Univariate Normal Model

Our first example of a multiparameter model is the situation where we have a random sample $Y_1, \ldots, Y_n$ from a normal distribution, $N(\mu, \sigma^2)$, which has density

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right).$$

Here the parameter vector $\boldsymbol{\theta}$ is (μ, σ^2).

We start with this setting since normal models for data are so prevalent in practice, being used in regression, analysis of variance and many other settings.

Most of the interesting and/or challenging problems in statistics involve more than one parameter.

The principles we've talked about thus far apply regardless of the number of parameters the model has, but there are also some additional twists in multiparameter models.

Nuisance parameters

Sometimes only a subset of the parameters in $\boldsymbol{\theta}$ are of any interest. The others may be regarded as *nuisance* parameters.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_I, \boldsymbol{\theta}_N)$, where $\boldsymbol{\theta}_I$ contains the parameters of interest and $\boldsymbol{\theta}_N$ is the vector of nuisance parameters.

We may compute the posterior for the parameters of interest by averaging the posterior over all values of the nuisance parameters.

$$p_I(\boldsymbol{\theta}_I|\boldsymbol{y}) = \int p(\boldsymbol{\theta}_I, \boldsymbol{\theta}_N|\boldsymbol{y})\, d\boldsymbol{\theta}_N.$$

Conditional distributions

The conditional distribution of some subset of parameters given the values of other parameters is often of interest.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The conditional posterior of $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$ is

$$p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p_2(\boldsymbol{\theta}_2|\boldsymbol{y})},$$

where $p_2(\boldsymbol{\theta}_2|\boldsymbol{y})$ is the marginal posterior of $\boldsymbol{\theta}_2$.

We turn now to the normal model. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n$ are a random sample from $N(\theta_1, 1/\theta_2)$.

$$\Theta = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, \theta_2 > 0\}$$

Here I have chosen a parameterization where θ_1 is the mean of the distribution, and θ_2 is the so-called *precision*, or the *reciprocal of the variance*.

A motivation for this parameterization is that the *normal-gamma* family is then a conjugate family of priors for the model.

Normal-gamma family of distributions

Suppose that θ_2 has a gamma(a, b) distribution and that $\theta_1 | \theta_2$ is $N(\mu, (\tau\theta_2)^{-1})$. Then we say that the joint distribution of (θ_1, θ_2) is normal-gamma.

The distribution has four parameters, $a > 0$, $b > 0$, $\tau > 0$ and μ, which is unrestricted.

Note that this is an example of a hierarchical construction of a distribution. One only has to use a univariate distribution at each stage of the hierarchy.

The joint distribution of (θ_1, θ_2) is

$$\begin{aligned} p(\theta_1, \theta_2) &\propto \sqrt{\theta_2}\, \mathrm{exp}\left(-\frac{\tau\theta_2}{2}(\theta_1 - \mu)^2\right) \\ &\qquad \times \theta_2^{a-1} e^{-b\theta_2}. \end{aligned}$$

The likelihood function is

$$p(\boldsymbol{y}|\theta_1, \theta_2) \propto \theta_2^{n/2} \exp\left[-\frac{\theta_2}{2}\sum_{i=1}^{n}(y_i - \theta_1)^2\right].$$

If we use a normal-gamma prior, then the posterior is

$$\begin{aligned} p(\theta_1, \theta_2|\boldsymbol{y}) &\propto& \sqrt{\theta_2}\exp\left[-\frac{(\tau+n)\theta_2}{2}(\theta_1 - \mu')^2\right] \\ && \times\ \theta_2^{a+n/2-1} e^{-b'\theta_2}, \end{aligned}$$

where

$$\mu' = \frac{\tau\mu + n\bar{y}}{\tau + n}$$

and

$$b' = b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\tau n(\bar{y} - \mu)^2}{2(\tau + n)}.$$

We have the following summary of the posterior distribution:

- The marginal posterior distribution of θ_2 is gamma with parameters $a + n/2$ and b'.

- The conditional distribution of θ_1 given θ_2 is normal, but the marginal distribution of θ_1 is not normal.

- The marginal *prior* distribution of θ_1 is
$$p_1(\theta_1) \propto \left[1 + \frac{1}{2a} \cdot \frac{a\tau(\theta_1 - \mu)^2}{b}\right]^{-(2a+1)/2},$$
which is a t distribution with $2a$ degrees of freedom and precision $a\tau/b$.

What is the marginal *posterior* of θ_1?

Example 9 Box and Tiao, p. 83

Breaking strength (in grams) was measured for 20 samples of yarn taken randomly from spinning machines.

The 20 observations are assumed to be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2.

The following noninformative (and improper) prior is used:

$$p(\mu, \sigma) \propto \sigma^{-1} I_{(0,\infty)}(\sigma) I_{(-\infty,\infty)}(\mu).$$

The posterior depends on the data only through sufficient statistics, which in this case are

$$\bar{y} = 50 \quad \text{and} \quad \sum_{i=1}^{20} (y_i - \bar{y})^2 = 348.$$

The posterior distribution is

$$\begin{aligned} p(\mu, \sigma|\boldsymbol{y}) &\propto \sigma^{-21} \exp\left(-\frac{348}{2\sigma^2}\right) \\ &\times \exp\left[-\frac{20}{2\sigma^2}(\mu - 50)^2\right]. \end{aligned}$$

The mode of the posterior is

$$(\hat{\mu}, \hat{\sigma}) = \left(50, \sqrt{\frac{348}{21}}\right) = (50, 4.07).$$

A good way of summarizing the posterior is to show *contours* of the distribution.

