Stat 5102 Lecture Slides
Deck 4

Charles J. Geyer

School of Statistics

University of Minnesota

Bayesian Inference

Now for something completely different.

Everything we have done up to now is frequentist statistics. Bayesian statistics is very different.

Bayesians don't do confidence intervals and hypothesis tests. Bayesians don't use sampling distributions of estimators. Modern Bayesians aren't even interested in point estimators.

So what do they do? Bayesians treat parameters as random variables.

To a Bayesian probability is only way to describe uncertainty. Things not known for certain — like values of parameters — must be described by a probability distribution.

Bayesian Inference (cont.)

Suppose you are uncertain about something. Then your uncertainty is described by a probability distribution called your *prior distribution*.

Suppose you obtain some data relevant to that thing. The data changes your uncertainty, which is then described by a new probability distribution called your *posterior distribution*.

The posterior distribution reflects the information both in the prior distribution and the data.

Most of Bayesian inference is about how to go from prior to posterior.

Bayesian Inference (cont.)

The way Bayesians go from prior to posterior is to use the laws of conditional probability, sometimes called in this context *Bayes rule* or *Bayes theorem*.

Suppose we have a PDF g for the prior distribution of the parameter θ, and suppose we obtain data x whose conditional PDF given θ is f. Then the joint distribution of data and parameters is conditional times marginal

$$f(x \mid \theta) g(\theta)$$

This may look strange because, up to this point in the course, you have been brainwashed in the frequentist paradigm. Here both x and θ are random variables.

Bayesian Inference (cont.)

The correct posterior distribution, according to the Bayesian paradigm, is the conditional distribution of θ given x, which is joint divided by marginal

$$h(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int f(x \mid \theta) g(\theta) \, d\theta}$$

Often we do not need to do the integral. If we recognize that

$$\theta \mapsto f(x \mid \theta) g(\theta)$$

is, except for constants, the PDF of a brand name distribution, then that distribution must be the posterior.

Binomial Data, Beta Prior

Suppose the prior distribution for p is Beta(α_1, α_2) and the conditional distribution of x given p is Bin(n, p). Then

$$\begin{aligned} f(x \mid p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ g(p) &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1 - 1} (1-p)^{\alpha_2 - 1} \end{aligned}$$

and

$$f(x \mid p) g(p) = \binom{n}{x} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot p^{x + \alpha_1 - 1} (1-p)^{n - x + \alpha_2 - 1}$$

and this, considered as a function of p for fixed x is, except for constants, the PDF of a Beta$(x + \alpha_1, n - x + \alpha_2)$ distribution. So that is the posterior.

Binomial Data, Beta Prior (cont.)

A bit slower, for those for whom that was too fast. If we look up the Beta(α_1, α_2) distribution in the brand name distributions handout, we see the PDF

$$f(x) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1} \qquad 0 < x < 1$$

We want $g(p)$. To get that we must change f to g, which is trivial, and x to p, which requires some care. That is how we got

$$g(p) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p^{\alpha_1 - 1}(1 - p)^{\alpha_2 - 1}$$

on the preceding slide.

Binomial Data, Beta Prior (cont.)

Note that we don't change x to p in $f(x \mid p)$. There x is the variable and p is the constant, because x is in front of the bar and p is behind the bar. And that is just what we see for the formula for the PDF of the Bin(n, p) distribution in the brand-name distributions handout. There is nothing that needs to be changed.

Binomial Data, Beta Prior (cont.)

So if we get to

$$f(x \mid p)g(p) = \binom{n}{x} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1}$$

how do we recognize this as the PDF of a brand-name distribution, except for constants?

First, remember that p is the variable and x is fixed. In Bayesian inference the parameter is always the variable for the prior or posterior and the data are always fixed. The posterior is the conditional distribution of the parameter or parameters given the data.

Binomial Data, Beta Prior (cont.)

Second, remember that p is a continuous variable, so we are looking for the distribution of a continuous random variable!

Binomial Data, Beta Prior (cont.)

Third, when we try to find something that looks like

$$p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1} \qquad (*)$$

in the brand-name distributions handout, we find only the binomial PMF and the beta PDF. From the second point we don't want a PMF. Hence the posterior, if brand-name, must be beta. To find which beta, we need to match up

$$x^{\alpha_1-1}(1-x)^{\alpha_2-1}$$

(the non-constant part of the beta PDF in the brand-name distributions handout) with $(*)$. To do that we need to change x to p, which is right because in the posterior p is the variable. And we need to change the first parameter to $x+\alpha_1$ and the second parameter to $n-x+\alpha_2$. That's how we get Beta$(x+\alpha_1, n-x+\alpha_2)$ for the posterior.

Bayesian Inference (cont.)

Simple. It is just a “find the other conditional” problem, which we did when studying conditional probability first semester. But there are a lot of ways to get confused and go wrong if you are not careful.

Bayesian Inference (cont.)

In Bayes rule, “constants” meaning anything that doesn’t depend on the parameter are irrelevant. We can drop multiplicative constants that do not depend on the parameter from $f(x \mid \theta)$ obtaining the likelihood $L(\theta)$. We can also drop multiplicative constants that do not depend on the parameter from $g(\theta)$ obtaining the unnormalized prior. Multiplying them together gives the unnormalized posterior

$$\text{likelihood} \times \text{unnormalized prior} = \text{unnormalized posterior}$$

Bayesian Inference (cont.)

In our example we could have multiplied likelihood

$$p^x (1 - p)^{n-x}$$

times unnormalized prior

$$p^{\alpha_1 - 1} (1 - p)^{\alpha_2 - 1}$$

to get unnormalized posterior

$$p^x (1 - p)^{n-x} p^{\alpha_1 - 1} (1 - p)^{\alpha_2 - 1} = p^{x + \alpha_1 - 1} (1 - p)^{n - x + \alpha_2 - 1}$$

which, as before, can be recognized as an unnormalized beta PDF.

Bayesian Inference (cont.)

It is convenient to have a name for the parameters of the prior and posterior. If we call them parameters, then we get confused because they play a different role from the parameters of the distribution of the data.

The parameters of the distribution of the data, p in our example, the Bayesian treats as random variables. They are the random variables whose distributions are the prior and posterior.

The parameters of the prior, α_1 and α_2 in our example, the Bayesian treats as known constants. They determine the particular prior distribution used for a particular problem. To avoid confusion we call them *hyperparameters*.

Bayesian Inference (cont.)

Parameters, meaning the parameters of the distribution of the data and the variables of the prior and posterior are unknown constants. The Bayesian treats them as random variables because probability theory is the correct description of uncertainty.

Hyperparameters, meaning the parameters of the prior and posterior are known constants. The Bayesian treats them as non-random variables because there is no uncertainty about their values.

In our example, the hyperparameters of the prior are α_1 and α_2, and the hyperparameters of the posterior are $x + \alpha_1$ and $n - x + \alpha_2$.

Normal Data, Normal Prior

Suppose X_1, ..., X_n are $\mathcal{N}(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. Suppose we are being Bayesian and our prior distribution for μ is also normal, say $\mathcal{N}(\mu_0, \sigma_0^2)$. We cannot denote the hyperparameters μ_0 and σ_0^2 as μ and σ, because we would then get them confused with the parameters μ and σ of the distribution of the data.

The likelihood is

$$L_n(\mu) = \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}\right)$$

(Deck 3, Slide 11). The PDF of the unnormalized prior is

$$g(\mu) = \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

Normal Data, Normal Prior (cont.)

It simplifies the math if we introduce

$$\lambda = \frac{1}{\sigma^2} \quad \text{and} \quad \lambda_0 = \frac{1}{\sigma_0^2}$$

(reciprocal variance is called *precision* so λ and λ_0 are precision parameters) so

$$\begin{aligned} L_n(\mu) &= \exp\left(-\tfrac{n}{2}\lambda(\bar{x}_n - \mu)^2\right) \\ g(\mu) &= \exp\left(-\tfrac{1}{2}\lambda_0(\mu - \mu_0)^2\right) \\ L_n(\mu)g(\mu) &= \exp\left(-\tfrac{n}{2}\lambda(\bar{x}_n - \mu)^2 - \tfrac{1}{2}\lambda_0(\mu - \mu_0)^2\right) \end{aligned}$$

A Lemma

The function

$$x \mapsto \exp\left(ax^2 + bx + c\right) \tag{$**$}$$

having domain the whole real line is an unnormalized PDF if and only if $a < 0$, in which case is the unnormalized PDF of the normal distribution with mean, variance, and precision

$$\begin{aligned} \mu &= -\frac{b}{2a} \\ \sigma^2 &= -\frac{1}{2a} \\ \lambda &= -2a \end{aligned}$$

A Lemma (cont.)

