Bayesian Statistics

Chapter 5. Normal Model

Hojin Yang

Department of Statistics
Pusan National University

Introduction

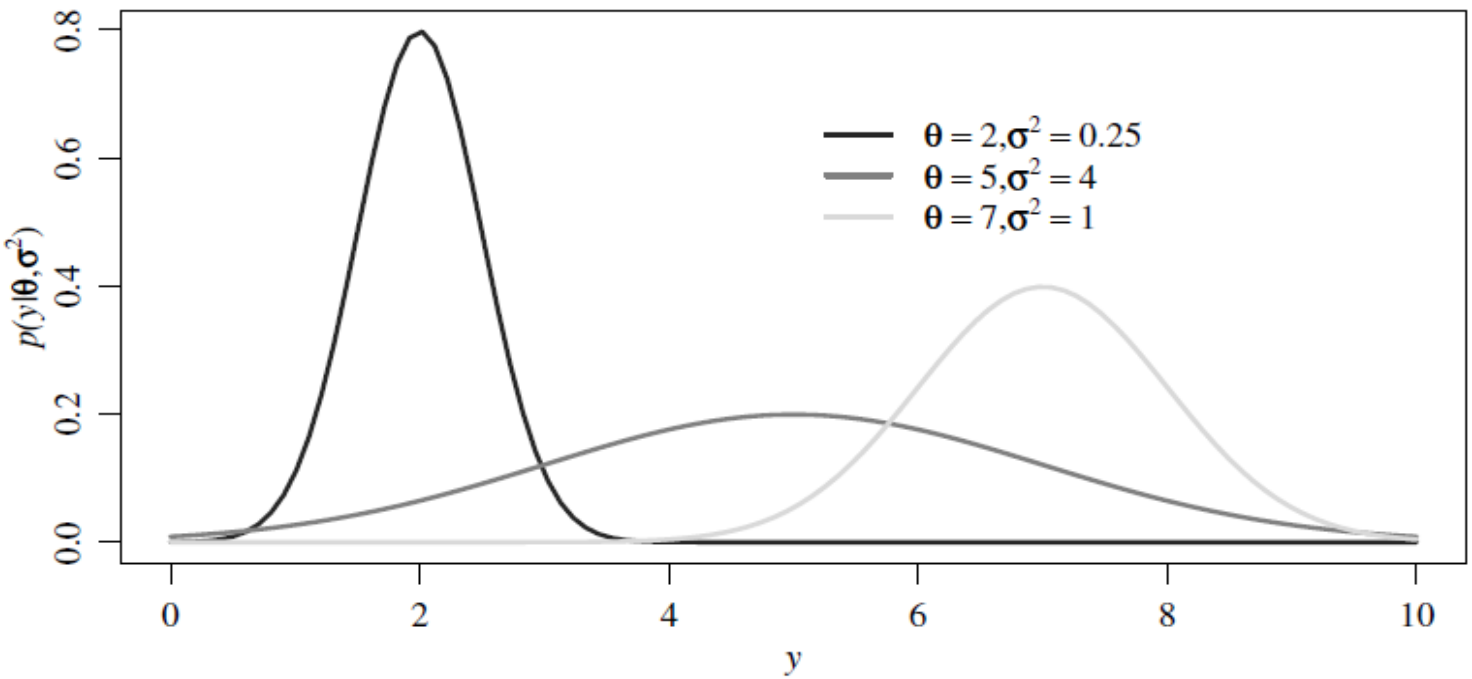
- Perhaps the most useful probability model for data analysis is the normal distribution
- There are several reasons for this, one being the central limit theorem, and another being that the normal model is a simple model with separate parameters for the population mean and variance
- In this chapter we discuss some of the properties of the normal distribution, and show how to make posterior inference on the population mean and variance parameters

5.1. Normal Model

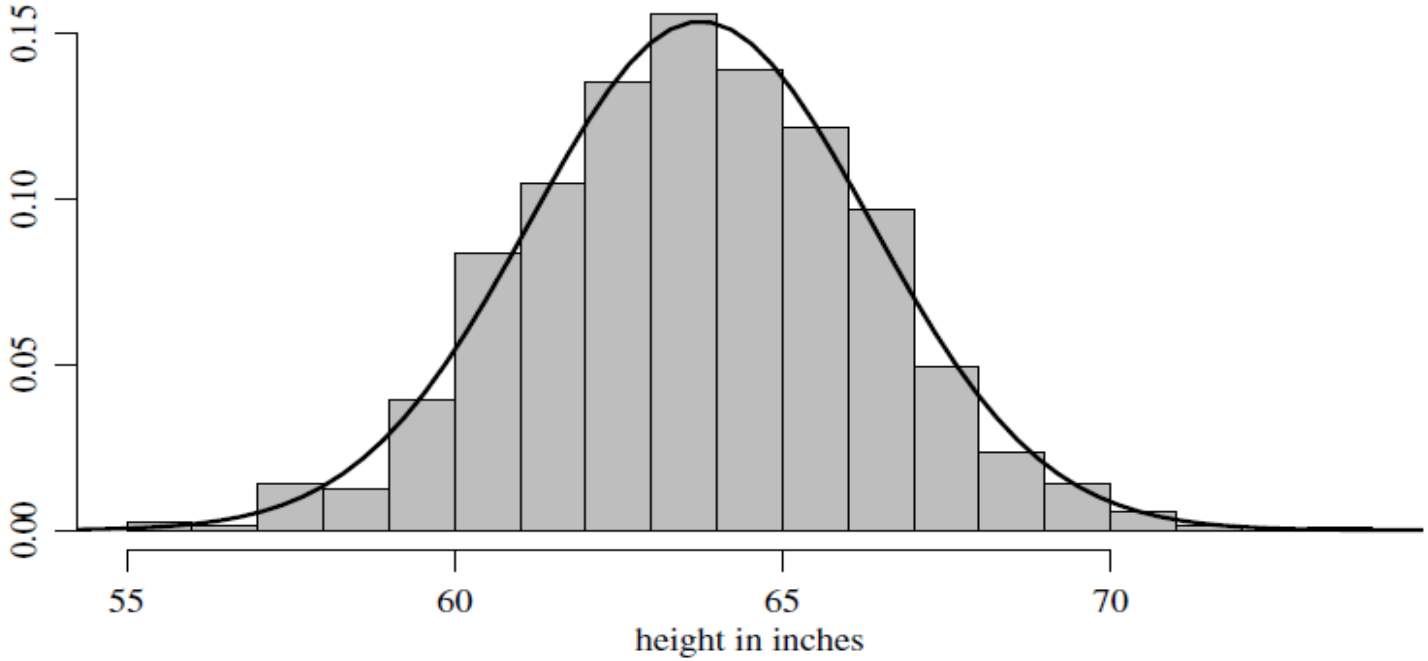
- A random variable Y is said to be normally distributed with mean θ and σ^2 if Y has the density