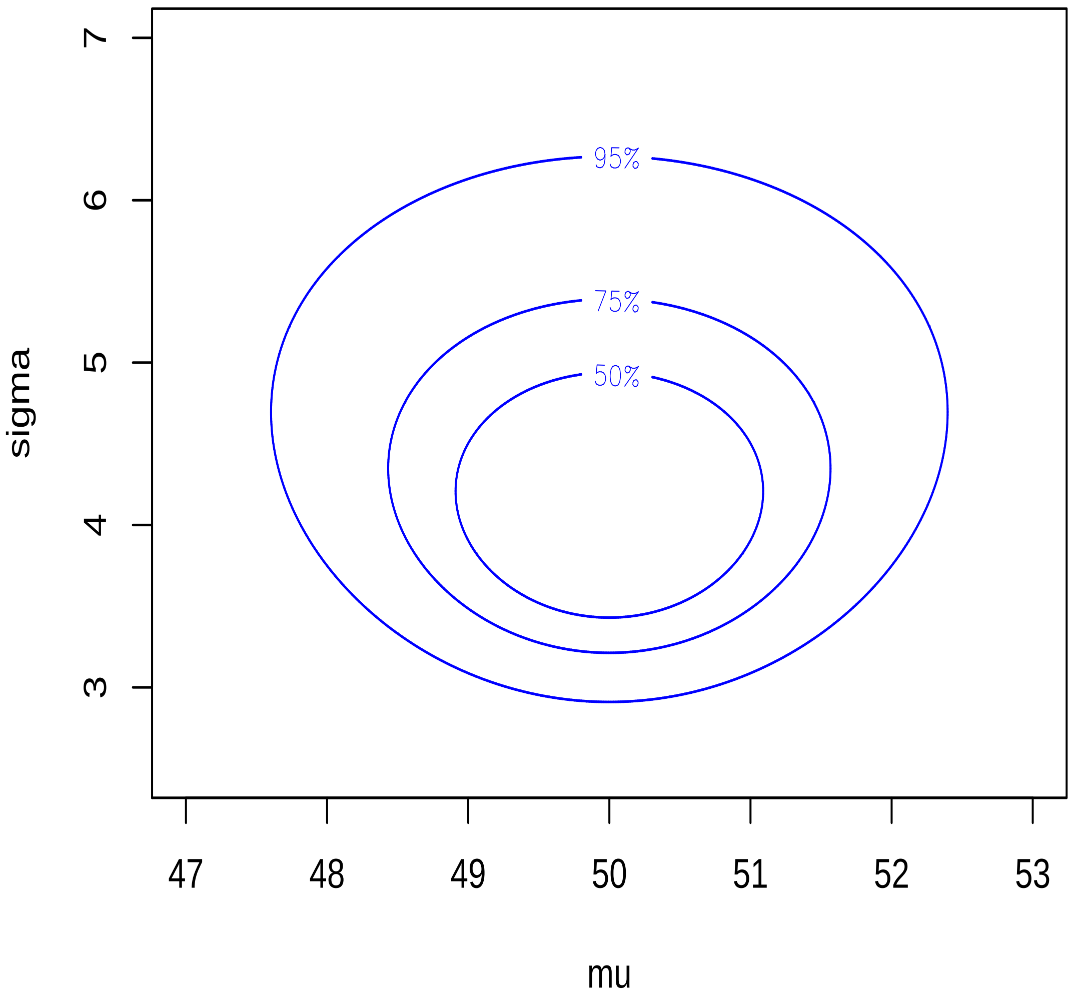
Jeffreys (*Theory of Probability*, 3rd edition) shows that, to a good approximation, the probability enclosed by the contour

$$\log(p(\mu, \sigma|\boldsymbol{y})) = \log(p(\hat{\mu}, \hat{\sigma}|\boldsymbol{y})) - \frac{1}{2}\chi^2(2, \alpha)$$

has probability $1-\alpha$, where $\chi^2(2, \alpha)$ is the $1-\alpha$ quantile of the χ^2 distribution with 2 degrees of freedom.

Box and Tiao suggest plotting 50, 75 and 95% contours.

Fifty, 75 and 95% contours for the breaking strength data



Also of interest are component distributions of the posterior. We may write

$$p(\mu, \sigma | \boldsymbol{y}) = p(\mu | \sigma, \boldsymbol{y}) p_2(\sigma | \boldsymbol{y}).$$

For the normal model with the noninformative prior on p. 99N, we have the following results:

(1) $p(\mu | \sigma, \boldsymbol{y})$ is $N(\bar{y}, \sigma^2/n)$.

(2) The marginal posterior of σ is

$$p_2(\sigma | \boldsymbol{y}) \propto \sigma^{-n} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right].$$

(3) The marginal posterior of $(\mu - \bar{y})/(s/\sqrt{n})$ is t_{n-1}.

There's an interesting symmetry between result (3) on the last page and a classical frequentist result.

