that $\Sigma(y_i-\mu)^2 = \Sigma(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2$.

Next, if we consider the marginal posterior distribution for the mean, we get

$$p(\mu|\sigma^2, \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum(y_i-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right)\right\}$$

$$\mu|\sigma^2, \mathbf{y} \propto N\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{\kappa_0}{\sigma^2}\mu_0}{\frac{n}{\sigma^2}+\frac{k_0}{\sigma^2}},\ \frac{1}{\frac{n}{\sigma^2}+\frac{k_0}{\sigma^2}}\right) \tag{8}$$

To obtain samples from the joint posterior distribution, first sample from σ^2 from the Equation 7 Inverse Gamma distribution.[6] Then, using this value for σ^2, sample from a normal distribution with the parameters from Equation 8. Then repeat this process many times to get samples of $\mu, \sigma|y$.

4.1.3. Non-Informative Prior

Non-Informative Prior: To get this distribution, we can let $\kappa_0, \nu_0 \to 0$, using the notation from the conjugate prior distribution described above. This gives us

$$p(\mu) \propto 1, \qquad p(\sigma^2) \propto (\sigma^2)^{-1}$$

Putting this together, you get a joint prior density of

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \quad \Leftrightarrow \quad p(\mu, \ln\sigma) \propto 1$$

where the second expression indicates the prior is flat on the $\mu, \ln\sigma$ scale.

Posterior: As a result, our posterior distributions will be[7]

$$\sigma^2|\mathbf{y} \sim \text{InvGamma}\left(\frac{n}{2},\ \frac{1}{2}\sum(y_i-\bar{y})^2\right)$$

$$\mu|\sigma^2, \mathbf{y} \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

So we just factored our posterior density into the product of conditional and marginal posterior densities: $p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y)$. So, to sample, just pull from the Inverse Gamma for σ^2, then use it to pull from the normal.

We already saw the marginal posterior density for σ^2 above. Now let's say we want the posterior *marginal* density of μ. We can integrate analytically to get

$$p(\mu|\mathbf{y}) = \int p(\mu, \sigma^2|\mathbf{y})\, d\sigma^2$$

$$\propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-\frac{n}{2}}, \qquad s^2 = \frac{1}{n-1}\sum(y_i-\bar{y})^2$$

[6] Alternatively, you could use the same parameters from Equation 7 to draw from the Gamma distribution, and then take the reciprocal to get a value for σ^2.

[7] The distribution for $\mu|\sigma^2, \mathbf{y}$ follows from the fact that this is equivalent to the one-parameter normal case with known variance

which we recognize to be the good old t-distribution:

$$\mu|\mathbf{y} \sim t_{n-1}\left(\bar{y}, \frac{s^2}{n}\right)$$

4.1.4. Semi-Conjugate Prior

Prior Distribution: This particular prior allows for a priori independence between μ and σ^2. To do so, take

$$\begin{aligned}\mu &\sim N(\mu_0, \tau_0^2)\\ \sigma^2 &\sim \text{InvGamma}\left(\frac{\nu_0}{2},\ \frac{\nu_0\sigma_0^2}{2}\right)\end{aligned}$$

Posterior: This will give us a joint posterior distribution of

$$\begin{aligned}p(\mu, \sigma^2|y) = (\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\bar{y})^2\right\}\exp\left\{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\}\\ \times(\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)}\exp\left\{\frac{-\nu_0\sigma_0}{2\sigma^2}\right\}\end{aligned}$$

which doesn't really simplify any further to something nice or standard. But, if we consider the conditional posterior density for μ, we do get something familiar:

$$\mu|\sigma^2, y \sim N\left(\frac{\frac{n}{\sigma^2}\bar{y}+\frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}},\ \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}\right)$$

But in order to use this, we need a value of σ^2 to condition on.

Marginal Posterior Density for σ^2: Next, we can integrate out μ from the joint posterior. This yields

$$\begin{aligned}p(\sigma^2|y) \propto \left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{1/2}(\sigma^2)^{-\left(\frac{n+\nu_0}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu_0)^2\right\}\\ \times\exp\left\{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\}\exp\left\{\frac{-\nu_0\sigma_0}{2\sigma^2}\right\}\end{aligned}$$

4.2. Multivariate Normal Model

Suppose y_i is a vector of length d with multivariate normal distribution,

$$y_i|\mu, \Sigma \sim N(\mu, \Sigma)$$

where μ is a column vector of length d and Σ is a $d \times d$ matrix. Then the likelihood function is

$$p(y_i|\mu, \Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\right)$$

Now suppose that we have n such random variables. Then the joint likelihood is written

$$p(y_1, \ldots, y_n|\mu, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\right)$$

5. Bayesian Regression

In this section, I'll let y be either a single observation or an $n \times 1$ vector, which will be clear from the context. To help establish that context, I'll let x_i denote covariates for a *single* observation, while X will denote a *matrix* of covariates with dimension $n \times p$ for n observations and p covariates.

5.1. Likelihood for Simple and Multiple Regression

In the simplest case, we will extend the simple $N(\mu, \sigma^2)$ model to allow for conditional means, as in simple regression. Therefore, our model for the ith observation will be

$$y_i = \beta x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

where x_i is a vector of p known covariates, and β is a p-dimensional coefficient vector. Typically, we'll want to estimate β and σ^2. In our notation, we can even be a bit broader and write

$$\begin{aligned} y &= X\beta + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 I_n) \\ y|\beta, X, \sigma^2 &\sim \mathrm{N}(X\beta, \sigma^2 I_n) \end{aligned}$$

where y is an $n \times 1$ vector of observations and X is an $n \times p$ matrix for n observations and p covariates. This setup also implies errors that are independent and have equal variance.

