Appendix 7

Introduction to Bayesian Analysis

Uncertainty is the refuge of hope. Henri Frederic Amiel. Version 17 Dec 22

The history of statistical methods in genetics closely parallels advances in computation. Before the widespread use of computers, method-of-moments approaches were common as they are relatively easy to obtain. Here, a summary statistic of the data is computed whose expected value is the parameter of interest (e.g., using the sample mean, $\overline{x}$, as an estimate of the true mean, μ_x, as $E[\overline{x}] = \mu_x$). In the mid-1970s, maximum-likelihood (ML) methods became much more common place, as they offer a very flexible platform for statistical analysis (estimation, determining precision, and hypothesis testing), but at the cost of numerically searching an often highly complex multidimensional likelihood surface (Appendix 4). Both these approaches typically return **point estimators** for the variables of interest, along with some measure of their uncertainty. As opposed to these classical (or **frequentist**) approaches, **Bayesian statistics** (which can be viewed as a natural extension of likelihood methods) is concerned with generating the *full distribution* for the parameters, $\boldsymbol{\Theta}$, given the data, $\mathbf{x}$, namely, obtaining the posterior distribution, $p(\boldsymbol{\Theta} \mid \mathbf{x})$. As such, Bayesian statistics provides a much more complete picture of the uncertainty in the estimation of the unknown parameters, especially after the confounding effects of nuisance parameters are removed.

Our treatment here is intentionally quite brief. A number of texts have presented excellent treatments of the statistical theory (e.g., Lindley 1965; Berger 1985; Carlin and Louis 2000; Lee 2012; Gelman et al. 2013). Blasco (2017) provided a very lucid introduction to applications in quantitative genetics, while Sorensen and Gianola (2002) offered a more comprehensive treatment. While very deep (and very subtle) differences in philosophy separate hard-core Bayesians from hard-core frequentists (Efron 1986; Glymour 1981), our treatment of Bayesian methods is motivated simply by their use as a powerful statistical tool. This appendix focuses on the basic theory, while the computational approaches that make these methods feasible are examined in Appendix 8.

WHY ARE BAYESIAN METHODS BECOMING MORE POPULAR?

In addition to providing a more formal framework for dealing with parameter uncertainty, two specific features have fueled the rapid growth of Bayesian approaches in genetics and genomics. First, under a Bayesian analysis, all parameters are random effects as opposed to fixed effects (Chapter 10). This has profound implications for degrees of freedom. Consider a gene expression study with 30,000 features (genes of interest), whose mRNA levels are contrasted over a set of 100 normal liver cells versus 100 cancerous ones. If we treat the differential expression level of any particular gene as a fixed effect (an unknown constant to be estimated) we will very quickly use all of the degrees of freedom, given the small sample size. Conversely, if these levels are treated as **random effects**, with the expression difference associated with a particular gene being a random variable drawn from some underlying (and unknown) distribution, then the only degrees of freedom lost will be those used to estimate the associated parameters for this underlying distribution (typically, its variance). Further, prediction of the random realization that corresponds to a particular gene borrows information over all the genes. Thus, a Bayesian analysis can handle high-dimensional experiments in which the number of parameters, p, greatly exceeds the number of observations, n, in a framework that fully manages the uncertainty over all these estimates. Second,

Bayesian methods are ***computationally feasible***, as approaches such as MCMC (Appendix 8) allow high-dimensional datasets to be analyzed in a computationally efficient manner. In settings with a large number of nuisance parameters or a high-dimensional dataset, a Bayesian approach not only has considerable appeal, it may be the only approach that is even feasible.

BAYES' THEOREM

The foundation of Bayesian statistics is Bayes' theorem, which was introduced in Chapter 3. From Equation 3.3b,

$$\Pr(\theta \mid x) = \frac{\Pr(x \mid \theta)\Pr(\theta)}{\Pr(x)} \tag{A7.1}$$

In Bayesian statistics, x represents an observable variable (the data), while θ represents a parameter describing the distribution of x. In this setting, $\Pr(\theta)$ is the **prior distribution** of possible parameter values, while $\Pr(\theta \mid x)$ is the subsequent **posterior distribution** of θ given the observed data x and the prior. In classical statistics, the unknown parameters are treated as fixed and the data are considered random, whereas under a Bayesian analysis, the data are considered fixed and the unknown parameters that generated the data are considered random.

Equation A7.1 also holds for continuous random variables, with the probability density function, p, replacing the discrete probability value, $\Pr$. In particular, the continuous multivariate version of Bayes' theorem is

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{\int p(\mathbf{x}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta}} \tag{A7.2}$$

where $\boldsymbol{\Theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ is a vector of n (potentially) continuous variables. As with the univariate case, $p(\boldsymbol{\Theta})$ is the assumed prior distribution of the unknown parameters, while $p(\boldsymbol{\Theta} \mid \mathbf{x})$ is the posterior distribution given the prior, $p(\boldsymbol{\Theta})$, and the data, $\mathbf{x}$.

The origin of Bayes' theorem has a fascinating history (Stigler 1983). It is named after the Rev. Thomas Bayes, a priest who never published a mathematical paper during his lifetime. The paper in which the theorem appears was posthumously read before the Royal Society by his friend Richard Price in 1764. Stigler suggests it was first discovered by Nicholas Saunderson, a blind mathematician and optician who, at age 29, became Lucasian Professor of Mathematics at Cambridge (the position held earlier by Issac Newton). This is an example of **Stigler's Law of Eponymy** (Stigler 1980), wherein no discovery or invention is named after its first discoverer (an **eponym**). As is fitting, Stigler's law is self-consistent, as this phenomenon was previously mentioned by Merton (1965).

Example A7.1. Consider a recessive color locus in cattle in which the genotypes *BB* and *Bb* are black, while *bb* is red. Two black-coated parents are crossed, and produce some red offspring, which implies that both parents must be *Bb*. A black-coated son of theirs is crossed to n red dams (*bb*), and all of his offspring are black. What is the posterior probability that he is *BB*?

To solve this problem using Bayes' theorem, we first define the indicator random variable

$$\theta = \begin{cases} 0 & \text{son is } Bb \\ 1 & \text{son is } BB \end{cases}$$

Given that both parents are *Bb*, the expected priors for their offspring are $1/4$ for *BB* and $1/2$ for *Bb*, resulting in a $3/4$ prior for a black-coated offspring. Further, from conditional probability (Equation 3.3a), the prior that a black offspring is *BB* is

$$\Pr(BB \mid \text{Black}) = \frac{\Pr(BB, \text{Black})}{\Pr(\text{Black})} = \frac{\Pr(BB)}{\Pr(\text{Black})} = \frac{1/4}{3/4} = 1/3$$

where we used the fact that all *BB* are black, so that Pr(*BB*, Black) = Pr(*BB*). Hence, the prior becomes

$$\Pr(\theta) = \begin{cases} 0 & \text{is } 2/3 \\ 1 & \text{is } 1/3 \end{cases}$$

Further

$$\Pr(\text{all } n \text{ offspring are black} \mid \text{sire is } Bb) = (1/2)^n$$

and

$$\Pr(\text{all } n \text{ black}) = \Pr(\text{all black} \mid BB)*\Pr(BB) + \Pr(\text{all black} \mid Bb)*\Pr(Bb)$$

$$= 1 \cdot 1/3 + (1/2)^n \cdot (2/3)$$

If we combine the above values, Bayes' theorem yields

$$\Pr(\theta = 1 \mid n \text{ black offspring}) = \frac{\Pr(n \mid \theta = 1)\Pr(\theta = 1)}{\Pr(n)} = \frac{1 \cdot (1/3)}{1 \cdot (1/3) + (1/2)^n \cdot (2/3)}$$

which returns values of 0.5, 0.67, 0.8, 0.89, 0.94, and 0.998 for $n = 1, 2, 3, 4, 5$, and 10, respectively.