Frequentists know that the sampling distribution of

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

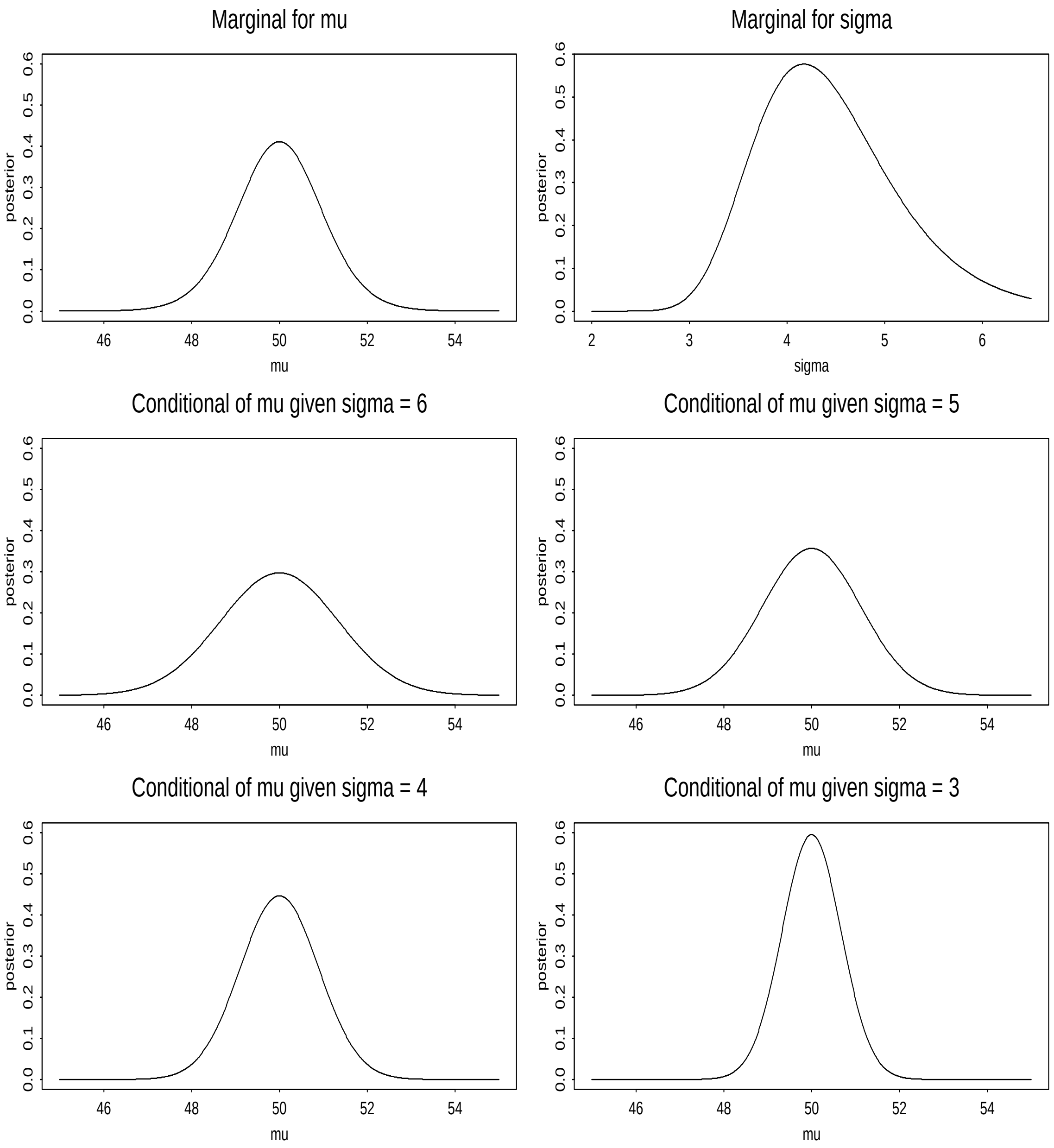
given that μ is the true population mean is t with $n-1$ degrees of freedom.

Result (3) says that the posterior distribution of

$$\frac{\mu - \bar{y}}{s/\sqrt{n}}$$

given the data is t with $n-1$ degrees of freedom.