5.2. Flat Prior

In the case of the flat, non-informative prior, we recall the multiparameter normal case and set

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-1}.$$

Now, recycling and adapting everything we did above, this we'll give us a joint posterior distribution of

$$\begin{aligned} p(\beta, \sigma^2|y) &\propto p(y|\beta, \sigma^2)p(\beta, \sigma^2) \\ &\propto (\sigma^2)^{-1}(\sigma^2)^{-\frac{n}{2}} \exp\left\{\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\} \\ &\propto (\sigma^2)^{-\left(\frac{n+2}{2}\right)} \exp\left\{y^T y - 2y^T X\beta + \beta^T X^T X\beta\right\} \end{aligned}$$

Now, let's say we want to compute the conditional posterior density for β. In that case, any unnecessary σ or y terms can be considered constants and wash out. This allows us

to rewrite the last line, make substitutions, and complete the square to get[8]

$$\begin{aligned} & & p(\beta|\sigma^2, y) &\propto \exp\left\{\beta^T X^T X \beta - 2y^T X \beta\right\} \\ V_\beta &= (X^T X)^{-1} & & \\ \hat{\beta} &= (X^T X)^{-1} X^T y \qquad \Rightarrow & &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta^T V_\beta \beta - 2\hat{\beta}^T V_\beta \beta)\right\} \\ & & &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta}) V_\beta^{-1} (\beta - \hat{\beta})\right\} \end{aligned}$$

which happens to be the density of a multivariate normal distribution. This means

$$\beta|\sigma^2, y, X \sim \text{MVN}_p(\hat{\beta},\ \sigma^2 V_\beta)$$

Now we have to handle the conditional posterior density of σ^2:

$$\begin{aligned} p(\sigma^2|y, X) &= \frac{p(\beta, \sigma^2|y)}{p(\beta|\sigma^2, y)} \\ &\propto \frac{(\sigma^2)^{-\left(\frac{n+2}{2}\right)} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\right\}}{(\det|\sigma^2 V_\beta|)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\ V_\beta^{-1}\ (\beta - \hat{\beta})\right\}} \end{aligned}$$

Since β appears nowhere in the LHS, we can plug in anything, like, say, $\beta = \hat{\beta}$, and is guaranteed to cancel. This simplifies to

$$\begin{aligned} p(\sigma^2|y, X) &\propto (\sigma^2)^{-\left(\frac{n-p}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i - x_i\beta)^2\right\} \\ \Rightarrow \quad \sigma^2|y, X &\sim \text{InvGamma}\left(\frac{n-p}{2},\ \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right) \end{aligned}$$

Like the multiparameter normal with the non-informative prior, the marginal posterior density is easy enough to get through integration

$$\begin{aligned} p(\beta|y, X) &= \int_0^\infty p(\beta, \sigma^2|y, X)\ d\sigma^2 \\ &\propto \left[1 + \frac{1}{\sum(y_i - x_i\beta)}(\beta - \hat{\beta})'\ V_\beta^{-1}\ (\beta - \hat{\beta})\right]^{-\frac{n}{2}} \end{aligned}$$

which is a *scaled multivariate t-distribution.*

Next, we can ask about the *posterior predictive distribution* which is simple enough to integrate analytically and get:

$$\begin{aligned} p(\tilde{y}|\tilde{X}, y, X) &= \int\int p(\tilde{y}|\tilde{X}, \beta, \sigma^2) p(\beta|\sigma^2, y, X) p(\sigma^2|y, X)\ d\beta\ d\sigma^2 \\ \Rightarrow \quad \tilde{y}|\tilde{X}, y, X &\sim \text{MVT}_{n-p}(\tilde{X}\hat{\beta},\ S^2(I + \tilde{X} V_\beta \tilde{X}')) \end{aligned}$$

[8]The substitution $\hat{\beta} = (X'X)^{-1}Xy$ is the standard MLE estimate for β in regression.

5.3. Informative Prior with iid $\beta_j \in \beta$

Now let's say we want to build some more information into our coefficient vector, β. Specifically, we can model each component of the beta vector, $\beta_j \in \beta$, to be iid and distributed

$$\beta_j \sim N(\beta_{0j}, \tau^2).$$

As before, we'll take a flat, non-informative prior for the variance of the error terms:

$$p(\sigma^2) \propto (\sigma^2)^{-1}$$

where β and σ^2 are independent.