FROM LIKELIHOOD TO BAYESIAN ANALYSIS

The method of maximum likelihood (Appendix 4) and Bayesian analysis are closely related. Suppose $\ell(\boldsymbol{\Theta} \mid \mathbf{x})$ is the assumed likelihood function. Under ML estimation, we would compute the **mode** of the likelihood function (the maximal value of ℓ, as a function of $\boldsymbol{\Theta}$ given the data $\mathbf{x}$), and use the local curvature around the mode to construct confidence intervals. Hypothesis testing follows using likelihood-ratio (LR) statistics. The strengths of ML estimation rely on its *large-sample* properties, namely, that when the sample size is sufficiently large, we can assume both normality of the estimators and that most LR tests follow χ^2 distributions. These features, nice as they are, may not hold for small samples. Conversely, a Bayesian analysis is *exact* for any sample size, given a specified prior.

To transition from a likelihood to a Bayesian analysis, we start with some prior distribution, $p(\boldsymbol{\Theta})$, that captures our initial knowledge (or best guess) about the possible values of the unknown parameters. From Bayes' theorem, the data (likelihood) is combined with the prior to produce a posterior distribution,

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) = \frac{1}{p(\mathbf{x})} \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{A7.3a}$$

$$= (\text{normalizing constant}) \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{A7.3b}$$

$$= \text{constant} \cdot \text{likelihood} \cdot \text{prior} \tag{A7.3c}$$

as $p(\mathbf{x} \mid \boldsymbol{\Theta}) = \ell(\boldsymbol{\Theta} \mid \mathbf{x})$ is simply the likelihood function (Appendix 4) and $1/p(\mathbf{x})$ is a constant (with respect to $\boldsymbol{\Theta}$). Consequently, the posterior distribution is often written as

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) \propto \ell(\boldsymbol{\Theta} \mid \mathbf{x})\ p(\boldsymbol{\Theta}) \tag{A7.3d}$$

where the symbol $\propto$ means "proportional to" (equal up to a constant). Note that the constant $p(\mathbf{x})$ normalizes $p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})$ to one (making the posterior is a formal probability distribution), and hence can be obtained by integration

$$p(\mathbf{x}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})\, d\boldsymbol{\Theta} \tag{A7.4}$$

The dependence of the posterior on the prior (which can easily be assessed by trying different priors) provides an indication of how much information on the unknown parameter values is contained in the data (the curvature of the likelihood surface). If the posterior is highly dependent on the prior, then the data likely has little signal (a **flat likelihood surface**), while if the posterior is largely unaffected by different priors, then the data are likely highly informative (a sharply peaked likelihood surface). To see this, taking logs on Equation A7.3c yields

$$\log(\text{posterior}) = \log(\text{likelihood}) + \log(\text{prior}) + \text{constant} \tag{A7.5}$$

When the likelihood signal is strong, it largely dominates the prior in the resulting posterior, but when a likelihood is weak, the prior can dominate.

Marginal Posterior Distributions

Often only a subset of the unknown parameters is of concern, with the rest being **nuisance parameters** that are of no interest, but still must be fitted in the model. A strong feature of Bayesian analysis is that we can account for all the uncertainty introduced into the parameters of interest by any uncertainty in the values of nuisance parameters. This is accomplished by integrating the nuisance parameters out of the posterior distribution to generate a **marginal posterior distribution** for the parameters of interest. For example, suppose the mean and variance of data coming from a normal distribution are unknown, but our real interest is only in the variance. Estimating the mean introduces additional uncertainty into our variance estimate, which is not fully captured by standard classical approaches. Under a Bayesian analysis, the marginal posterior distribution for σ^2 is simply

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

The resulting marginal posterior for σ^2 captures all of the uncertainty in the estimation of μ that influences the uncertainty in σ^2. This is an especially nice feature when a large number of nuisance parameters must be estimated.

The marginal posterior may involve several parameters (generating **joint marginal posteriors**). Suppose we write the vector of unknown parameters as $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_{nu})$, where $\boldsymbol{\Theta}_{nu}$ is the vector of nuisance parameters. Integrating over $\boldsymbol{\Theta}_{nu}$ yields the desired marginal for the vector $\boldsymbol{\Theta}_1$ of parameters of interest as

$$p(\boldsymbol{\Theta}_1 \,|\, \mathbf{y}) = \int_{\boldsymbol{\Theta}_{nu}} p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_{nu} \,|\, \mathbf{y})\, d\boldsymbol{\Theta}_{nu} \tag{A7.6}$$

While these complex integrals appear quite daunting (and indeed almost always are from an analytic standpoint), generating draws from the marginal distribution is usually very straightforward using MCMC methods (which are examined in Appendix 8).

SUMMARIZING THE POSTERIOR DISTRIBUTION

How do we extract a Bayesian estimator for some unknown parameter, θ? If our mindset is to use some sort of point estimator (as is usually done in classical statistics), then there are a number of candidates. We could follow maximum likelihood and use the **mode of the posterior distribution** (its maximal value)

$$\widehat{\theta} = \max_{\theta} [\, p(\,\theta \,|\, \mathbf{x}\,)] \tag{A7.7a}$$

We could take the **expected value of** θ (its mean) given the posterior

$$\widehat{\theta} = E[\,\theta \,|\, \mathbf{x}\,] = \int \theta\, p(\,\theta \,|\, \mathbf{x}\,) d\theta \tag{A7.7b}$$

Another candidate is the **median of the posterior**, which is more robust than the mean to outliers. Here the estimator satisfies $\Pr(\theta > \widehat{\theta} \,|\, \mathbf{x}) = \Pr(\theta < \widehat{\theta} \,|\, \mathbf{x}) = 0.5$, hence

$$\int_{\widehat{\theta}}^{+\infty} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \int_{-\infty}^{\widehat{\theta}} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \frac{1}{2} \tag{A7.7c}$$

However, using any of the above estimators, or even all three simultaneously, loses the full power of a Bayesian analysis, as ***the full estimator is the entire posterior density itself***. If we cannot obtain the full form of the posterior distribution, then these estimates of general features of the distribution can be presented. However, as we will see in Appendix 8, we can generally obtain the full posterior by simulation using MCMC sampling, and hence the Bayesian estimate of a parameter is often presented as a frequency histogram (potentially smoothed) of the MCMC-generated samples from the posterior distribution (an **empirical posterior**). Typically, when such a histogram is displayed, it is usually accompanied by one or more of the summary statistics given by Equation A7.7a—A7.7c (as well as other metrics, such as the variance and skewness).

Highest Density Regions (HDRs)

Given the posterior distribution, the construction of uncertainty intervals is straightforward. For example, a $100(1-\alpha)\%$ "confidence interval" is given by any $(L_{\alpha/2}, H_{\alpha/2})$ satisfying

$$\int_{L_{\alpha/2}}^{H_{\alpha/2}} p(\theta \,|\, \mathbf{x})\, d\theta = 1 - \alpha$$

To reduce the set of possible candidate intervals, one typically uses **highest density regions**, or **HDRs**, where, for a single parameter, the HDR $100(1-\alpha)$ region(s) are the shortest intervals giving an area of $(1-\alpha)$. More generally, if multiple parameters are being estimated, the HDR region(s) are those with the smallest *volume* in the parameter space. HDRs are also referred to as **Bayesian confidence intervals** or (better yet) **credible intervals**.