Component distributions for breaking strength example



Marginal for mu

Marginal for sigma

Conditional of mu given sigma = 6

Conditional of mu given sigma = 5

Conditional of mu given sigma = 4

Conditional of mu given sigma = 3

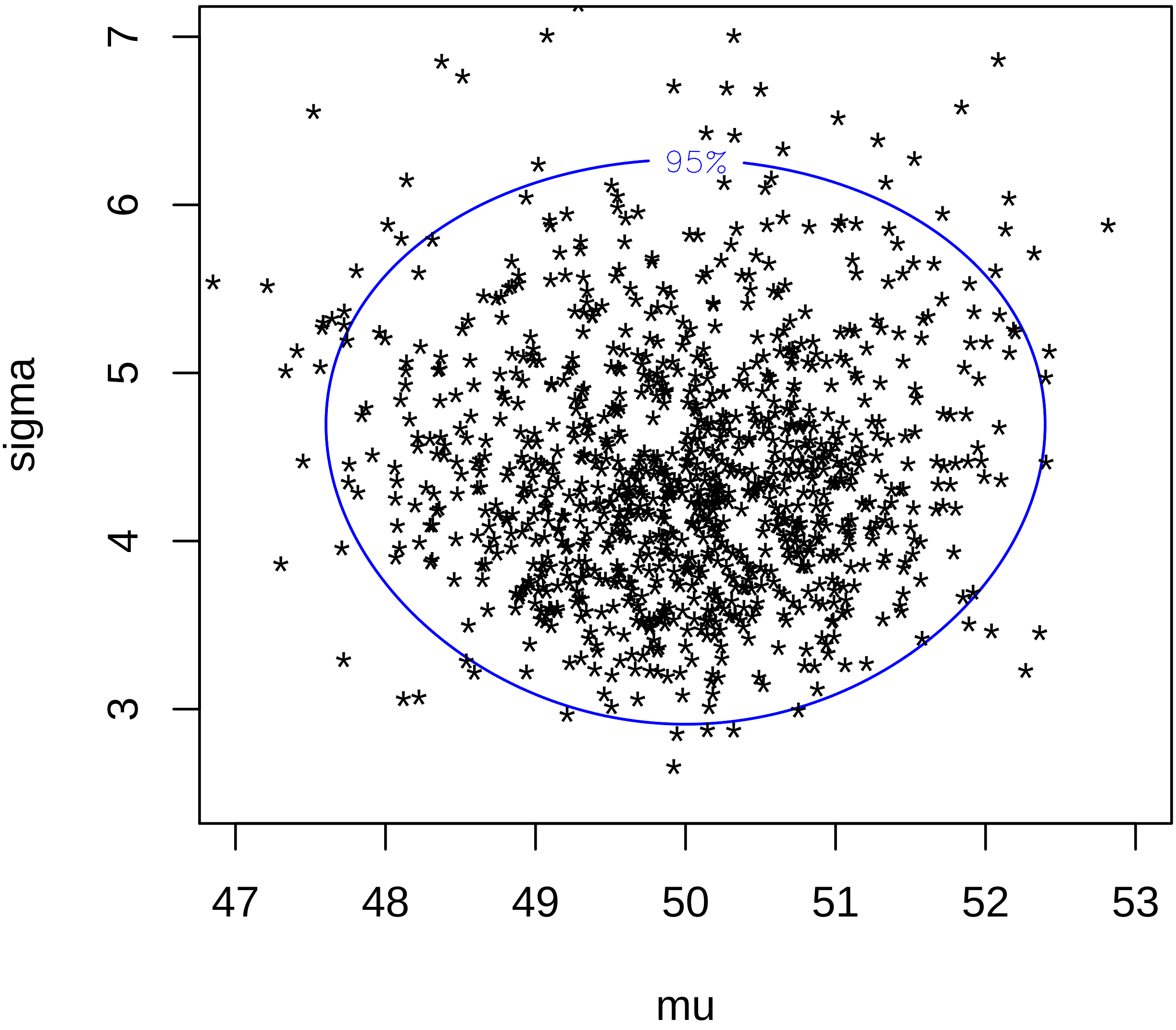
Conditional distributions can be useful in generating observations from the posterior. We can use the model in Example 9 to illustrate this idea.

- The marginal posterior for σ is given in (2) on p. 102N. This implies that the posterior distribution of $1/\sigma^2$ is
$$\text{gamma}\left[\frac{(n-1)}{2}, \frac{(n-1)s^2}{2}\right].$$
This fact allows us to generate a value of σ by first generating $1/\sigma^2$ from the appropriate gamma distribution.

- Given a value of σ, we can now generate a value of μ from $N(\bar{y}, \sigma^2/n)$.

- Repeating this process many times would give us a good impression of the distribution of (μ, σ).

Result of generating 1000 observations from the posterior in Example 9



mean of generated values of μ: 49.99
mean of generated values of σ: 4.46

Mean squared error of Bayes estimators

Frequentists evaluate inference methods, including point estimators, by considering *how they work in the long run in repeated sampling from the same population.*

The *mean squared error* of an estimator is a way of judging such long run performance.

Bayesian purists are not necessarily interested in long run performance. They want a procedure that works best for the data at hand, and believe that Bayesian methodology does that.

> But if the Bayesian method works best for *every* data set encountered, shouldn't it be best on average as well?

The answer is yes, in a certain sense.

In the context of point estimation, “doing best for the data at hand” means finding a best estimate for the given posterior.

One way in which this is done is to find an estimator that minimizes *posterior risk*.

Given a point estimate of θ, call it $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y})$, its posterior risk may be defined as

$$r(\widehat{\theta}) = \int_{\Theta} (\widehat{\theta} - \theta)^2 p(\theta|\boldsymbol{y})\, d\theta.$$

It’s easy to argue that, if this risk exists finite, then it is minimized by taking $\widehat{\theta}$ equal to the *posterior mean*.

We have already said that the posterior mean is a popular Bayes point estimate. This new interpretation gives it a little added credibility.

An example of a frequentist measure of performance is $E[r(\widehat{\theta})]$, since this averages over *different data sets*.

Explicitly this expectation is

$$E[r(\widehat{\theta})] = \int_{\mathcal{Y}} r(\widehat{\theta}(\boldsymbol{y}))m(\boldsymbol{y})\,d\boldsymbol{y},$$

where $\mathcal{Y}$ is the sample space, or the set of all possible values for data vector $\boldsymbol{Y}$.

Denote the posterior mean by $\bar{\theta}(\boldsymbol{y})$. As we said before $r(\bar{\theta}(\boldsymbol{y})) \leq r(\widehat{\theta}(\boldsymbol{y}))$ for every $\boldsymbol{y}$ and any $\widehat{\theta}(\boldsymbol{y})$.

So, *it must be true that $\bar{\theta}(\boldsymbol{Y})$ is the minimizer of $E[r(\widehat{\theta}(\boldsymbol{Y}))]$ over all functions $\widehat{\theta}$ of $\boldsymbol{Y}$.*

Any frequentist that adopted the measure of performance $E[r(\widehat{\theta})]$ would concede that the posterior mean was the best estimator of θ. On this *Bayesians and frequentists could agree*.