This gives a posterior conditional distribution for β of

$$\begin{aligned} p(\beta|y, X, \sigma^2) &\propto p(y|\beta, \sigma^2, X)p(\beta, \sigma^2) \\ &\propto p(y|\beta, \sigma^2, X)p(\beta)p(\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2}(y^* - X^*\beta)\ \Sigma^{-1}(y^* - X^*\beta)\right\} \end{aligned}$$

$$y^* = \begin{pmatrix} Y \\ \beta_0 \end{pmatrix}, \qquad X^* = \begin{pmatrix} X \\ I_p \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \tau^2 I_p \end{pmatrix}$$

This implies for the conditional posterior distribution of β that

$$\beta|y, X, \sigma^2 \sim \text{MVN}_p(\hat{\beta}, V_\beta)$$

$$\hat{\beta} = (X^{*'}\Sigma X^*)^{-1}X^{*'}\Sigma^{-1}y^*, \qquad V_\beta = (X^{*'}\Sigma^{-1}X^*)^{-1}$$

5.4. Extensions

Above, we took a very simple approach to the error terms in the likelihood, which gave us independent errors with equal variance. But we can generalize this a bit and allow correlation between error terms, and, thus, observed y by specifying instead:

$$y \sim \text{MVN}(X\beta,\ \sigma^2 I_n) \quad \Rightarrow \quad y \sim \text{MVN}_n(X\beta,\ \Sigma_y).$$

Similarly, we could extend our prior for β, where we assumed that all of the components $\beta_j \in \beta$ were iid. Instead, we could assume a distribution for the entire coefficient vector

$$\beta_j \sim N(\beta_{0j}, \tau^2) \quad \Rightarrow \quad \beta \sim \text{MVN}_p(\beta_0,\ \Sigma_\beta)$$

In both of these cases, however, things become substantially more complicated because we have to pre-specify covariance matrices, Σ_y and Σ_β. In practice, this is pretty tough to do a priori, unless we have a look at the data before we estimate parameters.

5.5. Optimization

In classical regression, to estimate $\hat{\beta}_{\text{MLE}}$, we will solve

$$\hat{\beta}_{\text{MLE}} = \arg\max_{\beta} p(y|\beta, X), \quad \Leftrightarrow \quad \hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

However, in Bayesian regression, the problem becomes

$$\hat{\beta} = \arg\max_{\beta} p(y|\beta, X)\ p(\beta), \quad \Leftrightarrow \quad \hat{\beta} = \arg\max_{\beta} \ln p(y|\beta, X) + \ln p(\beta)$$

$$\Rightarrow \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 - \ln p(\beta)$$

where the log of the prior acts as a penalty term. We'll examine this in more detail through an example.

Example *Ridge Regression (L2 Regression)*: Now let's consider the case where

$$\beta \sim \text{MVN}_p(\beta_0,\ \tau^2 I_p).$$

In models with many parameters (p large), it's common to set a prior where the coefficient vector is assumed to be the zero vector, so that our optimization problem becomes

$$\ln p(\beta) = K - \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \qquad \lambda = \frac{1}{2\tau^2}$$

Now it's clear to see how the $\ln p(\beta)$ acts as a penalty term. Large values of β_j are penalized more in the minimization, leading the β_j estimates to be shrunk back towards 0.

But we can make the penalty even more extreme by changing the λ term, which brings us to the next section.

5.6. Optimization for Non-Conjugate Priors

Laplace Prior and Lasso (or L2) Regression: We set a prior as follows leading to a corresponding minimization problem:

$$p(\beta_j) \propto e^{-\lambda|\beta_j|}$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The interesting thing about the Laplace-based model is that many $\hat{\beta}_j$ will be set *identically* to 0 because of the cusp that occurs in the prior density.

More General Priors and Bridge Regression: We can be even more general and set

$$p(\beta) \propto \exp\{-\lambda \mathcal{J}(\beta)\}$$

$$\Rightarrow \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \mathcal{J}(\beta)$$

Often, it's very common to restrict $\mathcal{J}(\beta)$ to

$$\mathcal{J}(\beta) = \sum_{j=1}^{p} |\beta_j|^q$$

which retains as special cases Lasso regression ($q = 1$) and Ridge regression ($q = 2$).

6. Mixture Models

Let's first discuss the idea of Mixture Models. If we set up a mixture model, then we assume that any given observation, $y_i \in y$, can come from one of k standard distributions. Each observation also has a certain probability of being pulled from any one of the k distributions. So to begin and nail down the intuition, let's begin with the simple two-parameter case.

6.1. Likelihood as a Mixture of Two Models

We assume that an observation can come from one of two distributions, A or B:

$$y_i \sim \begin{cases} A(\theta_A) & \text{with probability } \alpha \\ B(\theta_B) & \text{with probability } 1-\alpha \end{cases}$$

Let's assume that the probability density functions for A and B are denoted

$$\phi_A(y_i), \qquad \phi_B(y_i)$$

Then we can write the likelihood of a single observation and then of the data

$$p(y_i|\theta_A, \theta_B) = \alpha\phi_A(y_i) + (1-\alpha)\phi_B(y_i)$$
$$p(y_i|\theta_A, \theta_B) = \prod_{i=1}^{n} [\alpha\phi_A(y_i) + (1-\alpha)\phi_B(y_i)]$$

Now clearly, this is a totally non-standard distribution (most-likely). And on top of that, it's going to be really tough to estimate parameters. So to do that, we'll typically employ the technique of the EM algorithm to estimate parameters.