It is critical to note that there is a *profound difference* between a confidence interval (CI) from classical (frequentist) statistics and a Bayesian analysis. The interpretation of a classical confidence interval is that if we were to repeat the experiment a sufficiently large number of times, and construct CIs in the same fashion, the fraction of the resulting collection of CIs that enclose the unknown parameter approaches $(1-\alpha)$. Thus, the frequentist CI is a measure of the *frequency* of occurrences in independent experiments in which the CI encloses the true value (and hence the term frequentist for this type of statistics). In contrast, with a Bayesian HDR, there is a probability of $(1-\alpha)$ that the interval contains the true value of the unknown parameter. While at first blush these two interpretations of CIs appear to be essentially identical, they are not, and indeed they are fundamentally (but subtly) different. Often the CI and Bayesian intervals span essentially the same values, but again the interpretational difference remains. The key point is that the Bayesian prior allows us to make ***direct probability statements*** about θ, while under classical statistics we can only make statements about the behavior of the statistic if we consider ***repeating an experiment a large number of times***. Given the important conceptual difference between classical and Bayesian intervals, Bayesians typically avoid using the term *confidence interval*, using the term *credible interval* instead.

Bayes Factors and Hypothesis Testing

In the classical hypothesis-testing framework, we have two alternatives. The null hypothesis, H_0, that the unknown parameter, θ, belongs to some set or interval, Θ_0 ($\theta \in \Theta_0$), versus the alternative hypothesis, H_1, that θ belongs to the alternative set, Θ_1 ($\theta \in \Theta_1$). Θ_0 and Θ_1 contain no common elements ($\Theta_0 \cap \Theta_1 = \varnothing$) and the union of Θ_0 and Θ_1 contains the entire space of values for θ (i.e., $\Theta_0 \cup \Theta_1 = \Theta$).

In the classical frequentist framework, one uses the observed data to test the significance of a particular hypothesis, and (if possible) compute a p value (the probability, p, of observing

a value equal to, or more extreme than, that of the test statistic if the null hypothesis is indeed correct). Initially, one would think that the idea of a hypothesis test is trivial in a Bayesian framework, as using the posterior distribution provides the expected p values directly. For example,

$$\Pr(\theta > \theta_0) = \int_{\theta_0}^{\infty} p(\,\theta \,|\, \mathbf{x})\, d\theta \qquad \text{and} \qquad \Pr(\theta_0 < \theta < \theta_1) = \int_{\theta_0}^{\theta_1} p(\,\theta \,|\, \mathbf{x})\, d\theta$$

The fault in this logic under a Bayesian framework is that we also have *prior information* and Bayesian hypothesis testing addresses whether, *given the data*, we are more or less inclined to believe the hypothesis than was suggested from the prior. Hence, the ***prior probabilities influence hypothesis testing***. To formalize this idea, let

$$p_0 = \Pr(\theta \in \Theta_0 \,|\, \mathbf{x}) \qquad \text{and} \qquad p_1 = \Pr(\theta \in \Theta_1 \,|\, \mathbf{x}) \tag{A7.8a}$$

denote the probabilities, given the observed data, $\mathbf{x}$, that θ is in the null (p_0) and alternative (p_1) hypothesis sets. Note that these are *posterior* probabilities. Because $\Theta_0 \cap \Theta_1 = \oslash$ and $\Theta_0 \cup \Theta_1 = \Theta$, it follows that $p_0 + p_1 = 1$. Likewise, for the *prior* probabilities we have

$$\pi_0 = \Pr(\theta \in \Theta_0) \qquad \text{and} \qquad \pi_1 = \Pr(\theta \in \Theta_1) \tag{A7.8b}$$

Thus the **prior odds** of H_0 versus H_1 are π_0/π_1, while the **posterior odds** are p_0/p_1.

The **Bayes factor**, $\mathbf{B}_0$, in favor of H_0 versus H_1 is calculated by the ratio of the posterior odds divided by the prior odds,

$$B_0 = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0\pi_1}{p_1\pi_0} \tag{A7.9a}$$

The Bayes factor is loosely interpreted as the odds in favor of H_0 over H_1 as given by the data and our prior opinion. Because $\pi_1 = 1 - \pi_0$ and $p_1 = 1 - p_0$, we can also express this as

$$B_0 = \frac{p_0(1-\pi_0)}{\pi_0(1-p_0)} \tag{A7.9b}$$

By symmetry, note that the Bayes factor, B_1, in favor of H_1 versus H_0 is simply $B_1 = 1/B_0$.

Example A7.2. Suppose that the prior distribution of θ is such that $\Pr(\theta > \theta_0) = 0.10$, while for the posterior distribution $\Pr(\theta > \theta_0 \,|\, \mathbf{x}) = 0.05$. The latter is significant at the 5% level in a classical hypothesis-testing framework, but the data only doubles our confidence in the alternative hypothesis relative to our belief based on prior information. If $\Pr(\theta > \theta_0) = 0.50$ for the prior, then a 5% posterior probability would greatly increase our confidence in the alternative hypothesis. Consider the first case in this example, where the prior and posterior probabilities for the null were $\pi_0 = 0.1$ and $p_0 = 0.05$, respectively. The Bayes factor in favor of H_1 versus H_0 is

$$B_1 = \frac{\pi_0(1-p_0)}{p_0(1-\pi_0)} = \frac{0.1 \cdot 0.95}{0.05 \cdot 0.9} = 2.11$$

Similarly, for the second example, where the prior for the null was $\pi_0 = 0.5$,

$$B_1 = \frac{0.5 \cdot 0.95}{0.05 \cdot 0.5} = 19$$

Here, the data showed close to a 20-fold improvement (relative to the prior) in support of H_1. Bayes factors and p values represent fundamentally different approaches to an analysis and are not formally comparable. However, a *loose* interpretation is that a factor of 20 is akin to the level of support of a $p = 0.05$, and a factor of 100 to $p = 0.01$.

When the hypotheses are simple (i.e., single values), say $\Theta_0 = \theta_0$ vs. $\Theta_1 = \theta_1$, then

$$p_i \propto p(\theta_i)\, p(\mathbf{x} \,|\, \theta_i) = \pi_i\, p(\mathbf{x} \,|\, \theta_i) \quad \text{for} \quad i = 0, 1$$

Thus

$$\frac{p_0}{p_1} = \frac{\pi_0\, p(\mathbf{x} \,|\, \theta_0)}{\pi_1\, p(\mathbf{x} \,|\, \theta_1)} \tag{A7.10a}$$

and from Equation A7.9a, the Bayes factor (in favor of the null) reduces to

$$B_0 = \frac{p(\mathbf{x} \,|\, \theta_0)}{p(\mathbf{x} \,|\, \theta_1)} \tag{A7.10b}$$

which is simply a *likelihood ratio* (Appendix 4).

When hypotheses are **composite** (containing multiple elements), the situation is slightly more complicated. First, note that the prior distribution of θ conditioned on H_0 or H_1 is

$$p_i(\theta) = p(\theta)/\pi_i \qquad \text{for} \quad i = 0, 1 \tag{A7.11}$$

as the total probability $\theta \in \Theta_i = \pi_i$, so dividing by π_i normalizes the distribution to integrate to one. Thus,

$$\begin{aligned} p_i = \Pr(\theta \in \Theta_i \,|\, \mathbf{x}) &= \int_{\theta \in \Theta_i} p(\theta \,|\, \mathbf{x}) d\theta \\ &= \frac{1}{p(\mathbf{x})} \int_{\theta \in \Theta_i} p(\theta) p(\mathbf{x} \,|\, \theta) d\theta \\ &= \pi_i \int_{\theta \in \Theta_i} p(\mathbf{x} \,|\, \theta) p_i(\theta) d\theta \end{aligned} \tag{A7.12}$$

where the second step follows from Bayes' theorem, while the final step follows from Equation A7.11. The Bayes factor in favor of the null hypothesis becomes

$$B_0 = \left(\frac{p_0}{\pi_0}\right)\left(\frac{\pi_1}{p_1}\right) = \frac{\int_{\theta \in \Theta_0} p(\mathbf{x} \,|\, \theta) p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(\mathbf{x} \,|\, \theta) p_1(\theta) d\theta} \tag{A7.13}$$

which is a ratio of the weighted likelihoods of Θ_0 and Θ_1.