Let's take a closer look at $E[r(\widehat{\theta})]$. We have

$$\begin{aligned} E[r(\widehat{\theta})] &= \int_{\mathcal{Y}} r(\widehat{\theta}(\boldsymbol{y}))m(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int_{\mathcal{Y}}\int_{\Theta}(\widehat{\theta}(\boldsymbol{y})-\theta)^2 p(\theta|\boldsymbol{y})\,d\theta m(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int_{\Theta}\left[\int_{\mathcal{Y}}(\widehat{\theta}(\boldsymbol{y})-\theta)^2 p(\boldsymbol{y}|\theta)\,d\boldsymbol{y}\right]p(\theta)\,d\theta \\ &= \int_{\Theta} \text{MSE}(\widehat{\theta}(\boldsymbol{Y})|\theta)p(\theta)\,d\theta. \end{aligned}$$

> Therefore, $E[r(\widehat{\theta})]$ *is a weighted average of mean squared errors, with the largest weights being given to the values of θ that are most likely a priori.*

$E[r(\widehat{\theta})]$ is called a *Bayes risk*. The mean squared error is an example of a *risk* function.

Frequentists like to use $\text{MSE}(\widehat{\theta}(\boldsymbol{Y})|\theta)$ as a means of choosing a good estimator.

However, they face the problem that, generally speaking, there is no estimator that *uniformly* minimizes risk.

In other words, unless you restrict the class of estimators in some way, there exists no estimator $\widehat{\theta}_{\text{opt}}(\boldsymbol{Y})$ such that

$$\text{MSE}(\widehat{\theta}_{\text{opt}}(\boldsymbol{Y})|\theta) \leq \text{MSE}(\widehat{\theta}(\boldsymbol{Y})|\theta)$$

for all θ and all estimators $\widehat{\theta}$.

Given two "reasonable" estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, typically what happens is that the MSE of $\widehat{\theta}_1$ is *smaller* than that of $\widehat{\theta}_2$ for some values of θ, but for other values of θ the MSE of $\widehat{\theta}_1$ is *larger* than that of $\widehat{\theta}_2$.

Using Bayes risk gives a nice resolution to this problem by choosing the estimator that has the smallest *average* risk.

Of course, the average risk is dependent on the prior used, and so there is still no "magical" solution acceptable to all, unless everyone can agree on the prior.

How do the MSEs of classical frequentist estimators compare with those of Bayes estimators in the normal model?

Consider first estimators of the population variance σ^2. In the setting of Example 9, the posterior mode of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n+1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Two "classical" estimators of σ^2 are

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Of course, S^2 is unbiased and $\hat{\sigma}^2$ is the MLE. Let's find the mean squared error of each of the three.

We know from math stat that when we really have a random sample from the normal distribution, $(n-1)S^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom.

This means that

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1,$$

or $E(S^2) = \sigma^2$, and hence S^2 is unbiased. Also,

$$\text{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1),$$

or $\text{Var}(S^2) = 2\sigma^4/(n-1)$.

Since MSE is the sum of variance and squared bias, it follows that

$$\text{MSE}(S^2) = \frac{2\sigma^4}{(n-1)}.$$

Now consider any estimator of the form cS^2, where c is a constant. We have

$$\begin{aligned}
\text{MSE}(cS^2) &= \frac{2c^2\sigma^4}{(n-1)} + (c\sigma^2 - \sigma^2)^2 \\
&= \frac{2c^2\sigma^4}{(n-1)} + \sigma^4(1-c)^2 \\
&= \sigma^4\left[\frac{2c^2}{(n-1)} + (1-c)^2\right].
\end{aligned}$$

Therefore,

$$\text{MSE}(\hat{\sigma}^2) = \frac{\sigma^4(2n-1)}{n^2}$$

and

$$\text{MSE}(\tilde{\sigma}^2) = \frac{\sigma^4(2n+2)}{(n+1)^2}.$$

For all $n \geq 2$ we have

$$\text{MSE}(\tilde{\sigma}^2) < \text{MSE}(\hat{\sigma}^2) < \text{MSE}(S^2).$$

A natural question is “Why stop at $n + 1$? How about using an estimator of the form $(n - 1)S^2/(n + a)$ for $a = 2$ or 3 or whatever?”

> When you do so you find that the estimator with miminum mean squared error is the one with $a = 1$!

We’ll now consider a Bayesian estimator of the population mean that results from using the normal-gamma prior.

From pp. 97-98N, the posterior mean (and mode) of θ_1 is

$$\widehat{\theta}_1 = \frac{\tau\mu + n\bar{Y}}{\tau + n} = (1 - w)\mu + w\bar{Y},$$

where $w = n/(\tau + n)$.

It follows that

$$\begin{aligned} \text{MSE}(\widehat{\theta}_1) &= \text{Var}(\widehat{\theta}_1) + \left[E(\widehat{\theta}_1) - \theta_1\right]^2 \\ &= \frac{w^2\sigma^2}{n} + (1-w)^2(\mu - \theta_1)^2. \end{aligned}$$

The usual frequentist estimator of θ_1 is $\bar{Y}$, which is the MLE and also a *uniformly minimum variance unbiased estimator (UMVUE)* of θ_1.

The MSE of $\bar{Y}$ is equal to its variance, which is σ^2/n.