If $a \geq 0$, then $(**)$ does not go to zero as x goes to plus or minus infinity, hence the integral of $(**)$ does not exist and it cannot be normalized to make a PDF (5101, Deck 4, Slides 8 ff.)

A Lemma (cont.)

If $a < 0$, then we can equate $(**)$ to the unnormalized PDF of a normal distribution

$$\exp\left(ax^2 + bx + c\right) = d \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $d > 0$ is arbitrary. Taking logs we get

$$\begin{aligned} ax^2 + bx + c &= -\frac{(x-\mu)^2}{2\sigma^2} + \log(d) \\ &= -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} + \log(d) \end{aligned} \qquad (***)$$

A Lemma (cont.)

The only way $(***)$ can hold for all real x is if the coefficients of x^2 and x match on both sides

$$
\begin{aligned}
a &= -\frac{1}{2\sigma^2} \\
b &= \frac{\mu}{\sigma^2}
\end{aligned}
$$

and solving for μ and σ^2 gives the assertion of the lemma.

Normal Data, Normal Prior (cont.)

Returning to the problem where we had unnormalized posterior

$$\begin{aligned} &\exp\left(-\tfrac{n}{2}\lambda(\bar{x}_n - \mu)^2 - \tfrac{1}{2}\lambda_0(\mu - \mu_0)^2\right) \\ &\quad = \exp\left(-\tfrac{1}{2}n\lambda\mu^2 + n\lambda\bar{x}_n\mu - \tfrac{1}{2}n\lambda\bar{x}_n^2 - \tfrac{1}{2}\lambda_0\mu^2\lambda_0\mu_0\mu - \tfrac{1}{2}\lambda_0\mu_0^2\right) \\ &\qquad \exp\left(\left[-\tfrac{1}{2}n\lambda - \tfrac{1}{2}\lambda_0\right]\mu^2 + \left[n\lambda\bar{x}_n + \lambda_0\mu_0\right]\mu - \tfrac{1}{2}n\lambda\bar{x}_n^2 - \tfrac{1}{2}\lambda_0\mu_0^2\right) \end{aligned}$$

and we see we have the situation described by the lemma with

$$\begin{aligned} a &= -\frac{n\lambda + \lambda_0}{2} \\ b &= n\lambda\bar{x}_n + \lambda_0\mu_0 \end{aligned}$$

Normal Data, Normal Prior (cont.)

Hence by the lemma, the posterior is normal with hyperparameters

$$\begin{aligned} \mu_1 &= -\frac{b}{2a} \\ &= \frac{n \lambda \bar{x}_n + \lambda_0 \mu_0}{n \lambda + \lambda_0} \\ \lambda_1 &= -2a \\ &= n \lambda + \lambda_0 \end{aligned}$$

We give the hyperparameters of the posterior subscripts 1 to distinguish then from hyperparameters of the prior (subscripts 0) and parameters of the data distribution (no subscripts).

Bayesian Inference (cont.)

Unlike most of the notational conventions we use, this one (about subscripts of parameters and hyperparameters) is not widely used, but there is no widely used convention.

Toy Example

Suppose x has the distribution with PDF

$$f(x \mid \theta) = \frac{1 + 2\theta x}{1 + \theta}, \qquad 0 < x < 1$$

where $\theta > -1/2$ is the parameter. Suppose our prior distribution for θ is uniform on the interval (0, 2). Then the posterior is also concentrated on the interval (0, 2) and the unnormalized posterior is

$$\frac{1 + 2\theta x}{1 + \theta}$$

thought of a function of θ for fixed x.

Toy Example (cont.)

According to Mathematica, the normalized posterior PDF is

$$h(\theta \mid x) = \frac{1 + 2\theta x}{(1 + \theta)(2x(2 - \log(3)) + \log(3))}$$

Conjugate Priors

Given a data distribution $f(x \mid \theta)$, a family of distributions is said to be *conjugate* to the given distribution if whenever the prior is in the conjugate family, so is the posterior, regardless of the observed value of the data.

Trivially, the family of *all* distributions is always conjugate.

Our first example showed that, if the data distribution is binomial, then the conjugate family of distributions is beta.

Our second example showed that, if the data distribution is normal with known variance, then the conjugate family of distributions is normal.

Conjugate Priors (cont.)

How could we *discover* that binomial and beta are conjugate?

Consider the likelihood for arbitrary data and sample size

$$L_n(p) = p^x (1 - p)^{n-x}$$

If multiplied by another likelihood of the same family but different data and sample size, do we get the same form back? Yes!

$$p^{x_1}(1 - p)^{n_1 - x_1} p^{x_2}(1 - p)^{n_2 - x_2} = p^{x_3}(1 - p)^{n_3 - x_3}$$

where

$$\begin{aligned} x_3 &= x_1 + x_2 \\ n_3 &= n_1 + n_2 \end{aligned}$$

Conjugate Priors (cont.)

Hence, if the prior looks like the likelihood, then the posterior will too.

Thus we have discovered the conjugate family of priors. We only have to recognize

$$p \mapsto p^x (1 - p)^{n-x}$$

as an unnormalized beta distribution.

Conjugate Priors (cont.)

Note that beta with integer-valued parameters is also a conjugate family of priors in this case. But usually we are uninterested in having the smallest conjugate family. When we discover that a brand-name family is conjugate, we are happy to have the full family available for prior distributions.

Conjugate Priors (cont.)

Suppose the data are IID Exp(λ). What is the brand-name conjugate prior distribution? The likelihood is

$$L_n(\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) = \lambda^n \exp(-\lambda \bar{x}_n)$$

If we combine two likelihoods for two independent samples we get

$$\lambda^n \exp(-\lambda n \bar{x}_n) \times \lambda^n \exp(-\lambda n \bar{y}_n) = \lambda^n \exp\bigl(-\lambda n(\bar{x}_n + \bar{y}_n)\bigr)$$

where $\bar{x}_n$ and $\bar{y}_n$ are the means for the two samples. This has the same form as the log likelihood for one sample.

Conjugate Priors (cont.)

Hence, if the prior looks like the likelihood, then the posterior will too. We only have to recognize

$$g(\lambda) = \lambda^n \exp(-\lambda \bar{x}_n)$$

as the form

$$x^{\text{something}} \exp(-\text{something else} \cdot x)$$

of the gamma distribution. Thus $\Gamma(\alpha, \lambda)$ is the conjugate family.

Improper Priors

A *subjective Bayesian* is a person who really buys the Bayesian philosophy. Probability is the only correct measure of uncertainty, and this means that people have probability distributions in their heads that describe any quantities they are uncertain about. In any situation one must make one's best effort to get the correct prior distribution out of the head of the relevant user and into Bayes rule.

Many people, however, are happy to use the Bayesian paradigm while being much less fussy about priors. As we shall see, when the sample size is large, the likelihood outweighs the prior in determining the posterior. So, when the sample size is large, the prior is not crucial.

Improper Priors (cont.)

Such people are willing to use priors chosen for mathematical convenience rather than their accurate representation of uncertainty.

They often use priors that are very spread out to represent extreme uncertainty. Such priors are called “vague” or “diffuse” even though these terms have no precise mathematical definition.

In the limit as the priors are spread out more and more one gets so-called improper priors.

Improper Priors (cont.)

Consider our $\mathcal{N}(\mu_0, 1/\lambda_0)$ priors we used for μ when the data are normal with unknown mean μ and known variance.

What happens if we let λ_0 decrease to zero so the prior variance goes to infinity? The limit is clearly not a probability distribution. But let us take limits on unnormalized PDF

$$\lim_{\lambda_0 \downarrow 0} \exp\left(-\tfrac{1}{2}\lambda_0(\mu - \mu_0)^2\right) = 1$$

The limiting unnormalized prior something-or-other (we can't call it a probability distribution) is constant

$$g(\mu) = 1, \qquad -\infty < \mu < +\infty$$

Improper Priors (cont.)

What happens if we try to use this improper g in Bayes rule? It works!