THE CHOICE OF A PRIOR

Obviously, a critical feature of any Bayesian analysis is the choice of a prior. The key is that when the data have a sufficiently strong signal, even a poor choice of a prior will still not greatly influence the posterior. In a sense, it is an asymptotic (large-sample) property of Bayesian analysis in that all but pathological priors (those with zero probability where the true value lies) can be overcome by sufficient amounts of data. As mentioned above, one can check the impact of the prior by assessing the stability of posterior over a collection of diverse priors. The **location** of a parameter (mean or mode) and its **precision** (the reciprocal of the variance) of the prior is usually more critical than its actual shape in terms of conveying prior information. The shape (family) of the prior distribution is often chosen to facilitate calculation of the posterior, especially through the use of **conjugate priors** that, for a given likelihood function, return a posterior in the same distribution family as the prior (e.g., a gamma prior returns a gamma posterior when the likelihood is Poisson). We will return to conjugate priors, but first we will discuss other approaches for construction of priors.

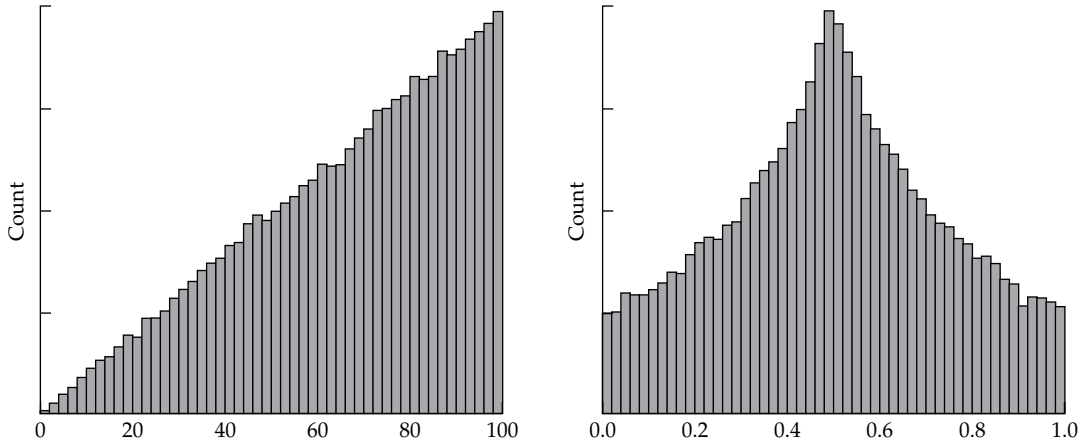

Figure A7.1 A uniform prior on one scale does not result in a flat prior on a transformed scale. Suppose a flat prior on (0,10000) is assumed for both the additive and residual variances. To mimic what happens under MCMC, we display these priors by using the resulting histograms generated from a large number of random draws, with a uniform expected to return a flat histogram. **Left:** The resulting prior for the standard deviation of either variance (the square root of a random draw). **Right:** The resulting prior for h^2, the ratio of a random draw for the additive variance divided by this value plus a random draw for the residual variance. Neither of these priors result in a uniform prior (namely, a flat histogram) on the transformed scale.

Diffuse Priors

One of the most commonly used priors is the **flat** or **diffuse** (also called **uninformative** or **naive**) prior, which is simply a constant

$$p(\theta) = \frac{1}{b-a} \qquad \text{for} \qquad a \leq \theta \leq b \tag{A7.14a}$$

This conveys that we have no a priori reason to favor any particular parameter value over another. With a flat prior, the posterior is just a constant C times the likelihood

$$p(\theta \,|\, \mathbf{x}) = C\, \ell(\theta \,|\, \mathbf{x}) \tag{A7.14b}$$

and we typically write that $p(\theta \,|\, \mathbf{x}) \propto \ell(\theta \,|\, \mathbf{x})$. In many cases, classical expressions from frequentist statistics are obtained by Bayesian analysis through assuming a flat prior (e.g., the posterior mode is the MLE).

If the variable (i.e., parameter) of interest ranges over $(0, \infty)$ or $(-\infty, +\infty)$, then, strictly speaking, a flat prior does not exist as, if the constant takes on any nonzero value, the integral does not exist. In such cases a flat prior (i.e., assuming $p[\theta \,|\, \mathbf{x}] \propto \ell[\theta \,|\, \mathbf{x}]$) is referred to as an **improper prior**, and care must be taken to ensure that the product of the prior and the likelihood results in a proper posterior (i.e., $\ell[\theta \,|\, \mathbf{x}]$ has a finite integral over the parameter range). This is by no means certain.

Another complication involved in using a uniform prior arises when the question of interest resides on a *different scale* than that used for the prior. A variable uniform on one scale may be far from uniform on a transformed scale. Figure A7.1 shows two examples based on the assumption that there was a flat prior on the variance. A uniform prior on the variance does *not* result in a uniform prior on the standard deviation (e.g., Van Dongen 2006). Likewise, if one assumes that the additive and residual variances have flat priors, this does not imply a flat prior for h^2, but rather a prior that is sharply peaked at $1/2$. When assuming a flat prior, care must be taken that it is truly uninformative on the appropriate scale of biological interest. Otherwise, the choice of what superficially appears as an unbiased prior may instead create a bias that the signal in the data must overcome.

The Jeffreys Prior

Jeffreys (1961) proposed a general prior based on the Fisher information information, F, of the likelihood. Recall (Equation A4.7b) that

$$F(\theta \,|\, \mathbf{x}) = -E\left[\frac{\partial^2 \ln \ell(\theta \,|\, \mathbf{x})}{\partial\, \theta^2}\right]$$

The **Jeffreys prior** is as follows:

$$p(\theta) \propto \sqrt{F(\theta \,|\, \mathbf{x})} \tag{A7.15}$$

A full discussion, with derivation, can be found in Lee (2012).

When there are k parameters, **F** is the $k \times k$ Fisher information matrix of the expected second partials, where the elements of **F** are calculated by

$$\mathbf{F}(\,\boldsymbol{\Theta} \,|\, \mathbf{x})_{ij} = -E_x\left[\frac{\partial^2 \ln \ell(\boldsymbol{\Theta} \,|\, \mathbf{x})}{\partial\theta_i\ \partial\theta_j}\right]$$

In this case, the Jeffreys prior becomes

$$p(\boldsymbol{\Theta}) \propto \sqrt{\det[\mathbf{F}(\theta \,|\, \mathbf{x})\,]} \tag{A7.16}$$

Example A7.3. Consider the likelihood of x successes in n independent draws from a binomial with a success parameter of θ,

$$\ell(\theta \,|\, \mathbf{x}) = C\theta^x(1-\theta)^{n-x}$$

where the constant C does not involve θ. Taking logs gives

$$L(\theta \,|\, \mathbf{x}) = \ln[\,\ell(\theta \,|\, \mathbf{x})\,] = \ln C + x\ln\theta + (n-x)\ln(1-\theta)$$

Thus

$$\frac{\partial L(\theta \,|\, \mathbf{x})}{\partial\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and likewise

$$\frac{\partial^2 L(\theta \,|\, \mathbf{x})}{\partial\theta^2} = -\frac{x}{\theta^2} - (-1)\cdot(-1)\frac{n-x}{(1-\theta)^2} = -\left(\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right)$$

Because $E[\,x\,] = n\theta$, then

$$-E\left[\frac{\partial^2 \ln \ell(\theta \,|\, \mathbf{x})}{\partial\, \theta^2}\right] = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = n\,\theta^{-1}(1-\theta)^{-1}$$

The resulting Jeffreys prior for this likelihood becomes

$$p(\theta) \propto \sqrt{\theta^{-1}(1-\theta)^{-1}} \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which is a U-shaped beta distribution with parameters $\alpha = \beta = 1/2$ (Equation A7.37a). This prior puts more weight on extreme values relative to assuming a uniform over (0,1), see Figure A7.3.