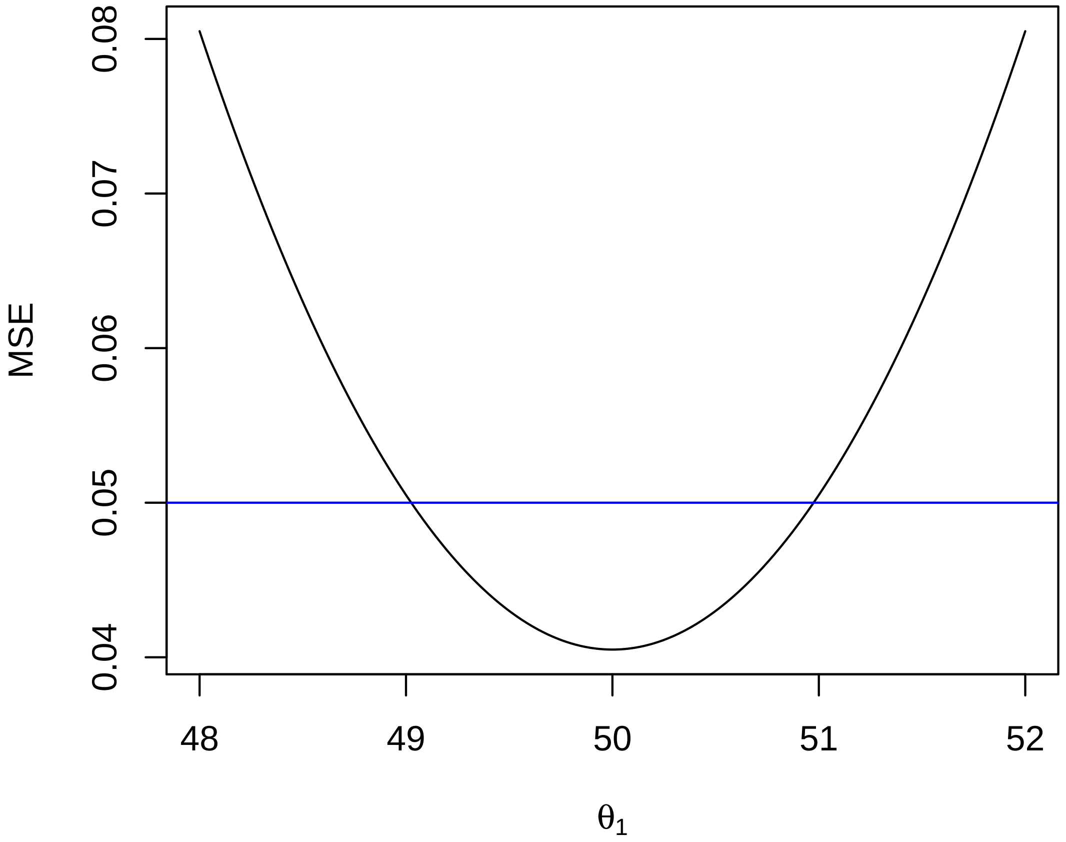
When is the MSE of $\widehat{\theta}_1$ smaller than that of $\bar{Y}$?

This happens when

$$(\mu - \theta_1)^2 < \frac{\sigma^2}{n} \cdot \frac{1+w}{1-w} = \sigma^2 \left(\frac{1}{n} + \frac{2}{\tau}\right),$$

i.e., *when the prior mean is a sufficiently good guess of θ_1.*

MSE of Posterior Mean and $\bar{Y}$ as a Function of True Mean θ_1



$$n = 20,\ w = 0.9,\ \sigma^2 = 1,\ \mu = 50$$

If the true mean is sufficiently close to the prior guess of 50, then the MSE of the Bayes estimator is smaller than that of $\bar{Y}$.

Large sample properties of Bayesian inference

A fundamental result in Bayesian analysis is that, under general conditions, the posterior distribution is approximately multivariate normal for large n.

Let $\widehat{\boldsymbol{\theta}}$ be the mode of the posterior distribution, let $p^*(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, and define the matrix $H(\boldsymbol{\theta})$ as follows:

$$H(\boldsymbol{\theta})_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log p^*(\boldsymbol{\theta}|\boldsymbol{y}).$$

The matrix $-H(\boldsymbol{\theta})$ is called the *Hessian*.

Under certain regularity conditions (to be discussed later) the posterior distribution is approximately $N(\widehat{\boldsymbol{\theta}}, H(\widehat{\boldsymbol{\theta}})^{-1})$ when n is sufficiently large.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-variate random vector.

The notation $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that $\boldsymbol{X}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

In this case $\boldsymbol{X}$ has density

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \\ &\quad \times \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \end{aligned}$$

where $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, meaning that

$$\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j).$$

The essential part of proving the asymptotic normality result is a Taylor series expansion. We expand $\log p^*(\boldsymbol{\theta}|\boldsymbol{y})$ in a Taylor series about $\widehat{\boldsymbol{\theta}}$.

$$\begin{aligned}\log p^*(\boldsymbol{\theta}|\boldsymbol{y}) &= \log p^*(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}) \\ &\quad +(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^T \frac{\partial}{\partial \boldsymbol{\theta}} \log p^*(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\ &\quad -\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^T H(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) \\ &\quad +R_n,\end{aligned}$$

where R_n is negligible relative to the other terms as $n \to \infty$.

The ith element of $\frac{\partial}{\partial \boldsymbol{\theta}} \log p^*(\boldsymbol{\theta}|\boldsymbol{y})$ is

$$\frac{1}{p^*(\boldsymbol{\theta}|\boldsymbol{y})} \cdot \frac{\partial}{\partial \theta_i} p^*(\boldsymbol{\theta}|\boldsymbol{y}).$$

The last quantity evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ is 0 since $\widehat{\boldsymbol{\theta}}$ maximizes p^*. It follows that

$$\begin{aligned} p^*(\boldsymbol{\theta}|\boldsymbol{y}) &= p^*(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}) \\ &\quad \times \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T H(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right] \\ &\quad \times \exp(R_n). \end{aligned}$$

So, for large n, the posterior is approximately proportional to a multivariate normal density with mean $\widehat{\boldsymbol{\theta}}$ and covariance $H(\widehat{\boldsymbol{\theta}})^{-1}$.