Likelihood times unnormalized improper prior is just the likelihood, because the improper prior is equal to one, so we have

$$L_n(\mu) g(\mu) = \exp\left(-\tfrac{n}{2} \lambda (\bar{x}_n - \mu)^2\right)$$

and this thought of as a function of μ for fixed data is proportional to a $\mathcal{N}\bigl(\bar{x}_n, 1/(n\lambda)\bigr)$ distribution.

Or, bringing back $\sigma^2 = 1/\lambda$ as the known variance

$$\mu \sim \mathcal{N}\left(\bar{x}_n, \frac{\sigma^2}{n}\right)$$

Improper Priors (cont.)

Interestingly, the Bayesian with this improper prior agrees with the frequentist.

The MLE is $\widehat{\mu}_n = \bar{x}_n$ and we know the exact sampling distribution of the MLE is

$$\widehat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where μ is the true unknown parameter value (recall σ^2 is known). To make a confidence interval, the frequentist would use the pivotal quantity

$$\widehat{\mu}_n - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

Improper Priors (cont.)

So the Bayesian and the frequentist, who are in violent disagreement about which of μ and $\widehat{\mu}_n$ is random — the Bayesian says $\widehat{\mu}_n$ is just a number, hence non-random, after the data have been seen and μ, being unknown, is random, since probability is the proper description of uncertainty, whereas the frequentist treats $\widehat{\mu}_n$ as random and μ as constant — agree about the distribution of $\widehat{\mu}_n - \mu$.

Improper Priors (cont.)

Also interesting is that although the limiting process we used to derive this improper prior makes no sense, the same limiting process applied to posteriors does make sense

$$\mathcal{N}\left(\frac{n\lambda\bar{x}_n + \lambda_0\mu_0}{n\lambda + \lambda_0}, \frac{1}{n\lambda + \lambda_0}\right) \to \mathcal{N}\left(\bar{x}_n, \frac{1}{n\lambda}\right), \qquad \text{as } \lambda_0 \downarrow 0$$

Improper Priors (cont.)

So how do we make a general methodology out of this?

We started with a limiting argument that makes no sense and arrived at posterior distributions that do make sense.

Let us call an *improper prior* any nonnegative function on the parameter space whose integral does not exist. We run it though Bayes rule just like a proper prior.

However, we are not really using the laws of conditional probability because an improper prior is not a PDF (because it doesn't integrate). We are using the form but not the content. Some people say we are using the *formal Bayes rule*.

Improper Priors (cont.)

There is no guarantee that

$$\text{likelihood} \times \text{improper prior} = \text{unnormalized posterior}$$

results in anything that can be normalized. If the right-hand side integrates, then we get a proper posterior after normalization. If the right-hand does not integrate, then we get complete nonsense.

You have to be careful when using improper priors that the answer makes sense. Probability theory doesn't guarantee that, because improper priors are not probability distributions.

Improper Priors (cont.)

Improper priors are very questionable.

- Subjective Bayesians think they are nonsense. They do not correctly describe the uncertainty of anyone.

- Everyone has to be careful using them, because they don't always yield proper posteriors. Everyone agrees improper posteriors are nonsense.

- Because the joint distribution of data and parameters is also improper, paradoxes arise. These can be puzzling.

However they are widely used and need to be understood.

Improper Priors (cont.)

For binomial data we know that beta is the conjugate family and likelihood times unnormalized prior is

$$p^{x+\alpha_1-1}(1-p)^{n-x+\alpha_2-1}$$

which is an unnormalized Beta$(x+\alpha_1, n-x+\alpha_2)$ PDF (slide 14).

The posterior makes sense whenever

$$\begin{aligned} x + \alpha_1 &> 0 \\ n - x + \alpha_2 &> 0 \end{aligned}$$

hence for some negative values of α_1 and α_2.

But the prior is only proper for $\alpha_1 > 0$ and $\alpha_2 > 0$.

Improper Priors (cont.)

Our inference looks the same either way

Data Distribution	Bin(n, p)
Prior Distribution	Beta(α_1, α_2)
Posterior Distribution	Beta$(x + \alpha_1, n - x + \alpha_2)$

but when either α_1 or α_2 is nonpositive, we say we are using an improper prior.

Objective Bayesian Inference

The subjective, personalistic aspect of Bayesian inference bothers many people. Hence many attempts have been made to formulate "objective" priors, which are supposed to be priors that many people can agree on, at least in certain situations.

Objective Bayesian inference doesn't really exist, because no proposed "objective" priors achieve wide agreement.

Flat Priors

One obvious “default” prior is flat (constant), which seems to give no preference to any parameter value over any other. If the parameter space is unbounded, then the flat prior is improper.

One problem with flat priors is that they are only flat for one parameterization.

Change of Parameter

Recall the change-of-variable formulas. Univariate: if $x = h(y)$, then

$$f_Y(y) = f_X[h(y)] \cdot |h'(y)|$$

Multivariate: if $\mathbf{x} = h(\mathbf{y})$, then

$$f_Y(\mathbf{y}) = f_X[h(\mathbf{y})] \cdot |\text{det}(\nabla h(\mathbf{y}))|$$

(5101, Deck 3, Slides 124–125). When you do a change-of-variable you pick up a “Jacobian” term.

This holds for change-of-parameter for Bayesians, because parameters are random variables.

Change of Parameter

If we use a flat prior for θ, then the prior for $\psi = \theta^2$ uses the transformation $\theta = h(\psi)$ with

$$h(x) = x^{1/2}$$
$$h'(x) = \tfrac{1}{2} x^{-1/2}$$

so

$$g_\Psi(\psi) = g_\Theta[h(\psi)] \cdot \tfrac{1}{2} \psi^{-1/2} = 1 \cdot \frac{1}{2\sqrt{\psi}}$$

And similarly for any other transformation.

You can be flat on one parameter, but not on any other. On which parameter should you be flat?

Jeffreys Priors

If flat priors are not "objective," what could be?

Jeffreys introduced the following idea. If $I(\theta)$ is Fisher information for a data model with one parameter, then the prior with PDF

$$g(\theta) \propto \sqrt{I(\theta)}$$

(where $\propto$ means proportional to) is objective in the sense that any change-of-parameter yields the Jeffreys prior for that parameter.

If the parameter space is unbounded, then the Jeffreys prior is usually improper.

Jeffreys Priors (cont.)

If we use the Jeffreys prior for θ, then the corresponding prior for a new parameter ψ related to θ by $\theta = h(\psi)$ is by the change-of-variable theorem

$$\begin{aligned} g_\Psi(\psi) &= g_\Theta[h(\psi)] \cdot |h'(\psi)| \\ &\propto \sqrt{I_\Theta[h(\psi)]} \cdot |h'(\psi)| \end{aligned}$$

The relationship for Fisher information is

$$I_\Psi(\psi) = I_\Theta[h(\psi)] \cdot h'(\psi)^2$$

(Deck 3, Slide 101). Hence the change-of-variable theorem gives

$$g_\Psi(\psi) \propto \sqrt{I_\Psi(\psi)}$$

which is the Jeffreys prior for ψ.

Jeffreys Priors (cont.)

The Jeffreys prior for a model with parameter vector $\boldsymbol{\theta}$ and Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ is

$$g(\boldsymbol{\theta}) \propto \sqrt{\det\Big(\mathbf{I}(\boldsymbol{\theta})\Big)}$$

and this has the same property as in the one-parameter case: each change-of-parameter yields the Jeffreys prior for that parameter.

Jeffreys Priors (cont.)

Suppose the data X are Bin(n, p). The Fisher information is

$$I_n(p) = \frac{n}{p(1-p)}$$

(Deck 3, Slide 51) so the Jeffreys prior is

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

which is a proper prior.

Jeffreys Priors (cont.)

Suppose the data X_1, ..., X_n are IID Exp(λ). The Fisher information is

$$I_n(\lambda) = \frac{n}{\lambda^2}$$

(Deck 3, Slide 51) so the Jeffreys prior is

$$g(p) \propto \frac{1}{\lambda}$$

which is an improper prior.

Jeffreys Priors (cont.)

Suppose the data X_1, ..., X_n are IID $\mathcal{N}(\mu, \nu)$. The Fisher information matrix is

$$I_n(\mu, \nu) = \begin{pmatrix} n/\nu & 0 \\ 0 & n/(2\nu^2) \end{pmatrix}$$

(Deck 3, Slide 89) so the Jeffreys prior is

$$g(\mu, \nu) \propto \nu^{-3/2}$$

which is an improper prior.