Example A7.4. Suppose our data consists of n independent draws from a normal distribution with an unknown mean and variance, μ and σ^2. In Example A4.3, we showed that the information matrix in this case is

$$\mathbf{F} = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Because the determinant of a diagonal matrix is the product of the diagonal elements, $\det(\mathbf{F}) \propto \sigma^{-6}$, giving the Jeffreys prior for μ and σ^2 as

$$p(\boldsymbol{\Theta}) \propto \sqrt{\sigma^{-6}} = \sigma^{-3}$$

Because the joint prior does not involve μ, this implies a flat prior for μ (i.e., $p[\mu] = c$). Note here that the prior distributions of μ and σ^2 are independent, as

$$p(\mu, \theta) = c \cdot \sigma^{-3} = p(\mu) \cdot p(\sigma^2)$$

POSTERIOR DISTRIBUTIONS UNDER NORMALITY ASSUMPTIONS

To introduce the basic ideas of Bayesian analysis, as well as treating a common assumption in quantitative genetics, consider the case where data are drawn from a normal (Gaussian) distribution, giving the likelihood function for the ith observation, x_i, as

$$\ell(\mu, \sigma^2 \mid x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{A7.17a}$$

If we assume independence, the resulting full likelihood for all n data points (with a sample mean of $\overline{x}$) is

$$\ell(\mu \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{A7.17b}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2\right)\right] \tag{A7.17c}$$

The form of the posteriors given these normal likelihoods is a function of the assumed priors. By using the appropriate conjugate priors, these posteriors follow fairly standard distributions, and hence are easier to work with, as we now demonstrate.

Gaussian Likelihood With Known Variance and Unknown Mean

As a starting point, assume that the variance, σ^2, is known, while the mean, μ, is unknown. For a Bayesian analysis, it remains to specify the prior for μ, $p(\mu)$. Suppose we assume a Gaussian prior, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, with

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \tag{A7.18}$$

The mean and variance of the prior, μ_0 and σ_0^2, are referred to as **hyperparameters**. Here, μ_0 specifies a prior location for the parameter (the unknown mean, μ), while σ_0^2 specifies our uncertainty in this prior location—the larger σ_0^2, the greater is our uncertainty. In the limit as $\sigma_0^2 \to \infty$, $p(\mu)$ approaches a flat (and in this case, improper) prior.

A useful device when calculating the posterior distribution is to ignore terms that are constants with respect to the unknown parameters. Suppose $\mathbf{x}$ denotes the data and $\boldsymbol{\Theta}_1$ is a vector of *known* model parameters, while $\boldsymbol{\Theta}_2$ is a vector of unknown parameters. If we can write the posterior as

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) = f(\mathbf{x}, \boldsymbol{\Theta}_1) \cdot g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{A7.19a}$$

then

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) \propto g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{A7.19b}$$

which follows because $f(\mathbf{x}, \boldsymbol{\Theta}_1)$ is constant with respect to $\boldsymbol{\Theta}_2$.

With the prior given by Equation A7.18, we can express the resulting posterior distribution as

$$\begin{aligned} p(\mu \,|\, \mathbf{x}) &\propto \ell(\mu \,|\, \mathbf{x}) \cdot p(\mu) \\ &\propto \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2\right)\right] \end{aligned} \tag{A7.20a}$$

We can factor out additional terms not involving μ to obtain

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{\mu^2}{2\sigma_0^2} + \frac{\mu\,\mu_0}{\sigma_0^2} + \frac{\mu n\overline{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right) \tag{A7.20b}$$

Factoring in terms of μ, the term in the exponential becomes

$$-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right) = -\frac{\mu^2}{\sigma_*^2} + \frac{2\mu\mu_*}{2\sigma_*^2} \tag{A7.21a}$$

where

$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_* = \sigma_*^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right) \tag{A7.21b}$$

Finally, by completing the square, we have

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left[-\frac{(\mu-\mu_*)^2}{2\sigma_*^2} + f(\mathbf{x}, \mu_0, \sigma^2, \sigma_0^2)\right] \tag{A7.21c}$$

Recalling Equation A7.19b, we can ignore the second term in the exponential (as it does not involve μ), and the resulting posterior for μ (given the observed data $\mathbf{x}$) becomes

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left[-\frac{(\mu-\mu_*)^2}{2\sigma_*^2}\right] \tag{A7.22a}$$

demonstrating that the posterior density function for μ is a normal with a mean of μ_* and a variance of σ_*^2, namely,

$$\mu \,|\, (\mathbf{x}, \sigma^2) \sim \mathrm{N}\left(\mu_*, \sigma_*^2\right) \tag{A7.22b}$$

Notice that the posterior density is in the same form as the prior. This occurred because the prior **conjugated** with the likelihood function—the product of the prior and likelihood returned a distribution in the same family as the prior (but with different distribution parameters). The use of such **conjugate priors** associated with a given family of likelihood functions is a key concept in Bayesian analysis, and we will explore it more fully below.

We are now in a position to inquire about the relative importance of the prior versus the data. Under the assumed prior, the mean (and in this case, the mode as well) of the posterior distribution is

$$\mu_* = \mu_0\,\frac{\sigma_*^2}{\sigma_0^2} + \overline{x}\,\frac{\sigma_*^2}{\sigma^2/n} \tag{A7.23}$$

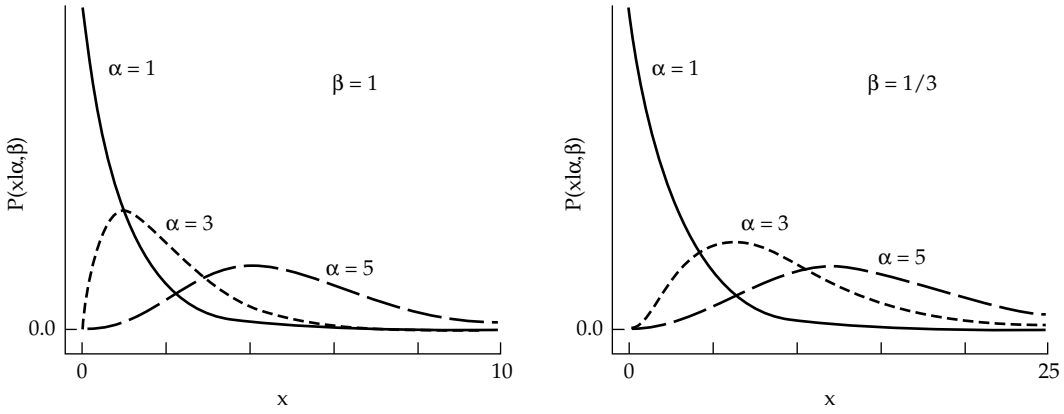

Figure A7.2 The effect of the **shape** (α) and **rate** ($\beta = 1/\lambda$, the inverse of the **scale**) **parameters** on the gamma distribution function. For $\alpha = 1$, the resulting distribution is the simple monotonically decreasing exponential, while for $\alpha > 1$, the distribution is unimodal. The effect of a change in the rate or scale is to keep the general shape but change the scaling with respect to x.