One regularity condition required for the previous result is that the second partial derivatives of $p^*(\boldsymbol{\theta}|\boldsymbol{y})$ exist and be continuous throughout a neighborhood of the true parameter vector $\boldsymbol{\theta}_0$.

For a complete set of sufficient conditions, see LeCam and Yang (1990).

Example 10 Asymptotic normality in binomial experiment

Suppose $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(θ), and suppose the prior is beta(a, b). If $y = \sum_{i=1}^n y_i$, then

$$p^*(\theta|\boldsymbol{y}) = C\theta^{y+a-1}(1-\theta)^{n-y+b-1}.$$

$$\frac{\partial}{\partial\theta}\log p^*(\theta|\boldsymbol{y}) = \frac{(y+a-1)}{\theta} - \frac{n-y+b-1}{1-\theta}$$

$$-\frac{\partial^2}{\partial\theta^2}\log p^*(\theta|\boldsymbol{y}) = \frac{(y+a-1)}{\theta^2} + \frac{(n-y+b-1)}{(1-\theta)^2}$$

The mode of the posterior is

$$\hat{\theta} = \frac{y+a-1}{n+a+b-2}.$$

Verify that

$$H(\widehat{\theta}) = \frac{(n+a+b-2)}{\widehat{\theta}(1-\widehat{\theta})}.$$

So, the posterior is approximately normal with mean $\widehat{\theta}$ and variance $\widehat{\theta}(1-\widehat{\theta})/(n+a+b-2)$.

Recall that the MLE of θ is $\widehat{\theta}_n = y/n$. We have seen previously that

$$\widehat{\theta} = \widehat{\theta}_n + O(1/n),$$

where $O(1/n)$ denotes a quantity that converges to 0 at the rate n^{-1}.

The asymptotic posterior variance is

$$\frac{\widehat{\theta}(1-\widehat{\theta})}{(n+a+b-2)} = \frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n} + O_p(n^{-2}),$$

where $O_p(n^{-2})$ is a random variable that converges to 0 at the rate n^{-2}.

Note that $I(\theta) = [\theta(1-\theta)/n]^{-1}$, and hence we may conclude that the posterior is approximately $N(\widehat{\theta}_n, I(\widehat{\theta}_n)^{-1})$.

The last result mentioned in Example 10 is not peculiar to the binomial model.

Under general conditions, the posterior is (for large n) approximately multivariate normal with mean vector equal to the MLE $\widehat{\boldsymbol{\theta}}_n$ and covariance matrix $I(\widehat{\boldsymbol{\theta}}_n)^{-1}$.

> *An interesting aspect of this result is that the approximation to the posterior depends in no way on the prior.*

The validity of the result rests on the fact that, for large n, the likelihood dominates the prior, in the sense that the likelihood is sharply peaked relative to the prior.

Another interesting aspect of the result on p. 124N is how it parallels the asymptotic normality of MLEs.

Frequentists make heavy use of the following result:

> *Under general conditions, the sampling distribution of the MLE $\widehat{\boldsymbol{\theta}}_n$ is approximately $N(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$ for large n.*

Therefore, as we have seen in other cases,

> *the frequency distribution of a statistic for a given parameter value has the same form as the distribution of a parameter conditional on the data.*

Example 11 Large sample posterior for normal model

We'll derive an approximation to the posterior of (μ, σ), the one based on the maximum likelihood estimates of μ and σ^2, which are $\bar{Y}$ and $\hat{\sigma}^2$.

Our main task is to find $I(\hat{\boldsymbol{\theta}})$. We have

$$\frac{\partial}{\partial \mu} \log p(\boldsymbol{y}|\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu),$$

$$\frac{\partial^2}{\partial \mu^2} \log p(\boldsymbol{y}|\mu, \sigma^2) = -\frac{n}{\sigma^2},$$

$$\frac{\partial}{\partial \sigma} \log p(\boldsymbol{y}|\mu, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2,$$

$$\frac{\partial^2}{\partial \sigma^2} \log p(\boldsymbol{y}|\mu, \sigma^2) = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2$$

and

$$\frac{\partial^2}{\partial\mu\partial\sigma}\log p(\boldsymbol{y}|\mu,\sigma^2) = -\frac{2}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu).$$

Evaluating the second derivatives at the MLEs gives the information matrix

$$I(\bar{y},\hat{\sigma}) = n\left[\begin{array}{cc} 1/\hat{\sigma}^2 & 0 \\ 0 & 2/\hat{\sigma}^2 \end{array}\right],$$

whose inverse is

$$I(\bar{y},\hat{\sigma})^{-1} = \frac{1}{n}\left[\begin{array}{cc} \hat{\sigma}^2 & 0 \\ 0 & \hat{\sigma}^2/2 \end{array}\right].$$

Since the off-diagonal term is 0, μ and σ are uncorrelated given the data, and the approximate posterior factors into the product of two normals.

So, given the data we have

$$\mu \mathrel{\dot{\sim}} N\left(\bar{y}, \frac{\widehat{\sigma}^2}{n}\right)$$

and, independently,

$$\sigma \mathrel{\dot{\sim}} N\left(\widehat{\sigma}, \frac{\widehat{\sigma}^2}{2n}\right).$$

Some situations where the asymptotic normality result *does not* hold.