Two-Parameter Normal

The likelihood for the two-parameter normal data distribution with parameters the mean μ and precision $\lambda = 1/\sigma^2$ is

$$L_n(\mu, \lambda) = \lambda^{n/2} \exp\left(-\frac{n v_n \lambda}{2} - \frac{n\lambda(\bar{x}_n - \mu)^2}{2} \right)$$

(Deck 3, Slides 10 and 80).

We seek a brand name conjugate prior family. This is no brand name bivariate distribution, so we seek a factorization

$$\text{joint} = \text{conditional} \times \text{marginal}$$

in which the marginal and conditional are brand name.

Two-Parameter Normal (cont.)

Finding the conjugate prior is equivalent to finding the posterior for a flat prior.

For fixed ν, we note that the likelihood is "e to a quadratic" in μ hence the posterior conditional for μ given ν that is normal. The normalizing constant is

$$1/\sqrt{2\pi\sigma^2} \propto \lambda^{1/2}$$

Hence we can factor the likelihood = unnormalized posterior into unnormalized marginal × unnormalized conditional as

$$\lambda^{(n-1)/2} \exp\left(-\frac{n v_n \lambda}{2}\right) \times \lambda^{1/2} \exp\left(-\frac{n\lambda(\bar{x}_n - \mu)^2}{2}\right)$$

and we recognize the marginal for λ as gamma.

Two-Parameter Normal (cont.)

We generalize this allowing arbitrary hyperparameters for the conjugate prior

$$\lambda \sim \text{Gam}(\alpha_0, \beta_0)$$
$$\mu \mid \lambda \sim \mathcal{N}(\gamma_0, \delta_0^{-1} \lambda^{-1})$$

the first is the marginal for λ and the second is the conditional for μ given λ. Note that μ and λ are dependent because the conditional depends on λ.

There are four hyperparameters: α_0, β_0, γ_0, and δ_0.

This is called the normal-gamma family of (bivariate) distributions.

Two-Parameter Normal (cont.)

Check how this conjugate family works.

Suppose X_1, ..., X_n are IID $\mathcal{N}(\mu, \lambda^{-1})$ and we use a normal-gamma prior. The likelihood is

$$\lambda^{n/2} \exp\left(-\tfrac{1}{2} n v_n \lambda - \tfrac{1}{2} n \lambda (\bar{x}_n - \mu)^2\right)$$

and the unnormalized prior is

$$\lambda^{\alpha_0 - 1} \exp\bigl(-\beta_0 \lambda\bigr) \cdot \lambda^{1/2} \exp\left(-\tfrac{1}{2} \delta_0 \lambda (\mu - \gamma_0)^2\right)$$

Hence the unnormalized posterior is

$$\lambda^{\alpha_0 + n/2 - 1/2} \exp\left(-\beta_0 \lambda - \tfrac{1}{2} \delta_0 \lambda (\mu - \gamma_0)^2 - \tfrac{1}{2} n v_n \lambda - \tfrac{1}{2} n \lambda (\bar{x}_n - \mu)^2\right)$$

Two-Parameter Normal (cont.)

We claim this is an unnormalized normal-gamma PDF with hyperparameters α_1, β_1, γ_1 and δ_1. It is obvious that the "λ to a power" part matches up with

$$\alpha_1 = \alpha_0 + n/2$$

Two-Parameter Normal (cont.)

It remains to match up the coefficients of λ, $\lambda\mu$, and $\lambda\mu^2$ in the exponent to determine the other three hyperparameters.

$$-\beta_1\lambda-\tfrac{1}{2}\delta_1\lambda(\mu-\gamma_1)^2 = -\beta_0\lambda-\tfrac{1}{2}\delta_0\lambda(\mu-\gamma_0)^2-\tfrac{1}{2}nv_n\lambda-\tfrac{1}{2}n\lambda(\bar{x}_n-\mu)^2$$

So

$$\begin{aligned} -\beta_1 - \tfrac{1}{2}\delta_1\gamma_1^2 &= -\beta_0 - \tfrac{1}{2}nv_n - \tfrac{1}{2}\delta_0\gamma_0^2 - \tfrac{1}{2}n\bar{x}_n^2 \\ \delta_1\gamma_1 &= \delta_0\gamma_0 + n\bar{x}_n \\ -\tfrac{1}{2}\delta_1 &= -\tfrac{1}{2}\delta_0 - \tfrac{1}{2}n \end{aligned}$$

Hence

$$\begin{aligned} \delta_1 &= \delta_0 + n \\ \gamma_1 &= \frac{\delta_0\gamma_0 + n\bar{x}_n}{\delta_0 + n} \end{aligned}$$

Two-Parameter Normal (cont.)

And the last hyperparameter comes from

$$-\beta_1 - \tfrac{1}{2}\delta_1\gamma_1^2 = -\beta_0 - \tfrac{1}{2}nv_n - \tfrac{1}{2}\delta_0\gamma_0^2 - \tfrac{1}{2}n\bar{x}_n^2$$

so

$$\begin{aligned}
\beta_1 &= \beta_0 + \tfrac{1}{2}n(v_n + \bar{x}_n^2) + \tfrac{1}{2}\delta_0\gamma_0^2 - \tfrac{1}{2}\delta_1\gamma_1^2 \\
&= \beta_0 + \tfrac{1}{2}n(v_n + \bar{x}_n^2) + \tfrac{1}{2}\delta_0\gamma_0^2 - \tfrac{1}{2}\frac{(\delta_0\gamma_0 + n\bar{x}_n)^2}{\delta_0 + n} \\
&= \beta_0 + \tfrac{1}{2}nv_n + \frac{(\delta_0\gamma_0^2 + n\bar{x}_n^2)(\delta_0 + n) - (\delta_0\gamma_0 + n\bar{x}_n)^2}{2(\delta_0 + n)} \\
&= \beta_0 + \frac{n}{2}\left(v_n + \frac{\delta_0(\bar{x}_n - \gamma_0)^2}{\delta_0 + n}\right)
\end{aligned}$$

Two-Parameter Normal (cont.)

And that finishes the proof of the following theorem. The normal-gamma family is conjugate to the two-parameter normal. If the prior is normal-gamma with hyperparameters α_0, β_0, γ_0, and δ_0, then the posterior is normal-gamma with hyperparameters

$$\begin{aligned}
\alpha_1 &= \alpha_0 + \frac{n}{2} \\
\beta_1 &= \beta_0 + \frac{n}{2} \left(v_n + \frac{\delta_0 (\bar{x}_n - \gamma_0)^2}{\delta_0 + n} \right) \\
\gamma_1 &= \frac{\gamma_0 \delta_0 + n \bar{x}_n}{\delta_0 + n} \\
\delta_1 &= \delta_0 + n
\end{aligned}$$

Two-Parameter Normal (cont.)

We are also interested in the other factorization of the normal-gamma conjugate family: marginal for μ and conditional for λ given μ. The unnormalized joint distribution is

$$\lambda^{\alpha-1} \exp\bigl(-\beta\lambda\bigr) \cdot \lambda^{1/2} \exp\left(-\tfrac{1}{2}\delta\lambda(\mu-\gamma)^2\right)$$

Considered as a function of λ for fixed μ, this is proportional to a $\text{Gam}(a, b)$ distribution with hyperparameters

$$\begin{aligned} a &= \alpha + 1/2 \\ b &= \beta + \tfrac{1}{2}\delta(\mu-\gamma)^2 \end{aligned}$$

Two-Parameter Normal (cont.)

The normalized PDF for this conditional is

$$\frac{1}{\Gamma(\alpha + 1/2)} (\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2)^{\alpha+1/2} \\ \times \lambda^{\alpha+1/2-1} \exp\left(-\left(\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2\right)\lambda\right)$$

We conclude the unnormalized marginal for μ must be

$$(\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2)^{-(\alpha+1/2)}$$

Two-Parameter Normal (cont.)

This marginal is not obviously a brand-name distribution, but we claim it is a location-scale transformation of a t distribution with noninteger degrees of freedom. Dropping constants, the unnormalized PDF of the t distribution with ν degrees of freedom is

$$(\nu + x^2)^{-(\nu+1)/2}$$

(brand name distributions handout). If we change the variable to $\mu = a + bx$, so $x = (\mu - a)/b$, we get

$$\left[\nu + \left(\frac{\mu - a}{b}\right)^2\right]^{-(\nu+1)/2} \propto [\nu b^2 + (\mu - a)^2]^{-(\nu+1)/2}$$

Two-Parameter Normal (cont.)