6.2. Expectation Maximization (EM) Algorithm

This method employs "missing data" in the sense of data points or characteristics of the data y that, *if* we had it, it would make our parameter estimation much easier. One example, let's say we had an indicator variable

$$I_i = \begin{cases} 1 & \text{if } y_i \text{ is in group } A \\ 0 & \text{if } y_i \text{ is in group } B \end{cases}$$

Complete Data Likelihood Step: From there, we can write out the *complete data likelihood* as a function of the observed data, the missing data, and the unknown parameters:

$$p(y, I|\theta_A, \theta_B) = \prod_{i=1}^{n} [\phi_A(y_i)]^{I_i} [\phi_B(y_i)]^{1-I_i} \alpha^{I_i}(1-\alpha)^{1-I_i}$$
$$\ell(y, I|\theta_A, \theta_B) = \sum_{i=1}^{n}(I_i) \ln \phi_A(y_i) + \sum_{i=1}^{n}(1-I_i) \ln \phi_B(y_i)$$
$$+ \sum_{i=1}^{n} I_i \ln \alpha + \sum_{i=1}^{n}(1-I_i) \ln(1-\alpha)$$

Expectation Step: Now it would be great if we could optimize the last log-likelihood on the last line, bu we don't *actually* have the values for I. Since we can't plug in actual values for each I_i, we can at least plug in the expected values *given* the observed data and the current values of the parameters.

$$\begin{aligned}\hat{I}_i = E[I_i|y,\theta] &= P(I_i = 1|y,\theta) \\ &= \frac{P(y_i|I_i = 1,\theta)P(I_i = 1|\theta)}{p(y_i|I_i = 1,\theta)P(I_i = 1|\theta) + P(y_i|I_i = 0,\theta)P(I = 0|\theta)}\end{aligned}$$

Maximization Step: Next, we plug $\hat{I}_i$ into the complete data likelihood and find the optimal estimates, $\hat{\theta}$, for all of our parameters, θ. We do that by plugging into the log-likelihood, taking the partial derivative with respect to all the parameters, setting each equation equal to zero, and solving out.

Having calculated the new parameter estimates, calculate the new $\hat{I}_i$ and iterate until convergence.

7. Hierarchical Models

7.1. General Form

Prior and Posterior: Hierarchical Models introduce difference levels of analysis. A common example is to have some likelihood $p(y|\theta)$ for observing the data, but have the parameter θ come from its own distribution, $p(\theta|\phi)$, where ϕ. Often, we'll take advantage of conditional independence in such models to simplify posterior analysis. For example:

$$\begin{aligned} p(\theta, \phi|y) &\propto p(y|\theta, \phi) \cdot p(\theta, \phi) \\ &\propto p(y|\theta) \cdot p(\theta|\phi) \cdot p(\phi) \end{aligned} \tag{9}$$

We see above that the prior distribution, which now must be specified for both θ and ϕ, was broken apart into $p(\theta, \phi) \propto p(\theta|\phi) \cdot p(\phi)$.

Conditional and Marginal Distributions: We saw above in Equation 9 how to write the joint posterior. Typically, we will break this up into a marginal and conditional posterior:

$$p(\theta, \phi|y) \propto p(\theta|\phi, y) \cdot p(\phi|y)$$

This method instructs us how to simulate from the posterior: draw ϕ from it's marginal posterior distribution ($p(\phi|y)$), then sample θ from it's marginal conditional distribution ($p(\theta|\phi, y)$).

Yet while it's often easy to get $p(\theta|\phi, y)$, we'll have to do a bit more work to get $p(\phi|y)$, whether that's brute force integration or algebraically:

$$p(\phi|y) \propto \int p(\theta, \phi|y)\, d\theta, \qquad \text{or} \qquad p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}$$

The second method works well for some nicely behaved distributions (but not always).

Note When possible, avoid assigning uniform prior distributions to the logarithm of the standard deviation (or approximate standard deviation) of the distribution on ϕ. That is,

$$\text{Don't set} \quad p(\ln \sigma_\phi) \propto 1$$

It can lead to an improper posterior density on ϕ—that is, $p(\phi|y) \propto \infty$. Rather, set a uniform prior on the standard deviation parameter itself.

7.2. Normal Model

For first model, we assume that there are J groups, each with their own mean (μ_j) but shared variance, where the μ_j are drawn from a common distribution

$$\begin{aligned} y_{ij} &\sim N(\mu_j, \sigma^2), \qquad i = 1, \ldots, n_j \\ \mu_j &\sim N(\mu_0, \tau^2), \qquad j = 1, \ldots, J \end{aligned}$$

We can think of σ^2—which we assume to be known—as the *within group* variance, while τ^2 is *between group*, which controls between-group sharing of information and the amount of shrinkage back to μ_0.

General Form of the Posterior: Let's examine how conditional independent allows us to simplify the posterior, keeping in mind that μ is a vector of the individual μ_j:

$$\begin{aligned} p(\mu, \mu_0, \tau^2|y) &\propto p(y|\mu, \mu_0, \tau^2) \cdot p(\mu, \mu_0, \tau^2) \\ &\propto p(y|\mu) \cdot p(\mu|\mu_0, \tau^2) \cdot p(\mu_0, \tau^2) \end{aligned}$$

Posterior Given Non-Informative Prior: Suppose we choose $p(\mu_0) \propto 1$ and $p(\tau^2) \propto 1$. Then we get for our posterior

$$\begin{aligned} p(\mu, \mu_0, \tau^2|y) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} (\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_{ij} - \mu_j)^2 \right\} \\ \times \prod_{j=1}^{J} (\tau^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\tau^2}(\mu_j - \mu_0)^2 \right\} \end{aligned}$$

Breaking Up the Posterior: We can analyze the conditional distributions of the parameters and get standard distributions. In fact, simple calculations will show that

$$\begin{aligned} \mu_j|\mu_0, \tau^2, y &\sim N\left(\frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu_0}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \ \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \right), \\ \mu_0|\tau^2, y &\sim N\left(\hat{\mu}, \ V_\mu\right), \\ \tau^2|y &\sim V_\mu^{1/2} \prod_{j=1}^{J} \left(\frac{\sigma^2}{n_j} + \tau^2 \right)^{1/2} \exp\left\{ -\frac{(\bar{y}_j - \hat{\mu})^2}{2\left((\sigma^2/n_j) + \tau^2\right)} \right\} \end{aligned}$$

where we have

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \qquad \hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{(\sigma^2/n_j)+\tau^2} \bar{y}_j}{\sum_{j=1}^{J} \frac{1}{(\sigma^2/n_j)+\tau^2}}, \qquad V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{(\sigma^2/n_j)+\tau^2}}$$

The only distribution that is nonstandard is $\tau^2|y$, which will require grid sampling.