- *Nonidentifiable models.* A model is *nonidentifiable* when the distribution of $\boldsymbol{Y}$ is the same for multiple values of $\boldsymbol{\theta}$. In this case, the data cannot distinguish between such $\boldsymbol{\theta}$ values.

- *Improper posteriors.* The posterior distribution must be proper in order for asymptotic normality to hold. An improper posterior will occur only when the *prior* is improper.

- *Priors that are 0 in a neighborhood of $\boldsymbol{\theta}_0$.* The prior probability that θ is in $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ should be larger than 0 for each $\epsilon > 0$.

- *True parameter is on boundary of Θ.* Suppose for example that $\Theta = [0, \infty)$ and $\theta_0 = 0$. Then the posterior will not be asymptotically normal because it will have no mass less than θ_0.

Inference for the mean when the data aren't normal

When it is assumed that the data are a random sample from $N(\theta_1, 1/\theta_2)$ and we use the normal-gamma prior, a $(1-\alpha)$100% HPD region for μ has the form

$$\mu' \pm t_{2a',\alpha/2}\sqrt{\frac{b'}{a'(\tau+n)}},$$

where $a' = a + n/2$ and $t_{2a'}$ is the $1-\alpha/2$ quantile of the t-distribution with mean 0, precision 1 and degrees of freedom $2a'$.

When n is large, this interval is approximately

$$\bar{y} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \tag{6}$$

We shall investigate if interval (6) is still valid when the data *are not* normally distributed.

Suppose we think the normal model is valid, but it really isn't. So, we are treating $\bar{Y}$ and S^2 as sufficient statistics when they may not be.

Let $\mu = E(Y_i)$, $\sigma^2 = \mathrm{Var}(Y_i)$, and

$$\mu_i = E\left[(Y_i - \mu)^i\right], \quad i = 3, 4.$$

It follows that, for large n, $(\bar{Y}, S^2)$, given μ, σ^2, μ_3 and μ_4, has a bivariate normal distribution with mean vector (μ, σ^2) and covariance matrix

$$\Sigma_n = \frac{1}{n}\begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}.$$

Here we will deal with the case where we *assume* that $\mu_3 = 0$.

It follows that, for large n and given μ, σ^2 and μ_4, $\bar{Y}$ and S^2 are independent with $\bar{Y} \sim N(\mu, \sigma^2/n)$ and $S^2 \sim N(\sigma^2, (\mu_4 - \sigma^4)/n)$.

We are in a situation that we have not yet encountered:

> The likelihood contains no information about one of the parameters in the model, namely μ_4.

This isn't necessarily a big deal for the Bayesian. We can proceed by *putting a prior on μ_4 and then integrating out μ_4 in the posterior.*

Suppose we use the noninformative and improper flat prior for μ, and assume that μ is independent of μ_4 and σ^2.

Now use a prior for (μ_4, σ^2) as follows:

$$\begin{aligned} p(\mu_4, \sigma^2) &= p(\mu_4|\sigma^2)p(\sigma^2) \\ &= \pi(\mu_4 - \sigma^4)p(\sigma^2), \end{aligned}$$

where π is a density such that $\pi(z) = 0$ for $z < 0$. (**Note:** $\mu_4 \geq \sigma^4$ by Jensen's inequality.)

For example, we could use the inverse gamma for π

$$\pi(z) \propto z^{-a-1}e^{-b/z}I_{(0,\infty)}(z).$$

The posterior is

$$\begin{aligned} p(\mu, \sigma^2, \mu_4|\bar{y}, s^2) &\propto \frac{1}{\sigma}\phi\left(\frac{\sqrt{n}(\mu - \bar{y})}{\sigma}\right) \\ &\quad \times \frac{1}{\sqrt{\mu_4 - \sigma^4}}\phi\left(\frac{\sqrt{n}(\sigma^2 - s^2)}{\sqrt{\mu_4 - \sigma^4}}\right) \\ &\quad \times \pi(\mu_4 - \sigma^4)p(\sigma^2). \end{aligned}$$

If we make the change of variable $z = \mu_4 - \sigma^4$ and integrate out z, we are left with a posterior of the form

$$\begin{aligned} p(\mu, \sigma^2 | \bar{y}, s^2) \;&\propto\; \frac{1}{\sigma}\phi\left(\frac{\sqrt{n}(\mu - \bar{y})}{\sigma}\right) \\ &\quad \times f(\sqrt{n}(\sigma^2 - s^2))p(\sigma^2). \end{aligned}$$

As an exercise, see what f is in the case where π is inverse gamma.

The conditional posterior of μ given σ^2 is $N(\bar{y}, \sigma^2/n)$. This means that a conditional (on σ) HPD region for μ has the form

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{7}$$

This fact means we can construct a credible region for (μ, σ) that has the shape of a trapezoid. Why?

Well,

$$P\left(\mu \in \bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \bigcap \sigma_1 \le \sigma \le \sigma_2 \middle| \text{data}\right) =$$

$$\int_{\sigma_1}^{\sigma_2} P\left(\mu \in \bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \middle| \bar{y}, s^2, \sigma\right) p(\sigma|s^2)\, d\sigma =$$

$$(1-\alpha)P\left(\sigma_1 \le \sigma \le \sigma_2 \middle| \text{data}\right).$$

Furthermore, since the marginal posterior of σ^2 is highly concentrated at s^2, it follows from (7) that an approximate HPD region for μ is

$$\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

For large n, why is $f(\sqrt{n}(\sigma^2 - s^2))p(\sigma^2)$ highly concentrated near $\sigma^2 = s^2$?