With a very diffuse prior on μ (i.e., $\sigma_0^2 \gg \sigma^2$), $\sigma_*^2 \rightarrow \sigma^2/n$ and $\mu_* \rightarrow \overline{x}$. Also note from Equation A7.21b that as we collect enough data (i.e., achieve a sufficiently large value of n), $\sigma_*^2 \rightarrow \sigma^2/n$ and again $\mu_* \rightarrow \overline{x}$, implying that the data, rather than the prior, will be the primarily influence on posterior when the value of n is sufficiently large.

Gamma, χ^2, Inverse-gamma, and χ^{-2} Distributions

Before examining the Gaussian likelihood with unknown variance, a brief aside is needed to develop the **inverse chi-square distribution**, denoted by χ^{-2}. We do this via the gamma and inverse-gamma distributions, as both χ^2 and χ^{-2} are special cases of these distributions.

To motivate the **gamma distribution**, first consider the simple exponential waiting-time distribution, where β is the **rate parameter** (the probability of a success in some small time unit, δ_t, is given by $\beta\,\delta_t$), then the probability density function (pdf) for the exponential is

$$p(x \mid \beta) = \beta e^{-\beta x} \quad \text{for} \quad 0 \leq x < \infty, \quad \beta > 0$$

Because the expected waiting time until a success is $\lambda = 1/\beta$, this can be reparameterized in terms of the **scale parameter** (waiting time) as

$$p(x \mid \beta) = \lambda^{-1} e^{-x/\lambda}$$

The sum of k exponentials with the same rate (or scale) parameter is called an **Erlang distribution**, and it was initially developed for certain problems in telephone queuing theory. Expressed in terms of the rate parameter, the resulting pdf becomes

$$p(x \mid k, \beta) = \frac{\beta^k}{(k-1)!}\, x^{k-1}\, e^{-\beta x} \quad \text{for} \quad 0 \leq x < \infty$$

where the integer k is called the **shape parameter**, with $k = 1$ recovering the exponential.

The gamma distribution follows by allowing the shape parameter to be any positive number, α, with $x \sim \text{Gamma}(\alpha, \beta)$ having its pdf defined by its shape (α) and rate (β) values,

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x} \quad \text{for } \alpha,\ \beta,\ x > 0 \tag{A7.24a}$$

Note that the factorial in the Erlang is replaced by the gamma function, $\Gamma(x)$, which is defined below (Equation A7.25a). Figure A7.2 shows how changes in these two parameters

Table A7.1 Summary of the functional forms (in terms of x) of various gamma-related distributions. See the text for further details.

Distribution	α	β	$p(x)$/constant
Gamma (α, β)			$x^{\alpha-1}e^{-\beta x}$
Chi-square, χ^2_n	$n/2$	1/2	$x^{n/2-1}e^{-x/2}$
Inverse-gamma (α, β)			$x^{-(\alpha+1)}e^{-\beta/x}$
Inverse chi-square, χ^{-2}_n	$n/2$	1/2	$x^{-(n/2+1)}e^{-1/(2x)}$
Scaled inverse chi-square, $\chi^{-2}_{(n,\sigma_0^2)}$	$n/2$	$\sigma_0^2/2$	$x^{-(n/2+1)}e^{-\sigma_0^2/(2x)}$

influence the shape of the distribution. Note that, as a function of x,

$$p(x \mid \alpha, \beta) \propto x^{\alpha-1}e^{-\beta x} \tag{A7.24b}$$

When expressed in terms of the scale ($\lambda = 1/\beta$) parameter, the pdf becomes

$$p(x \mid \alpha, \lambda) = \frac{\lambda^{-\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1}e^{-x/\lambda}$$

which yields

$$p(x \mid \alpha, \lambda) \propto x^{\alpha-1}e^{-x/\lambda} \tag{A7.24c}$$

Because both the rate and scale versions of the gamma distribution are widely used, take care to know which version your software package is using (for example, the default in R uses the scale parameter version). We can parameterize a gamma in terms of its mean and variance by noting that

$$\mu_x = \frac{\alpha}{\beta} = \alpha\ \lambda \qquad \text{and} \qquad \sigma_x^2 = \frac{\alpha}{\beta^2} = \alpha\,\lambda^2 \tag{A7.24d}$$

so that

$$\alpha = \frac{\mu_x^2}{\sigma_x^2} \qquad \text{and} \qquad \beta = \frac{\mu_x}{\sigma_x^2} \tag{A7.24e}$$

$\Gamma(\alpha)$, the **gamma function** evaluated at α (which normalizes the gamma distribution), is defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy \qquad \text{for } \alpha > 0 \tag{A7.25a}$$

This is the generalization of the factorial function from the integers to any positive number. If n is an integer, then $\Gamma(n) = (n-1)!$ Using integration by parts, one can show that Γ satisfies the following identities

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \Gamma(1) = 1, \quad \text{and} \quad \Gamma(1/2) = \sqrt{\pi} \tag{A7.25b}$$

The chi-square (χ^2) distribution is a special case of the gamma, as a χ^2 random variable with n degrees of freedom follows a gamma distribution with parameters $\alpha = n/2$ and $\beta = 1/2$ ($\lambda = 2$), namely, $\chi^2_n \sim \text{Gamma}(n/2, 1/2)$, giving the density function as

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)}\, x^{n/2-1}e^{-x/2} \tag{A7.26a}$$

Hence for $x \sim \chi^2_n$,

$$p(x) \propto x^{n/2-1}e^{-x/2} \tag{A7.26b}$$

The **inverse-gamma** distribution will prove useful as a conjugate prior for Gaussian likelihoods with unknown variance. It is defined by the distribution of the random variable $y = x^{-1}$, where $x \sim \text{Gamma}(\alpha, \beta)$. The resulting density function is

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x} \quad \text{for } \alpha,\ \beta,\ x > 0 \tag{A7.27a}$$

The mean and variance for this distribution are only defined (i.e., finite) if α is sufficiently large, with

$$\mu_x = \frac{\beta}{\alpha - 1} \quad \text{for } \alpha > 1 \qquad \text{and} \qquad \sigma_x^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2 \tag{A7.27b}$$

Note for the inverse gamma that

$$p(x \mid \alpha, \beta) \propto x^{-(\alpha+1)} e^{-\beta/x} \tag{A7.27c}$$

If $y \sim \chi_n^2$, then $x = 1/y$ follows an **inverse chi-square distribution**, which is denoted by $x \sim \chi_n^{-2}$. This is a special case of the inverse gamma, with (as for a normal χ^2) $\alpha = n/2$, $\beta = 1/2$. For $n > 4$ (i.e., $\alpha > 2$), the resulting density function is

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{-(n/2+1)} e^{-1/(2x)} \tag{A7.28a}$$

with a mean and variance of

$$\mu_x = \frac{1}{n-2} \qquad \text{and} \qquad \sigma_x^2 = \frac{2}{(n-2)^2(n-4)} \tag{A7.28b}$$

The **scaled inverse chi-square distribution** is more typically used in a Bayesian analysis, where the rate parameter, β (which equals 1/2 under a chi-square), is replaced by $\beta = \sigma_0^2/2$, making the resulting pdf

$$p(x \mid n) \propto x^{-(n/2+1)} e^{-\sigma_0^2/(2x)} \tag{A7.29a}$$

where the $1/(2x)$ term in the exponential is replaced by a $\sigma_0^2/(2x)$ term. The scaled inverse chi-square distribution thus involves two parameters (σ_0^2 and n), and is denoted by $\chi^{-2}_{(n,\sigma_0^2)}$ or SI$-\chi^2(n, \sigma_0^2)$. Note that if

$$x \sim \chi^{-2}_{(n,\sigma_0^2)}, \qquad \text{then} \qquad \sigma_0^2 x \sim \chi_n^{-2} \tag{A7.29b}$$

which shows that σ_0^2 is a scaling factor on a standard ($\beta = 1/2$) inverse chi-square.