So to identify the marginal we much match up

$$[\nu b^2 + (\mu - a)^2]^{-(\nu+1)/2}$$

and

$$(\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2)^{-(\alpha+1/2)} \propto [2\beta/\delta + (\mu - \gamma)^2]^{-(\alpha+1/2)}$$

and these do match up with

$$\begin{aligned} \nu &= 2\alpha \\ a &= \gamma \\ b &= \sqrt{\frac{\beta}{\alpha\delta}} \end{aligned}$$

And that finishes the proof of the following theorem. The other factorization of the normal-gamma family is gamma-t-location-scale.

Two-Parameter Normal (cont.)

If

$$\begin{aligned} \lambda &\sim \text{Gam}(\alpha, \beta) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma, \delta^{-1}\lambda^{-1}) \end{aligned}$$

then

$$\begin{aligned} (\mu - \gamma)/d &\sim t(\nu) \\ \lambda \mid \mu &\sim \text{Gam}(a, b) \end{aligned}$$

where

$$\begin{aligned} a &= \alpha + \tfrac{1}{2} \\ b &= \beta + \tfrac{1}{2}\delta(\mu - \gamma)^2 \\ \nu &= 2\alpha \\ d &= \sqrt{\frac{\beta}{\alpha\delta}} \end{aligned}$$

Two-Parameter Normal (cont.)

Thus the Bayesian also gets t distributions. They are marginal posteriors of μ for normal data when conjugate priors are used.

Two-Parameter Normal (cont.)

The unnormalized normal-gamma prior is

$$\lambda^{\alpha_0 - 1} \exp\bigl(-\beta_0 \lambda\bigr) \cdot \lambda^{1/2} \exp\left(-\tfrac{1}{2}\delta_0 \lambda (\mu - \gamma_0)^2\right)$$

Set $\beta_0 = \delta_0 = 0$ and we get the improper prior

$$g(\mu, \lambda) = \lambda^{\alpha_0 - 1/2}$$

Two-Parameter Normal (cont.)

The Jeffreys prior for the two-parameter normal with parameters μ and $\nu = 1/\lambda$ is

$$g(\mu, \nu) \propto \nu^{-3/2}$$

(slide 55). The change-of-variable to μ and λ gives Jacobian

$$\left| \det \begin{pmatrix} 1 & 0 \\ 0 & \partial\nu/\partial\lambda \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & 0 \\ 0 & -1/\lambda^2 \end{pmatrix} \right| = \lambda^{-2}$$

Hence the Jeffreys prior for μ and λ is

$$g(\mu, \lambda) = \left(\frac{1}{\lambda} \right)^{-3/2} \cdot \lambda^{-2} = \lambda^{3/2} \cdot \lambda^{-2} = \lambda^{-1/2}$$

This matches up with what we had on the preceding slide if we take $\alpha_0 = 0$.

Two-Parameter Normal (cont.)

The Jeffreys prior has $\alpha_0 = \beta_0 = \delta_0$. Then γ_0 is irrelevant.

This produces the posterior with hyperparameters

$$\begin{aligned} \alpha_1 &= \frac{n}{2} \\ \beta_1 &= \frac{n v_n}{2} \\ \gamma_1 &= \bar{x}_n \\ \delta_1 &= n \end{aligned}$$

Two-Parameter Normal (cont.)

The marginal posterior for λ is

$$\text{Gam}(\alpha_1, \beta_1) = \text{Gam}\left(\frac{n}{2}, \frac{n v_n}{2}\right)$$

The marginal posterior for μ is

$$(\mu - \gamma_1)/d \sim t(\nu)$$

where $\gamma_1 = \bar{x}_n$ and

$$d = \sqrt{\frac{\beta_1}{\alpha_1 \delta_1}} = \sqrt{\frac{n v_n / 2}{n/2 \cdot n}} = \sqrt{\frac{v_n}{n}}$$

and $\nu = 2\alpha_1 = n$.

Two-Parameter Normal (cont.)

In summary the Bayesian using the Jeffreys prior gets

$$
\begin{aligned}
\lambda &\sim \text{Gam}\left(\frac{n}{2}, \frac{n v_n}{2}\right) \\
\frac{\mu - \bar{x}_n}{\sqrt{v_n / n}} &\sim t(n)
\end{aligned}
$$

Two-Parameter Normal (cont.)

Alternatively, setting $\alpha_0 = -1/2$ and $\beta_0 = \delta_0 = 0$ gives

$$
\begin{aligned}
\alpha_1 &= \frac{n-1}{2} \\
\beta_1 &= \frac{n v_n}{2} = \frac{(n-1) s_n^2}{2} \\
\gamma_1 &= \bar{x}_n \\
\delta_1 &= n \\
d &= \sqrt{\frac{\beta_1}{\alpha_1 \delta_1}} = \sqrt{\frac{n v_n / 2}{(n-1)/2 \cdot n}} = \frac{s_n}{\sqrt{n}}
\end{aligned}
$$

where $s_n^2 = n v_n / (n-1)$ is the usual sample variance.

Two-Parameter Normal (cont.)

So the Bayesian with this improper prior almost agrees with the frequentist. The marginal posteriors are

$$\lambda \sim \text{Gam}\left(\frac{n-1}{2}, \frac{(n-1)s_n^2}{2}\right)$$

$$\frac{\mu - \bar{x}_n}{s_n/\sqrt{n}} \sim t(n-1)$$

or

$$\frac{\mu - \bar{x}_n}{s_n/\sqrt{n}} \sim t(n-1)$$

$$(n-1)s_n^2\lambda \sim \text{chi}^2(n-1)$$

But there is no reason for the Bayesian to choose $\alpha_0 = -1/2$ except to match the frequentist.

Bayesian Point Estimates

Bayesians have little interest in point estimates of parameters. To them a parameter is a random variable, and what is important is its distribution. A point estimate is a meager bit of information as compared, for example, to a plot of the posterior density.

Frequentists too have little interest in point estimates except as tools for constructing tests and confidence intervals.

However, Bayesian point estimates are something you are expected to know about if you have taken a course like this.

Bayesian Point Estimates (cont.)

Bayesian point estimates are properties of the posterior distribution. The three point estimates that are widely used are the posterior mean, the posterior median, and the posterior mode.

We already know what the mean and median of a distribution are.

A *mode* of a continuous distribution is a local maximum of the PDF. The distribution is *unimodal* if it has one mode, *bimodal* if two, and *multimodal* if more than one.

When we say *the* mode (rather than *a* mode) in reference to a multimodal distribution, we mean the highest mode.

Bayesian Point Estimates (cont.)

Finding the modes of a distribution is somewhat like maximum likelihood, except one differentiates with respect to the variable rather than with respect to the parameter. For a Bayesian, however, the variable is the parameter. So it is just like maximum likelihood except that instead of maximizing

$$L_n(\theta)$$

one maximizes

$$L_n(\theta)g(\theta)$$

or one can maximize

$$\log\Bigl[L_n(\theta)g(\theta)\Bigr] = l_n(\theta) + \log g(\theta)$$

(log likelihood + log prior).

Bayesian Point Estimates (cont.)

Suppose the data x is Bin(n, p) and we use the conjugate prior Beta(α_1, α_2), so the posterior is Beta$(x + \alpha_1, n - x + \alpha_2)$ (slide 6).

Looking up the mean of a beta distribution on the brand name distributions handout, we see the posterior mean is

$$E(p \mid x) = \frac{x + \alpha_1}{n + \alpha_1 + \alpha_2}$$

Bayesian Point Estimates (cont.)

The posterior median has no simple expression. We can calculate it using the R expression

```
qbeta(0.5, x + alpha1, n - x + alpha2)
```

assuming `x`, `n`, `alpha1`, and `alpha2` have been defined.

Bayesian Point Estimates (cont.)

The posterior mode is the maximizer of

$$h(p \mid x) = p^{x + \alpha_1 - 1} (1 - p)^{n - x + \alpha_2 - 1}$$

or of

$$\log h(p \mid x) = (x + \alpha_1 - 1) \log(p) + (n - x + \alpha_2 - 1) \log(1 - p)$$

which has derivative

$$\frac{d}{dp} \log h(p \mid x) = \frac{x + \alpha_1 - 1}{p} - \frac{n - x + \alpha_2 - 1}{1 - p}$$

setting this equal to zero and solving for p gives

$$\frac{x + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}$$

for the posterior mode.

Bayesian Point Estimates (cont.)