Posterior Predictive Sampling: There are two ways we can take posterior predictive draws, depending on our goal:

1. Samples from currently existing group: In this case, we consider what would happen if there were more observations in group j. Here's how we implement:
   a) Draw τ^2 from it's posterior distribution using grid sampling.
   b) Draw μ_0 from the conditional posterior distribution of $\mu_0|\tau^2, y$.
   c) Sample μ_j—for the group j you are considering—from $p(\mu_j|\mu_0, \tau^2, y)$ using the values of μ_0 and τ^2 just obtained.
   d) Sample from the distribution for $\tilde{y} \sim N(\mu_j, \sigma^2)$, where μ_j is the value just obtained.
   e) Repeat the process.
2. Samples from entirely new group $\tilde{j}$: Here, we consider what would happen if we introduce a new group entirely. This will get a very diffuse posterior predictive distribution.
   a) Draw τ^2 from it's posterior distribution using grid sampling.
   b) Draw μ_0 from the conditional posterior distribution of $\mu_0|\tau^2, y$.
   c) Sample $\tilde{\mu}$ from $\mathrm{N}(\mu_0, \tau^2)$ using the μ_0 just obtained.
   d) Sample $\tilde{y} \sim \mathrm{N}(\tilde{\mu}, \sigma^2)$.
   e) Repeat the process.

7.3. Normal Model with the EM Algorithm

We saw the EM Algorithm with Mixture Models; however, this technique, where we suppose that there is "missing data," can also useful in Hierarchical Models. Let's return to the normal model with known variance:

$$\begin{aligned} y_{ij} &\sim N(\mu_j, \sigma^2) \\ \mu_j &\sim N(\mu_0, \tau^2) \end{aligned}$$

In the last section, we considered μ, μ_0 and τ^2 as our unknown parameters, but suppose we consider only (μ_0, τ^2) as unknown and let all of the components $\mu_j \in \mu$ be "missing data." Then we could say that the estimation of μ_0 and τ^2 would be really easy if we knew the μ_j. This leads us to the EM algorithm:

1. Expectation Step: We want to compute the expectation of the complete data log-likelihood. This differs a bit from the Expectation Step in the Mixture Model implementation, but it's a straightforward generalization.

   $$Q(y_{\text{obs}}|\theta) = E_{y_{\text{mis}}}\left[\ell(y_{\text{obs}}, y_{\text{mis}}|\theta)\right]$$

   In our normal example, this means

   $$\begin{aligned} p(y, \mu|\mu_0, \tau^2) &\propto \prod_{j=1}^{m} (\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\mu_j - \mu_0)^2\right\} \\ &\quad \times \prod_{j=1}^{m}\prod_{i=1}^{n_j} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \mu_j)^2\right\} \\ \Rightarrow \quad Q(y_{\text{obs}}|\theta) &= E[\ell(y, \mu|\mu_0, \tau^2)] = E[\ln p(y, \mu|\mu_0, \tau^2)] \end{aligned}$$

2. Maximization Step: Next, we calculate the $\hat{\theta}$ that maximizes $Q(y_{\text{obs}}|\theta)$

7.4. Binomial Hierarchical Model

Let's suppose our vector observations, denoted y, is binomially distributed. Moreover, each observation, y_i, has its own parameter θ_i drawn from a distribution—this is the hierarchical part. So together, we have

$$\begin{aligned} y_i &\sim \text{Binom}(n_i, \theta_i) \\ \theta_i &\sim \text{Beta}(\alpha, \beta). \end{aligned}$$

Let's specify the generic posterior, taking advantage of the conditional independence built into the hierarchical model, noting that θ is a vector of θ_i:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(y|\theta, \alpha, \beta) \cdot p(\theta, \alpha, \beta) \\ &\propto p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta) \end{aligned}$$

We can substitute in for the particular distributions, leaving our prior distribution on α and β unspecified:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto \left[\prod_{i=1}^{n} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i}\right] \cdot \left[\prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\right] \cdot p(\alpha, \beta) \\ &\propto p(\alpha, \beta) \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]^n \prod_{i=1}^{n} \theta_i^{y_i+\alpha-1}(1-\theta_i)^{n_i - y_i + \beta - 1} \end{aligned}$$

Let's get the conditional posterior distribution of $\theta|\alpha, \beta, y$:

$$\begin{aligned} p(\theta_i|\alpha, \beta, y) &\propto \theta_i^{y_i+\alpha-1}(1-\theta_i)^{n_i - y_i + \beta - 1} \\ \theta_i|\alpha, \beta, y &\sim \text{Beta}(y_i + \alpha,\ n_i - y_i + \beta) \end{aligned}$$