Gaussian Likelihood With Unknown Variance: Scaled Inverse-χ^2 Priors

Suppose data are drawn from a normal distribution with a known mean, μ, but unknown variance, σ^2. The resulting likelihood function can be expressed as

$$\ell(\sigma^2 \mid \mathbf{x}, \mu) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \tag{A7.30a}$$

where

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{A7.30b}$$

Notice that because we condition on $\mathbf{x}$ and μ (i.e., their values are known), S^2 is a constant. Further observe that, as a function of the unknown variance, σ^2, the likelihood is

proportional to a scaled inverse χ^2 distribution (Equation A7.29a). If we take the prior for the unknown variance also as a scaled inverse χ^2 with hyperparameters ν_0 and σ_0^2, the posterior becomes

$$\begin{aligned} p(\sigma^2 \,|\, \mathbf{x}, \mu) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{nS^2}{2\sigma^2}\right)(\sigma^2)^{-\nu_0/2-1} \cdot \exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-(n+\nu_0)/2-1} \exp\left(-\frac{nS^2+\sigma_0^2}{2\sigma^2}\right) \end{aligned} \tag{A7.31a}$$

Equation A7.29a shows the resulting posterior is also a scaled inverse χ^2 distribution with parameters $\nu_n = (n + \nu_0)$ and $\sigma_n^2 = (nS^2 + \sigma_0^2)$. Hence,

$$\text{the prior } \sigma^2 \sim \chi^{-2}_{\nu_0,\sigma_0^2} \quad \text{yields the posterior} \quad \sigma^2 \,|\, (\mathbf{x}, \mu) \sim \chi^{-2}_{\nu_n,\sigma_n^2} \tag{A7.31b}$$

Student's *t* Distribution

The final distribution needed for a Bayesian analysis of a Gaussian likelihood is the ***t*** (or **Student's *t***) distribution. Suppose that $x_i \sim N(\mu, \sigma^2)$, so for n independent draws, $\overline{x} \sim N(\mu, \sigma^2/n)$. This implies that $(\overline{x} - \mu)/\sqrt{\sigma^2/n} \sim U$, where $U \sim N(0, 1)$ denotes a unit normal. Likewise, the sample variance, $\text{Var}(x)$, follows a scaled chi-square distribution, with $\text{Var}(x) \sim (n-1)\sigma^2\chi^2_{n-1}$ (Equation A5.15a). When the estimated variance, $\text{Var}(x)$, is used in place of the true variance, σ^2, the quantity $(\overline{x}-\mu)/\sqrt{\text{Var}(x)/n}$ follows a t distribution with $n-1$ degrees of freedom, giving rise to the very familiar ***t*-test**. Notice that

$$t_{n-1} = \left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}\right)\left(\frac{1}{\sqrt{\text{Var(x)}/\sigma^2}}\right) = \frac{U}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

Thus, a t_ν random variable follows the distribution of a unit normal divided by the square root of a chi-square with ν degrees of freedom,

$$t_\nu = \frac{U}{\sqrt{\chi^2_\nu/\nu}} \tag{A7.32a}$$

Note that $E(\chi^2_\nu) = \nu$, so $E(\chi^2_\nu/\nu) = 1$. Relative to a normal, a t distribution is more peaked and has heavier tails, and this kurtosis becomes more pronounced as ν decreases. Indeed, the tails fall off sufficiently slowly that a t random variable with two degrees of freedom has an infinite variance, while a t with four (or fewer) degrees of freedom has an infinite fourth moment. The coefficient of kurtosis (Equation 2.12a) for a t with $\nu > 4$ degrees of freedom is $k_4 = 6/(\nu-4)$, which approaches the value (zero) for a normal random variable for large values of ν. For $\nu > 30$, the t essentially becomes a unit normal distribution.

As with a unit normal, one can also add scale and location to a standard t_ν-distributed random variable, thus generating a three-parameter family of distributions,

$$t_\nu(\mu, \sigma) = \mu + \sigma \cdot t_\nu \tag{A7.32b}$$

The resulting mean and variance this distribution are

$$E[t_\nu(\mu,\sigma)] = \mu \quad \text{and} \quad \sigma^2[t_\nu(\mu,\sigma)] = \sigma^2\,\frac{\nu}{\nu-2} \quad \text{for } \nu > 2 \tag{A7.32c}$$

Hence, the choice of μ and σ control, respectively, the location and scale (uncertainty about the location), while ν controls the kurtosis, with heavy tails for values of ν that are small and little kurtosis for $\nu > 20$. The resulting probability density function thus becomes

$$p(x \,|\, \nu, \mu, \sigma) = \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)\sigma\sqrt{\pi\,\nu}}\left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2} \tag{A7.32d}$$

The role of the t distribution in Bayesian statistics is twofold. First, it is often used as a ***more robust prior***, as its heavier tails may better account for outliers. Using a t distribution with low degrees of freedom (often $\nu = 5$) offers a prior that is similar to a normal but allows for more frequent extreme values. The second scenario is that the marginal posterior for μ of a Gaussian likelihood with a normal prior on the mean and an inverse chi-square prior on the variance is a t distribution. This arises after the joint posterior is integrated over all possible σ^2 values (i.e., over an inverse chi-square).

General Gaussian Likelihood: Unknown Mean and Variance

If we put all these pieces together, the posterior density for draws from a normal with the mean and variance both unknown is obtained as follows. First, we write the joint prior by conditioning on the variance,

$$p(\mu, \sigma^2) = p(\mu \,|\, \sigma^2) \cdot p(\sigma^2) \tag{A7.33a}$$

As above, we assume a scaled inverse chi-square distribution for the variance and, conditioned on the variance, a Gaussian prior for the mean with hyperparameters of μ_0 and σ^2/κ_0, namely,

$$\sigma^2 \sim \chi^{-2}_{\nu_0, \sigma_0^2} \qquad \text{and} \qquad \mu \,|\, \sigma^2 \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \tag{A7.33b}$$

We write the variance for the conditional mean prior in this way because σ^2 is known (as we condition on it) and we scale σ^2 by the hyperparameter, κ_0. The resulting marginal posterior becomes

$$\sigma^2 \,|\, \mathbf{x} \sim \chi^{-2}_{\nu_n, \sigma_n^2} \quad \text{and} \quad \mu \,|\, \mathbf{x} \sim \mathrm{t}_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right) \tag{A7.34}$$

where $\mathrm{t}_n(\mu, \sigma^2)$ denotes a t distribution with n degrees of freedom, mean μ, and scale parameter σ^2, and where

$$\nu_n = \nu_0 + n, \qquad \kappa_n = \kappa_0 + n \tag{A7.35a}$$

$$\mu_n = \mu_0 \frac{\kappa_0}{\kappa_n} + \overline{x} \frac{n}{\kappa_n} = \mu_0 \frac{\kappa_0}{\kappa_0 + n} + \overline{x} \frac{n}{\kappa_0 + n} \tag{A7.35b}$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0 \sigma_0^2 + \sum_{i=1}^{n} (\, x_i - \overline{x}\,)^2 + \frac{\kappa_0 n}{\kappa_n} (\, \overline{x} - \mu_0)^2\right) \tag{A7.35c}$$

CONJUGATE PRIORS

The above use of prior densities that conjugate the likelihood allowed us to develop analytic expressions of the posterior density. As we will see in Appendix 8, this is critical in developing Gibbs samplers for problems of interest. Table A7.2 summarizes the conjugate priors for several common likelihood functions, with the various families of distributions discussed below.