The formula

$$\frac{x + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}$$

is only valid if it gives a number between zero and one. If $x + \alpha_1 < 1$ then the posterior PDF goes to infinity as $p \downarrow 0$. If $n - x + \alpha_2 < 1$ then the posterior PDF goes to infinity as $p \uparrow 1$.

Bayesian Point Estimates (cont.)

Suppose $\alpha_1 = \alpha_2 = 1/2$, $x = 0$, and $n = 10$.

```
Rweb:> alpha1 <- alpha2 <- 1 / 2
Rweb:> x <- 0
Rweb:> n <- 10
Rweb:> (x + alpha1) / (n + alpha1 + alpha2)
[1] 0.04545455
Rweb:> qbeta(0.5, x + alpha1, n - x + alpha1)
[1] 0.02194017
Rweb:> (x + alpha1 - 1) / (n + alpha1 + alpha2 - 2)
[1] -0.05555556
```

Posterior mean: 0.045. Posterior median: 0.022. Posterior mode: 0.

Bayesian Point Estimates (cont.)

Suppose $\alpha_1 = \alpha_2 = 1/2$, $x = 2$, and $n = 10$.

```
Rweb:> alpha1 <- alpha2 <- 1 / 2
Rweb:> x <- 2
Rweb:> n <- 10
Rweb:> (x + alpha1) / (n + alpha1 + alpha2)
[1] 0.2272727
Rweb:> qbeta(0.5, x + alpha1, n - x + alpha1)
[1] 0.2103736
Rweb:> (x + alpha1 - 1) / (n + alpha1 + alpha2 - 2)
[1] 0.1666667
```

Posterior mean: 0.227. Posterior median: 0.210. Posterior mode: 0.167.

Bayesian Point Estimates (cont.)

In one case, the calculations are trivial. If the posterior distribution is symmetric and unimodal, for example normal or t-location-scale, then the posterior mean, median, mode, and center of symmetry are equal.

When we have normal data and use the normal-gamma prior, the posterior mean, median, and mode is

$$\gamma_1 = \frac{\gamma_0 \delta_0 + n \bar{x}_n}{\delta_0 + n}$$

Bayesian Point Estimates (cont.)

The posterior mean and median are often woofed about using decision-theoretic terminology. The posterior mean is the *Bayes estimator that minimizes squared error loss*. The posterior median is the *Bayes estimator that minimizes absolute error loss*. The posterior mode is the Bayes estimator that minimizes the loss

$$t \mapsto E\{1 - I_{(-\epsilon,\epsilon)}(t - \theta) \mid \text{data}\}$$

when ϵ is infinitesimal.

Bayesian Asymptotics

Bayesian asymptotics are a curious mix of Bayesian and frequentist reasoning.

Like the frequentist we assume there is a true unknown parameter value $\boldsymbol{\theta}_0$ and X_1, X_2, $\ldots$ IID from the distribution having parameter value $\boldsymbol{\theta}_0$. Like the Bayes we calculate posterior distributions

$$h_n(\boldsymbol{\theta}) = h(\boldsymbol{\theta} \mid x_1, \ldots, x_n) \propto L_n(\boldsymbol{\theta}) g(\boldsymbol{\theta})$$

and look a posterior distributions of θ for larger and larger sample sizes.

Bayesian Asymptotics (cont.)

Bayesian asymptotic analysis is similar to frequentist analysis but more complicated. We omit the proof and just give the results.

If the prior PDF is continuous and strictly positive at $\boldsymbol{\theta}_0$, and all the frequentist conditions for asymptotics of maximum likelihood are satisfied and some extra assumptions about the tails of the posterior are small enough, the the Bayesian agrees with the frequentist

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \approx \mathcal{N}\Bigl(0, \mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\Bigr)$$

Of course, the Bayesian and frequentist disagree about what is random on the left-hand side. The Bayesian says $\boldsymbol{\theta}$ and the frequentist says $\hat{\boldsymbol{\theta}}_n$. But they agree about the asymptotic distribution.

Bayesian Asymptotics (cont.)

Several important points here.

As the sample size gets large, the influence of the prior diminishes (so long as the prior PDF is continuous and positive near the true parameter value).

The Bayesian and frequentist disagree about philosophical woof. They don't exactly agree about inferences, but they do approximately agree when the sample size is large.

Bayesian Credible Intervals

Not surprisingly, when a Bayesian makes an interval estimate, it is based on the posterior.

Many Bayesians do not like to call such things "confidence intervals" because that names a frequentist notion. Hence the name "credible intervals" which is clearly something else.

One way to make credible intervals is to find the marginal posterior distribution for the parameter of interest and find its $\alpha/2$ and $1 - \alpha/2$ quantiles. The interval between them is a $100(1 - \alpha)\%$ Bayesian credible interval for the parameter of interest called the *equal tailed interval*.

Bayesian Credible Intervals (cont.)

Another way to make credible intervals is to find the marginal posterior distribution $h(\theta \mid \mathbf{x})$ for the parameter of interest and find the level set

$$A_\gamma = \{\, \theta \in \Theta : h(\theta \mid \mathbf{x}) > \gamma \,\}$$

that has the required probability

$$\int_{A_\gamma} h(\theta \mid \mathbf{x}) \, d\theta = 1 - \alpha$$

A_γ is a $100(1 - \alpha)$% Bayesian credible region, not necessarily an interval, for the parameter of interest called the *highest posterior density region* (HPD region).

These are not easily done, even with a computer. See computer examples web pages for example.

Bayesian Credible Intervals (cont.)

Equal tailed intervals transform by invariance under monotone change of parameter. If (a, b) is an equal tailed Bayesian credible interval for θ and $\theta = h(\psi)$, where h is an increasing function, then $\bigl(h(a), h(b)\bigr)$ is an equal tailed Bayesian credible interval for ψ with the same confidence level.

The analogous fact does not hold for highest posterior density regions because the change-of-parameter involves a Jacobian.

Despite the fact that HPD regions do not transform sensibly under change-of-parameter, they seem to be preferred by most Bayesians.

Bayesian Hypothesis Tests

Not surprisingly, when a Bayesian does a hypothesis test, it is based on the posterior.

To a Bayesian, a hypothesis is an event, a subset of the sample space. Remember that after the data are seen, the Bayesian considers only the parameter random. So the parameter space and the sample space are the same thing to the Bayesian.

The Bayesian compares hypotheses by comparing their posterior probabilities.

All but the simplest such tests must be done by computer.

Bayesian Hypothesis Tests (cont.)

Suppose the data x are Bin(n, p) and the prior is Beta(α_1, α_2), so the posterior is Beta$(x + \alpha_1, n - x + \alpha_2)$.

Suppose the hypotheses in question are

$$\begin{aligned} H_0 &: p > 1/2 \\ H_1 &: p \le 1/2 \end{aligned}$$

We can calculate the probabilities of these two hypotheses by the the R expressions

```
pbeta(0.5, x + alpha1, n - x + alpha2)
pbeta(0.5, x + alpha1, n - x + alpha2, lower.tail = FALSE)
```

assuming `x`, `n`, `alpha1`, and `alpha2` have been defined.

Bayesian Hypothesis Tests

```
Rweb:> alpha1 <- alpha2 <- 1 / 2
Rweb:> x <- 2
Rweb:> n <- 10
Rweb:> pbeta(0.5, x + alpha1, n - x + alpha2)
[1] 0.9739634
Rweb:> pbeta(0.5, x + alpha1, n - x + alpha2, lower.tail = FALSE)
[1] 0.02603661
```

Bayesian Hypothesis Tests (cont.)

Bayes tests get weirder when the hypotheses have different dimensions.

In principle, there is no reason why a prior distribution has to be continuous. It can have degenerate parts that put probability on sets a continuous distribution would give probability zero. But many users find this weird. Hence the following scheme, which is equivalent, but doesn't sound as strange.

Bayes Factors

Let $\mathcal{M}$ be a finite or countable set of models. For each model $m \in \mathcal{M}$ we have the prior probability of the model $h(m)$. It does not matter if this prior on models is unnormalized.

Each model m has a parameter space Θ_m and a prior

$$g(\theta \mid m), \qquad \theta \in \Theta_m$$

The spaces Θ_m can and usually do have different dimensions. That's the point. These within model priors must be normalized proper priors. The calculations to follow make no sense if these priors are unnormalized or improper.

Each model m has a data distribution

$$f(x \mid \theta, m)$$

which may be a PDF or PMF.