Now for the posterior distribution of α and β we'll need to make a choice for our prior. We'd probably like to do something non-informative, like $p(\alpha, \beta) \propto 1$, but that would lead to an improper posterior. Gelman suggests that we use $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$, which is justified in Exercise 5.7 in the book. This leads to

$$\begin{aligned} p(\alpha, \beta|y) &\propto \int p(\theta, \alpha, \beta|y)\, d\theta \\ &\propto (\alpha+\beta)^{-5/2} \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]^n \int \prod_{i=1}^{n} \theta_i^{y_i+\alpha-1}(1-\theta_i)^{n_i - y_i + \beta - 1}\, d\theta_i \end{aligned}$$

This looks like a beast of an integral, but notice that what's inside the integral is the beta distribution, unnormalized by a constant. So it's value must be one divided by that normalizing constant,[9] which we can write

$$p(\alpha, \beta|y) \propto (\alpha+\beta)^{-5/2} \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]^n \prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha)\Gamma(n_i - y_i + \beta)}{\Gamma(n_i + \alpha + \beta)}$$

To implement, we sample α and β using grid sampling, then sample each $\theta_i|\alpha, \beta$.

[9] Alternatively, we could have computed $p(\alpha, \beta|y)$ by the other method: $p(\alpha, \beta|y) = p(\theta, \alpha, \beta|y)/p(\theta|\alpha, \beta|y)$, which would have given us the same result (and is easy to compute).

8. Markov Chain Monte-Carlo (MCMC) Algorithms

Markov Chain Monte-Carlo Algorithms allow us to sample from the posterior distribution of non-standard distributions. We saw grid-sampling earlier as a way to do so, but also noted the limitations when we have posteriors with great than two parameters.

We also saw Newton Raphson and the EM algorithm as ways to estimate parameters, but they only give point estimates as opposed to full posterior distributions. MCMC algorithms, which is a *stochastic* optimization algorithm, attempt to overcome these difficulties by sampling sequentially from a variety of univariate distributions.

8.1. Gibbs Sampler

Our goal is to obtain samples for $\theta|y$, where θ is a vector of parameters, $(\theta_1, \ldots, \theta_k)$. Here's the procedure:

1. Start with a set of arbitrary values for the parameter
$$\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$$

2. Next, we sample iteratively from conditional posterior distributions of each parameter—given current values of other parameters—as follows:
$$\begin{aligned}
\theta_1^{(t+1)} &\sim p(\theta_1|\theta_2^{(t)}, \ldots, \theta_k^{(t)}, y) \\
\theta_2^{(t+1)} &\sim p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, y) \\
&\vdots \qquad\qquad \vdots \\
\theta_k^{(t+1)} &\sim p(\theta_k|\theta_1^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}, y)
\end{aligned}$$

   As noted above, the conditional distributions will often be standard distributions that are relatively simple since we can absorb $k-1$ parameters into the $\propto$ each time we need to get a draw for θ_i.

3. We iterate through this, and eventually the distribution will converge to a range of values for each θ_i. But the question is "How exactly do we know when convergence occurs?"

   To answer this question, we typically run multiple Markov chains from multiple starting values, $\theta^{(0)}$. Once the chains multiple chains start moving together, we declare convergence and remove the "burn-in" before the chains converged.

4. After we remove the burn-in, we still have the problem of autocorrelated samples. Since the Gibbs Sampler generates a Markov chain, autocorrelation is baked into the procedure. So to obtain independent samples, we will need to evaluate the autocorrelation of our chains, "thin" them by only taking every mth value or so, where m is chosen so that there is negligible autocorrelation in the resulting sample.

8.2. Metropolis Algorithm

The Metropolis algorithm is an MCMC method that uses rejection sampling to draw from a posterior distribution. This is particularly when the conditional posterior distributions are non-standard, so that the Gibbs sampler has trouble. Here are the steps involved:

1. Initialize at a starting point, θ^0. Then for each $t \in \{1, 2, \ldots\}$, sample a *proposal* value, which I'll denote θ^*, from some proposal or "jumping" distribution, $g(\theta^*|\theta^{t-1})$.[10] Typically, $g(\cdot|\cdot)$ will be a standard distribution.

2. Form the ratio of the densities

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} \tag{10}$$

3. Then, once we have a value for r, we set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min\{r, 1\}. \\ \theta^{t-1} & \text{otherwise} \end{cases} \tag{11}$$

   Intuitively, we accept a draw θ^* if it increases the posterior density. If not, we randomly accept it, but only with probability proportional to how good it is. In this way, this method acts like a stochastic stepwise mode-finding algorithm.

Note, for the Metropolis algorithm, we *require* that the proposal (or "jumping") distribution, $g(\cdot|\cdot)$, is symmetric, i.e.

$$g(\theta^*|\theta^{t-1}) = g(\theta^{t-1}|\theta^*), \qquad \forall \theta^*, \theta^{t-1}, t$$

The common and convenient normal distribution obeys this rather nice property and can be used with this algorithm.

8.3. Metropolis-Hastings Algorithm

The *Metropolis-Hastings Algorithm* is a direct generalization of the Metropolis algorithm. It relaxes the restrictive requirement that $g(\cdot|\cdot)$, our proposal distribution, be symmetric.