The Beta and Dirichlet Distributions

With a binomial, each trial (observation) has two possible outcomes and the likelihood is a function of the sample size (number of trials), n, and a single success probability, p (as the two outcomes on any given trial have probabilities of p and $1 - p$). The generalization of this model is the multinomial distribution (Equation 2.20a), where now each trial has k possible outcomes, which requires $k - 1$ success probabilities to describe the likelihood. In particular, for a total of n observations, the probability that n_1 are in category 1, n_2 in category 2, $\cdots$, and n_k in category k is

$$p(n_1, \cdots n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!}\, p_1^{n_1} \cdots p_k^{n_k} \quad \text{where} \quad \sum_i n_i = n \quad \text{and} \quad \sum_i p_i = 1 \tag{A7.36a}$$

Table A7.2 Conjugate priors for common likelihood functions. If one uses the distribution family of the conjugate prior with its paired likelihood function, then the resulting posterior is in the same distribution family as the prior (albeit, of course, with different parameters).

Likelihood	Conjugate prior	Equation
Binomial	Beta	A7.37a
Multinomial	Dirichlet	A7.36b
Poisson	Gamma	A7.26a
Normal		
μ unknown, σ^2 known	Normal	A7.17a
μ known, σ^2 unknown	Inverse chi-square	A7.29a
Multivariate normal		
$\boldsymbol{\mu}$ unknown, $\mathbf{V}$ known	Multivariate normal	9.24
$\boldsymbol{\mu}$ known, $\mathbf{V}$ unknown	Inverse-Wishart	A7.40

The conjugate prior for the multinomial likelihood is the **Dirichlet distribution**. If we let $\mathbf{x} = (x_1, x_2, \cdots, x_k)$ denote the k success probabilities, when pdf for $\mathbf{x} \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_k)$ is

$$p(x_1, \cdots x_k \,|\, \alpha_1, \cdots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\, x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \tag{A7.36b}$$

where

$$\alpha_0 = \sum_{i=1}^{k} \alpha_i \quad \text{with} \quad \alpha_i > 0, \quad \text{and} \quad \sum_{i=1}^{k} x_i = 1 \quad \text{with} \quad 0 \le x_i \le 1 \tag{A7.36c}$$

At first glance, this looks like the multinomial density function (with $\alpha_i - 1 = n_i$). The difference is that the multinomial is calculated over a set of discrete random variables (n_i), thus returning the expected probabilities for any vector of discrete numbers of counts (successes) in each category. Conversely, the Dirichlet treats an equivalent of the vector of outcomes (generalized to non-integers) as fixed and returns the continuous distribution for all possible configurations of the *success parameters* given this data, which means that the data (α_i) is fixed, and the success parameters (x_i) are random. A few key moments of this distribution are

$$\mu_{x_i} = \frac{\alpha_i}{\alpha_0}, \quad \sigma^2(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{and} \quad \sigma(x_i, x_j) = -\frac{\alpha_i\, \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{A7.36d}$$

An important special case of the Dirichlet (for $k = 2$ classes) is the **beta distribution**, whose pdf is given by

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha - 1}(1 - x)^{\beta - 1} \quad \text{for} \quad 0 \le x \le 1, \quad \alpha,\, \beta > 0 \tag{A7.37a}$$

which has a mean and a variance of

$$\mu = \frac{\alpha}{\alpha + \beta} \qquad \text{and} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta - 1)} \tag{A7.37b}$$

As Figure A7.3 illustrates, the beta distribution is *extremely flexible*, and can be flat, unimodal, U- , or L-shaped, depending on the choice of α and β.

Wishart and Inverse-Wishart Distributions

The **Wishart distribution** can be thought of as the multivariate extension of the χ^2 distribution. Suppose $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are independent and identically distributed vectors, with

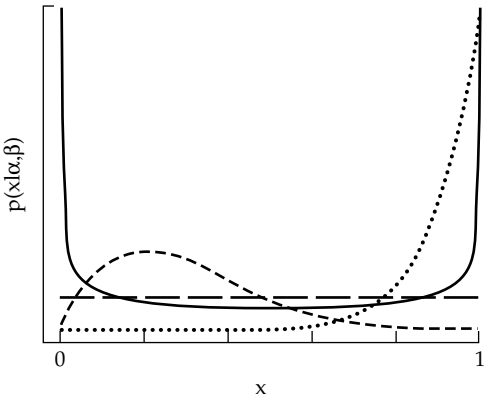

Figure A7.3 For $\alpha = \beta = 1$ (long-dashed curve), the beta distribution is simply the uniform distribution over $(0, 1)$. The pdf for the beta distribution can also be U-shaped ($\alpha = \beta = 0.5$; solid curve), unimodal ($\alpha = 2, \beta = 5$; short-dashed curve), or L-shaped ($\alpha = 10, \beta = 1$; dotted curve). Because the beta distribution is symmetric in α and β, switching their parameter values generates a distribution of the same shape translated about 0.5.

$\mathbf{x}_i \sim \text{MVN}_k(\mathbf{0}, \mathbf{V})$. Using these n draws, and assuming that the mean is known to be zero, the resulting random ($k \times k$ symmetric, positive definite) sample covariance matrix, **W**, is given by

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i \, \mathbf{x}_i^T \sim \text{W}_n(\mathbf{V}) \tag{A7.38}$$

This sum is defined as a Wishart distribution with n degrees of freedom and a (matrix) parameter **V**. Recalling that the sum of n squared unit normals follows a χ^2_n distribution, the Wishart is the extension to the multivariate normal. Indeed, for $k = 1$ with $\mathbf{V} = (1)$, the Wishart is simply a χ^2_n distribution, as $\sum x_i^2 \sim \chi^2_n$, because $x_i \sim N(0, 1)$.

The Wishart is the sampling distribution for covariance matrices (just like the χ^2 is associated with the distribution of a sample variance for data drawn from a normal; Equation A5.13c). The pdf of the Wishart distribution is

$$p(\mathbf{W} \mid \mathbf{V}) = 2^{-nk/2} \pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{-n/2} \mid \mathbf{W} \mid^{(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{V}^{-1}\mathbf{W}\right]\right)}{\prod_{i=1}^{k} \Gamma\left(\frac{n+1-i}{2}\right)} \tag{A7.39}$$

Recall that the trace (tr) of a matrix is just the sum of its diagonal elements, $\text{tr}(\mathbf{A}) = \sum A_{ii}$ (Chapter 9). Odell and Feiveson (1966) presenteds an algorithm for generating random draws from the Wishart.

If $\mathbf{Z} \sim \mathbf{W}_n(\mathbf{V})$, then $\mathbf{Z}^{-1} \sim \mathbf{W}_n^{-1}\left(\mathbf{V}^{-1}\right)$, where $\mathbf{W}^{-1}$ denotes the **inverse-Wishart distribution**. The density function for an inverse-Wishart distributed random matrix, **W**, is

$$p(\mathbf{W} \mid \mathbf{V}) = 2^{-nk/2} \pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{n/2} \mid \mathbf{W} \mid^{-(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{V}\mathbf{W}^{-1}\right]\right)}{\prod_{i=1}^{k} \Gamma\left(\frac{n+1-i}{2}\right)} \tag{A7.40}$$

which is the distribution of the inverse of the sample covariance matrix.