Bayes Factors (cont.)

The unnormalized posterior for everything, models and parameters within models, is

$$f(x \mid \theta, m) g(\theta \mid m) h(m)$$

To obtain the conditional distribution of x given m, we must integrate out the nuisance parameters θ

$$\begin{aligned} q(x \mid m) &= \int_{\Theta_m} f(x \mid \theta, m) g(\theta \mid m) h(m) \, d\theta \\ &= h(m) \int_{\Theta_m} f(x \mid \theta, m) g(\theta \mid m) \, d\theta \end{aligned}$$

These are the unnormalized posterior probabilities of the models. The normalized probabilities are

$$p(m \mid x) = \frac{q(x \mid m)}{\sum_{m \in \mathcal{M}} q(x \mid m)}$$

Bayes Factors (cont.)

It is considered useful to define

$$b(x \mid m) = \int_{\Theta_m} f(x \mid \theta, m) g(\theta \mid m) \, d\theta$$

so

$$q(x \mid m) = b(x \mid m) h(m)$$

Then the ratio of posterior probabilities of models m_1 and m_2 is

$$\frac{p(m_1 \mid x)}{p(m_2 \mid x)} = \frac{q(x \mid m_1)}{q(x \mid m_2)} = \frac{b(x \mid m_1)}{b(x \mid m_2)} \cdot \frac{h(m_1)}{h(m_2)}$$

This ratio is called the *posterior odds* of the models (a ratio of probabilities is called an *odds*) of these models.

The *prior odds* is

$$\frac{h(m_1)}{h(m_2)}$$

Bayes Factors (cont.)

The term we have not yet named in

$$\frac{p(m_1 \mid x)}{p(m_2 \mid x)} = \frac{b(x \mid m_1)}{b(x \mid m_2)} \cdot \frac{h(m_1)}{h(m_2)}$$

is called the *Bayes factor*

$$\frac{b(x \mid m_1)}{b(x \mid m_2)}$$

the ratio of posterior odds to prior odds.

The prior odds tells how the prior compares the probability of the models. The Bayes factor tells us how the data shifts that comparison going from prior to posterior via Bayes rule.

Bayes Factors (cont.)

Suppose the data x are Bin(n, p) and the models (hypotheses) in question are

$$\begin{aligned} m_1 &\colon p = 1/2 \\ m_2 &\colon p \neq 1/2 \end{aligned}$$

The model m_1 is concentrated at one point $p = 1/2$, hence has no nuisance parameter. Hence $g(\theta \mid m_1) = 1$. Suppose we use the within model prior Beta(α_1, α_2) for model m_2.

Then

$$b(x \mid m_1) = f(x \mid 1/2) = \binom{n}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{n-x} = \binom{n}{x} \left(\frac{1}{2}\right)^n$$

Bayes Factors (cont.)

And

$$\begin{aligned} b(x \mid m_2) &= \int_0^1 f(x \mid p) g(p \mid m_2) \, dp \\ &= \int_0^1 \binom{n}{x} \frac{1}{B(\alpha_1, \alpha_2)} p^{x + \alpha_1 - 1} (1 - p)^{n - x + \alpha_2 - 1} \, dp \\ &= \binom{n}{x} \frac{B(x + \alpha_1, n - x + \alpha_2)}{B(\alpha_1, \alpha_2)} \end{aligned}$$

where

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

by the theorem for the beta distribution (brand name distributions handout).

Bayes Factors (cont.)

```
Rweb:> alpha1 <- alpha2 <- 1 / 2
Rweb:> x <- 2
Rweb:> n <- 10
Rweb:> p0 <- 1 / 2
Rweb:> b1 <- dbinom(x, n, p0)
Rweb:> b2 <- choose(n, x) * beta(x + alpha1, n - x + alpha2) /
+      beta(alpha1, alpha2)
Rweb:> ##### Bayes factor
Rweb:> b1 / b2
[1] 0.5967366
Rweb:> ##### frequentist P-value
Rweb:> 2 * pbinom(x, n, p0)
[1] 0.109375
```

Bayes Factors (cont.)

For comparison, we calculated not only the Bayes factor 0.597 but also the frequentist P-value 0.109. Bayes factors and P-values are sort of comparable, but as can be seen are not identical. In fact, it is a theorem that in situations like this the Bayes factor is always larger than the P-value, at least asymptotically. This makes Bayesian tests more conservative, less likely to reject the null hypothesis, than frequentists.

Either the frequentists are too optimistic or the Bayesians are too conservative, or perhaps both. The jury is still out on that.

Bayes Factors (cont.)

Now we try two-sample procedures. The data are X_i distributed Bin(n_i, p_i) for $i = 1$, 2. We start with the two-tailed test with models

$$
\begin{aligned}
m_1 &\colon p_1 = p_2 \\
m_2 &\colon p_1 \neq p_2
\end{aligned}
$$

Suppose for model 2 we use independent priors on the parameters with p_i distributed Beta(α_1, α_2). Then

$$
\begin{aligned}
b(\mathbf{x} \mid m_2) &= \prod_{i=1}^{2} \int_0^1 \binom{n_i}{x_i} \frac{1}{B(\alpha_1, \alpha_2)} p^{x_i + \alpha_1 - 1} (1 - p_i)^{n_i - x_i + \alpha_2 - 1} \, dp_i \\
&= \prod_{i=1}^{2} \binom{n_i}{x_i} \frac{B(x_i + \alpha_1, n_i - x_i + \alpha_2)}{B(\alpha_1, \alpha_2)}
\end{aligned}
$$

Bayes Factors (cont.)

It is not obvious how to choose the within model prior for model m_1. Here is one idea. Consider the unnormalized joint prior for model m_2

$$p_1^{\alpha_1 - 1} (1 - p_1)^{\alpha_2 - 1} p_2^{\alpha_1 - 1} (1 - p_2)^{\alpha_2 - 1}$$

and set $p_1 = p_2$ obtaining

$$p^{2 \alpha_1 - 2} (1 - p)^{2 \alpha_2 - 2}$$

This is the restriction of the unnormalized PDF for the prior for m_2 to the support of m_1. We recognize that as an unnormalized Beta$(2 \alpha_1 - 1, 2 \alpha_2 - 1)$ prior. However, this doesn't work if $\alpha_1 \le 1/2$ or $\alpha_2 \le 1/2$, because the result is improper. So it doesn't work for the Jeffreys prior.

Bayes Factors (cont.)

Uncertain about what prior to use for m_2, call it Beta(α_3, α_4)

Using this prior for m_1 we get

$$
\begin{aligned}
b(\mathbf{x} \mid m_1) &= \int_0^1 \binom{n_1}{x_1}\binom{n_2}{x_2}\frac{1}{B(\alpha_3, \alpha_4)} \\
&\qquad p^{x_1+x_2+\alpha_3-1}(1-p)^{n_1+n_2-x_1-x_2+\alpha_4-1}\,dp \\
&= \binom{n_1}{x_1}\binom{n_2}{x_2}\frac{B(x_1+x_2+\alpha_3, n_1+n_2-x_1-x_2+\alpha_4)}{B(\alpha_3, \alpha_4)}
\end{aligned}
$$

Bayes Factors (cont.)

Now consider the one-tailed test for the same data with models

$$\begin{aligned} m_1 &\colon p_1 \geq p_2 \\ m_2 &\colon p_1 < p_2 \end{aligned}$$

Using Beta(α_1, α_2) for both parameters

$$b(\mathbf{x} \mid m_j) = \\ \iint_{m_j} \prod_{i=1}^{2} \binom{n_i}{x_i} \frac{1}{B(\alpha_1, \alpha_2)} p^{x_i + \alpha_1 - 1} (1 - p_i)^{n_i - x_i + \alpha_2 - 1} \, dp_1 \, dp_2$$

but we cannot do the integrals except by simulation.

Bayes Factors (cont.)

We simulate from the posterior we would use if we weren't doing tests p_1 and p_2 are independent and p_i is Beta$(x_i + \alpha_1, n_i - x_i + \alpha_2)$.

We rewrite the integrals on the preceding page as expectations

$$b(\mathbf{x} \mid m_j) = \text{Pr}(m_j \mid x_1, x_2) \prod_{i=1}^{2} \binom{n_i}{x_i} \frac{B(x_i + \alpha_1, n_i - x_i + \alpha_2)}{B(\alpha_1, \alpha_2)}$$

Ordinary Monte Carlo

The "Monte Carlo method" refers to the theory and practice of learning about probability distributions by simulation rather than calculus. In ordinary Monte Carlo (OMC) we use IID simulations from the distribution of interest. Suppose X_1, X_2, $\ldots$ are IID simulations from some distribution, and suppose we want to know an expectation

$$\mu = E\{g(X_i)\}.$$

Then the law of large numbers (LLN) then says

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

converges in probability to μ.