Except for one step, the algorithm functions *exactly* the same way as the Metropolis Algorithm. We only have to modify the "attractiveness" ratio, r, given by Equation 10. Here's the new ratio for the new Metropolis-Hastings Algorithm:

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} \cdot \frac{g(\theta^{t-1}|\theta^*)}{g(\theta^*|\theta^{t-1})} \tag{12}$$

We accept or reject draws θ^* from $g(\cdot|\cdot)$ just as before, in Expression 11.

[10] The dependence on the previous value, θ^{t-1}, makes our method an Mark Chain Monte Carlo method.

8.4. Gibbs Sampler as a Special Case of the M-H Algorithm

The Gibbs sampler can be viewed as a special case of the more general Metropolis-Hastings Algorithm. To see why, suppose we *could* draw directly from the posterior distribution, $p(\theta|y)$, that we are targeting with our Metropolis-Hastings algorithm. (Obviously unnecessary to use the M-H Algorithm in that case, but play along).

Then we can take the posterior as our proposal distribution, $g(\theta^*|\theta^{t-1}) = p(\theta^*|y)$. In that case, the ratio in Equation 12 will *always* be one, so that every proposed θ^* is accepted as in the Gibbs sampler.

9. EM Algorithm

Suppose we hope to estimate the posterior, $p(\phi|y)$, for a vector of parameters, ϕ, despite the posterior distribution being nonstandard. Then suppose we introduce some "missing data," denoted γ, in the hope that knowing γ would make the estimation of the posterior of ϕ easier.

More explicitly, it may be tough to find $p(\phi|y)$, but by introducing γ, we may be able to work with $p(\gamma|\phi, y)$ and $p(\phi|\gamma, y)$. In particular,

1. Start with a guess for ϕ and γ.
2. Replace the components of γ with their conditional expectation, $E[\gamma|\phi]$.
3. Maximize the posterior likelihood given γ, which is denoted $p(\phi|\gamma, y)$, and replace ϕ with the maximized values.
4. Repeat Steps 1 and 2 until convergence.

A. Finding Parameters for and Drawing from Non-Standard Distributions

Noting that some of the resulting posterior distributions won't necessarily be standard, it's important to know how to find parameters and get probabilities from these so-called *non-standard distributions*. The simplest is the grid method, and while easy to do from a computational standpoint, it can be difficult when it comes to selecting a proper grid initially.

A.1. Grid Method in One Dimension

Suppose we want to get samples from a non-standard distribution. Then we follow this process to generates probabilities and then samples for θ.

1. Pick a grid of possible values for θ. A fair amount of trial and error will be involved here.[11]
2. Calculate $m(\theta) = \frac{p(\theta,y)}{\sum p(\theta,y)}$, where the sum is the sum over the grid. This generates rough approximations of the true probabilities.
3. Sample one grid value with probabilities proportional to $m(\theta)$.
4. Repeat the previous step many times to generate a large sample.

A.2. Grid Method in Two Dimensions

Here, suppose there are two parameters which define a non-standard distribution. We decompose the grid method into two parts:

1. Pick a grid of values for α and β, our two parameters. Again, this will probably take some trial and error.
2. Calculate the probabilities $m(\alpha, \beta) = \frac{p(\alpha,\beta|y)}{\sum p(\alpha,\beta|y)}$.
3. Calculate marginal and conditional posterior probabilities:

$$m(\alpha) = \sum_{\beta} m(\alpha, \beta), \qquad m(\beta|\alpha) = \frac{m(\alpha, \beta)}{m(\alpha)}$$

4. Sample a grid value α_i with probability proportional to $m(\alpha)$.
5. Based on our sampled value of α_i, sample grid value β_i with probability proportional to $p(\beta|\alpha_i)$.
6. Repeat the last couple steps many times to get distributions.

[11] You choose the value of θ that maximize the likelihood or (more often) log-likelihood.

A.3. Newton's Method in One Dimension

Suppose we have some function $f(x)$, and we want to find a root of this function. Then we start with a linear approximation to the Taylor's Series expansion of $f(x)$ about some starting point, x_k:

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + O(x^2).$$

Since we want to find a root, we set the last equation equal to zero and solve for a new guess

$$\begin{aligned} f(x) = 0 &= f(x_k) + f'(x_k)(x - x_k) + O(x^2) \\ \Rightarrow x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)} \end{aligned}$$

Iterating this process, we can get better and better approximations of the root of $f(x)$.[12]

Now the question is how do we apply this to finding parameters? Well that's quite simple. Instead of trying finding a root for $f(x)$, we typically try to maximize the likelihood—which involves finding a root for $\ell'(\theta)$, the derivative of the log-likelihood. So let's go through the steps to find a parameter:

$$\begin{aligned} & \hat{\theta} = \arg\min_\theta L(\theta) = \arg\min_\theta \ell(\theta) \\ & 0 = \ell'(\hat{\theta}) \qquad \text{Solve for } \hat{\theta} \\ \text{Newton Method} \qquad & 0 = \ell'(\theta_k) + \ell''(\theta_k)(\theta_{k+1} - \theta_k) + O(\theta^2) \\ & \theta_{k+1} = \theta_k - \frac{\ell'(\theta_k)}{\ell''(\theta_k)} \end{aligned}$$

Again, iterating over this process, we can get better and better estimates of $\hat{\theta}$.[13]

A.4. Newton's Method in Multiple Dimensions

Again, let's start with the general case, then move to our special case of finding parameters. So we have a vector valued function, and want to expand it into a Taylor Series approximation.:

$$f(x + h) = f(x) + f'(x)h + O(h^2)$$

$$x = (x_1, \ldots, x_n)^T, \qquad h = (h_1, \ldots, h_n)^T, \qquad f(x) = (f_1(x), \ldots, f_n(x))^T$$

$$f'(x) = J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

[12] Note that there might not be a single root. In fact the Newton-Raphson algorithm is very sensitive to the starting point.

[13] The same cautions about the sensitivity to starting point still apply.

Then, you use

$$0 = f(x) + Jh$$
$$x_{k+1} = x_k - J^{-1} f(x_k)$$

Now let's adopt this to our problem of a two-parameter optimization. Suppose we are faced with

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta} L(\alpha, \beta) = \arg\max_{\alpha,\beta} \ell(\alpha, \beta)$$
$$\text{Solve} \qquad 0 = \frac{\partial \ell(\hat{\alpha}, \hat{\beta})}{\partial \alpha}, \qquad 0 = \frac{\partial \ell(\hat{\alpha}, \hat{\beta})}{\partial \beta} \tag{13}$$

Awesome, so we have a vector valued function that we wish to find a root for. So let's do it:

$$\ell(\alpha, \beta) = \begin{pmatrix} \frac{\partial \ell(\alpha,\beta)}{\partial \alpha} & \frac{\partial \ell(\alpha,\beta)}{\partial \beta} \end{pmatrix}$$
$$J = \ell'(\alpha, \beta) = \begin{pmatrix} \frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha^2} & \frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha \partial \beta} & \frac{\partial^2 \ell(\alpha,\beta)}{\partial \beta^2} \end{pmatrix}$$

From there, we get an estimate by iteration over

$$(\alpha_{k+1}, \beta_{k+1}) = (\alpha_k, \beta_k) - J^{-1} \ell(\alpha_k, \beta_k)$$

For more general cases, where the number of parameters is greater than one or two, simply expand the number partial derivatives and maximization criteria that we labeled Equations 13. The only difference is that the Jacobian and function of the parameters would have a higher dimension.

A.5. Gradient Descent

Again, suppose we want to find point estimates of the components in a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ by minimizing some cost function, $\mathcal{J}(\theta)$.[14] Then we update our parameters all at once as follows:

$$\begin{aligned}
\theta_1^{(t+1)} &= \theta_1^{(t)} - \alpha \frac{\partial}{\partial \theta_1^{(t)}} \left[\mathcal{J}(\theta) \right] \\
\vdots \qquad & \qquad \vdots \\
\theta_n^{(t+1)} &= \theta_n^{(t)} - \alpha \frac{\partial}{\partial \theta_n^{(t)}} \left[\mathcal{J}(\theta) \right]
\end{aligned}$$

where α is the learning rate. This parameter controls the size of our steps and influences the aggressiveness of our descent. We keep updating and iterating through this process until convergence.

Moreover, notice how this process self-adjusts in the magnitude of the update. In particular, as you get close to the minimum, the derivatives approach zero. So even with α constant, the shrinking derivative means that the gradient descent algorithm will take smaller steps.

Note At each iteration, we update all the components of θ *simultaneously*. That is, the equations above are a package deal. We don't update one parameter at a time as we did with Gibbs Sampling. Instead, we take all the partial derivatives of the cost function, $\mathcal{J}(\theta)$, then plug into the RHS of the above expressions.

[14] We can easily make the necessary changes if we want to maximize. In particular, we can minimize the negative of some target function. (The name "cost function" isn't really appropriate if we're trying to maximize.)

B. Parametric Bootstrap

In this method, which is in the realm of classical rather than Bayesian statistics, you have some data $\mathbf{y}$, which you use to estimate a parameter θ using Maximum Likelihood Estimation. Using this θ_{MLE}, you then simulate a new random sample of data using the likelihood,

$$p(\mathbf{y}|\theta_{\text{MLE}})$$

C. Jacobian Transformation

This method is useful for determining the probability distribution of some variable $\phi = h(\theta)$,[15] especially when we know the probability distribution of θ, $p(\theta)$. Then we apply

$$f(\phi) = f(\theta)\left|\frac{d\theta}{d\phi}\right|$$

where f is the pdf and the term following $f(\theta)$ is known as the *Jacobian*.

Proof. Just to sketch the proof, let's start with what we know: $f(\theta)$ and $\phi = h(\theta)$. We want to find

$$f(\phi) = \frac{d}{d\phi}\left[F(\phi)\right]$$

where F is the cdf. We can nicely rewrite and rearrange this as

$$\begin{aligned} f_\Phi(\phi) &= \frac{d}{d\phi}\left[F_\Phi(\phi)\right] = \frac{d}{d\phi}\left[P(\Phi < \phi)\right] \\ &= \frac{d}{d\phi}\left[P(h(\theta) < \phi)\right] = \frac{d}{d\phi}\left[P(\theta < h^{-1}(\phi))\right] \\ &= \frac{d}{d\phi}\int_{-\infty}^{h^{-1}(\phi)} f_\Theta(\theta)\, d\theta \\ &= f_\Theta\left(h^{-1}(\phi)\right)\left|\frac{d}{d\phi}\left[h^{-1}(\phi)\right]\right| \end{aligned}$$

And if we use the fact that $\theta = h^{-1}(\phi)$ and drop some of the pedantic subscripting, we can write a more compact version:

$$f(\phi) = f(\theta)\left|\frac{d\theta}{d\phi}\right|$$

□

[15] Note that ϕ must be a function of θ and θ only.