Ordinary Monte Carlo (cont.)

The central limit theorem (CLT) says

$$\sqrt{n}(\widehat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \text{var}\{g(X_i)\}$$

which can be estimated by the empirical variance

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \bigl(g(X_i) - \widehat{\mu}_n\bigr)^2$$

An asymptotic 95% confidence interval for μ is

$$\widehat{\mu}_n \pm 1.96 \frac{\widehat{\sigma}_n}{\sqrt{n}}$$

Ordinary Monte Carlo (cont.)

The theory of OMC is just the theory of frequentist statistical inference. The only differences are that

- the “data” X_1, . . ., X_n are computer simulations rather than measurements on objects in the real world,
- the “sample size” n is the number of computer simulations rather than the size of some real world data, and
- the unknown parameter μ is in principle completely known, given by some integral, which we are unable to do.

Ordinary Monte Carlo (cont.)

Everything works just the same when the data $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$, which are computer simulations are vectors. But the functions of interest $g(\mathbf{X}_1)$, $g(\mathbf{X}_2)$, $\ldots$ are scalars.

OMC works great, but is very hard to do when $\mathbf{X}_i$ is vector.

Hence Markov chain Monte Carlo (MCMC).

Markov Chain Monte Carlo

A Markov chain is a dependent sequence of random variables X_1, X_2, $\ldots$ or random vector $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ having the property that the future is independent of the past given the present: the conditional distribution of $\mathbf{X}_{n+1}$ given $\mathbf{X}_1$, $\ldots$, $\mathbf{X}_n$ depends only on $\mathbf{X}_n$.

The Markov chain has *stationary transition probabilities* if the conditional distribution of $\mathbf{X}_{n+1}$ given $\mathbf{X}_n$ is the same for all n. Every Markov chain used in MCMC has this property.

The joint distribution of $\mathbf{X}_1$, $\ldots$, $\mathbf{X}_n$ is determined by the *initial distribution* of the Markov chain — marginal distribution of $\mathbf{X}_1$ — and the *transition probabilities* — the conditional distributions of $\mathbf{X}_{n+1}$ given $\mathbf{X}_n$.

Markov Chain Monte Carlo (cont.)

A *scalar functional* of a Markov chain $\mathbf{X}_1$, $\mathbf{X}_2$, ... is a time series $g(\mathbf{X}_1)$, $g(\mathbf{X}_2)$, ..., where g is a scalar-valued function on the state space of the Markov chain.

An initial distribution of a Markov chain is called *stationary* or *invariant* or *equilibrium* if the resulting Markov chain is a stationary stochastic process, in which case any scalar functional is a stationary time series (5101, deck 2, slides 101–109). This means the joint distribution of the k-tuple

$$(\mathbf{X}_{i+1}, \mathbf{X}_{i+2}, \ldots, \mathbf{X}_{i+k})$$

is the same for all i, and that this holds for all positive integers k. Similarly, the joint distribution of the k-tuple

$$\Bigl(g(\mathbf{X}_{i+1}), g(\mathbf{X}_{i+2}), \ldots, g(\mathbf{X}_{i+k})\Bigr)$$

is the same for all i and k and all real-valued functions g.

Markov Chain Monte Carlo (cont.)

A Markov chain is *stationary* if its initial distribution is stationary.

This is different from having stationary transition probabilities. All chains used in MCMC have stationary transition probabilities. None are exactly stationary.

Markov Chain Monte Carlo (cont.)

To be (exactly) stationary, must start the chain with simulation from the equilibrium (invariant, stationary) distribution. Can never do that except in toy problems.

If chain is stationary, then every iterate $\mathbf{X}_i$ has the same marginal distribution, which is the equilibrium distribution.

If chain is not stationary but has a unique equilibrium distribution, which includes chains used in MCMC, then the marginal distribution $\mathbf{X}_i$ converges to the equilibrium distribution as $i \to \infty$.

Markov Chain Monte Carlo (cont.)

Suppose $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ are simulation from a Markov chain having a unique equilibrium distribution, and suppose we want to know an expectation

$$\mu = E_{\text{equilib}}\{g(\mathbf{X})\},$$

where the subscript "equilib" refers to this unique equilibrium distribution. Then the law of large numbers (LLN) for Markov chains then says

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i)$$

converges in probability to μ.

Markov Chain Monte Carlo (cont.)

The central limit theorem (CLT) for Markov chains says

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \text{var}_{\text{stationary}}\{g(\mathbf{X}_i)\} + 2 \sum_{k=1}^{\infty} \text{cov}_{\text{stationary}}\{g(\mathbf{X}_i), g(\mathbf{X}_{i+k})\}$$

which can be estimated in various ways, where the subscript "stationary" refers the stationary chain whose initial distribution is the unique equilibrium distribution.

Although the iterates of the Markov chain are neither independent nor identically distributed — the chain converges to equilibrium but never gets there — the CLT still holds if the infinite sum defining σ^2 is finite and chain is reversible.

Markov Chain Monte Carlo (cont.)

All chains used in MCMC can be made reversible (a term we don't define). All the chains produced by the R function `metrop` in the contributed package `mcmc`, which are the only ones we show in the handout, are reversible.

Verifying the infinite sum is finite is theoretically possible for some simple applications, but generally so hard that it cannot be done. Nevertheless, we expect the CLT to hold in practice and it does seem to whenever enough simulation is done to check.

Batch Means

In order to make MCMC practical, need a method to estimate the variance σ^2 in the CLT, then can proceed just like in OMC.

If $\widehat{\sigma}_n^2$ is a consistent estimate of σ^2, then an asymptotic 95% confidence interval for μ is

$$\widehat{\mu}_n \pm 1.96 \frac{\widehat{\sigma}_n}{\sqrt{n}}$$

The method of batch means estimates the asymptotic variance for a stationary time series.

Batch Means (cont.)

Markov chain CLT says

$$\widehat{\mu}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Suppose b evenly divides n and we form the means

$$\widehat{\mu}_{b,k} = \frac{1}{b} \sum_{i=bk+1}^{bk+b} g(X_i)$$

for $k = 1, \ldots, m = n/b$. Each of these "batch means" satisfies

$$\widehat{\mu}_{k,b} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{b}\right)$$

if b is sufficiently large.

Batch Means (cont.)

Thus empirical variance of the sequence of batch means

$$\frac{1}{m} \sum_{k=1}^{m} (\widehat{\mu}_{b,k} - \widehat{\mu}_n)^2$$

estimates σ^2 / b. And b / n times this estimates σ^2 / n, the asymptotic variance of $\widehat{\mu}_n$.

Metropolis Algorithm

Suppose we are interested in a continuous random vector having unnormalized PDF h. This means

$$h(\mathbf{x}) \geq 0, \qquad \mathbf{x} \in \mathbb{R}^d$$

and

$$\int h(\mathbf{x}) \, d\mathbf{x} < \infty$$

(this being a d-dimensional integral).

Here is how to simulate one step of a Markov chain having equilibrium distribution having unnormalized PDF h.

Metropolis Algorithm (cont.)

Suppose the current state of the Markov chain is $\mathbf{X}_n$. Then the next step of the Markov chain $\mathbf{X}_{n+1}$ is simulated as follows.

- Simulate $\mathbf{Y}_n$ having a $\mathcal{N}(0, \mathbf{M})$ distribution.
- Calculate
$$r = \frac{h(\mathbf{X}_n + \mathbf{Y}_n)}{h(\mathbf{X}_n)}$$
- Simulate U_n having a Unif$(0, 1)$ distribution.
- If $U_n < r$, set $\mathbf{X}_{n+1} = \mathbf{X}_n + \mathbf{Y}_n$. Otherwise set $\mathbf{X}_{n+1} = \mathbf{X}_n$.

Metropolis Algorithm (cont.)

Only thing that can be adjusted and must be adjusted is "proposal" variance matrix $\mathbf{M}$.

For simplicity, handout uses $\mathbf{M} = \tau^2 \mathbf{I}$, where $\mathbf{I}$ is identity matrix. Then only one constant τ to adjust.

Now go to handout.