$$p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\theta}{\sigma})^2}, \quad -\infty < y < \infty.$$

- Some normal densities



- Remember about this distribution include the followings
- Dist is symmetric about θ and the mode, median and mean are all equal to θ
- About 95% of the population lies within two standard deviations of the mean (more precisely, 1.96 standard deviations)
- If $X \sim N(\mu, \tau^2)$, $Y \sim N(\theta, \sigma^2)$ and X and Y are independent, then $aX + bY \sim N(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$
- The dnorm, rnorm, pnorm, and qnorm commands in R take the standard deviation σ as their argument, not the variance σ^2, which can drastically change your results



- The importance of the normal distribution stems primarily from the central limit theorem
- The sum (or mean) of a set of random variables is approximately normally distributed
- This means that the normal sampling model will be appropriate for data that result from the additive effects of a large number of factors

5.2. Inference for Mean, Conditional on Variance

- Let $\{Y_1, \ldots, Y_n | \theta, \sigma\} \sim N(\theta, \sigma^2)$. Joint sampling dist is

$$\begin{aligned} p(y_1, \ldots, y_n | \theta, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \theta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \theta}{\sigma}\right)^2} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2} \sum \left(\frac{y_i - \theta}{\sigma} \right)^2 \right\} \end{aligned}$$

- Expanding the quadratic term in the exponent, we see that $p(y_1, \ldots, y_n | \theta, \sigma^2)$ depends on $y_1, \ldots, y_n$ through

$$\sum_{i=1}^{n} \left(\frac{y_i - \theta}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum y_i^2 - 2\frac{\theta}{\sigma^2} \sum y_i + n\frac{\theta^2}{\sigma^2}.$$

- We know $\{\sum y_i^2, \sum y_i\}$ make up a two-dimensional sufficient statistic
- Knowing the values of these quantities is equivalent to knowing the values of $\bar{y}$ and s^2, and $\{s^2, \bar{y}\}$ are also a sufficient statistic
- Inference for this two-parameter model can be broken down into two one parameter problems
- We will begin with the problem of making inference for θ when σ^2 is known, while using a conjugate prior distribution for θ

- For any (conditional) prior distribution $p(\theta|\sigma^2)$, the posterior distribution will satisfy

$$p(\theta|y_1,\ldots,y_n,\sigma^2) \propto p(\theta|\sigma^2) \times e^{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2}$$
$$\propto p(\theta|\sigma^2) \times e^{c_1(\theta-c_2)^2}.$$

- Recall that a class of prior distributions is conjugate for a sampling model if the resulting posterior distribution is in the same class
- If $p(\theta|\sigma^2)$ is to be conjugate, it must include quadratic terms like $e^{c_1(\theta-c_2)^2}$
- The simplest such class of probability densities on R is the normal family of densities

- Let's evaluate this claim: If $\theta \sim N(\mu_0, \tau_0^2)$, then

$$\begin{aligned} p(\theta|y_1,\ldots,y_n,\sigma^2) &= p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2)/p(y_1,\ldots,y_n|\sigma^2) \\ &\propto p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2) \\ &\propto \exp\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\}\exp\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\} \end{aligned}$$

- Adding the terms in the exponents and ignoring the $-1/2$ for the moment, we have

$$\frac{1}{\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}(\sum y_i^2 - 2\theta\sum y_i + n\theta^2) = a\theta^2 - 2b\theta + c, \text{ where}$$

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \quad b = \frac{\mu_0}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}, \quad \text{and } c = c(\mu_0, \tau_0^2, \sigma^2, y_1, \ldots, y_n).$$

- Now let's see if $p(\theta|\sigma^2, y_1, \ldots, y_n)$ takes the form of a normal density

$$\begin{aligned} p(\theta|\sigma^2, y_1, \ldots, y_n) &\propto \exp\{-\frac{1}{2}(a\theta^2 - 2b\theta)\} \\ &= \exp\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\} \\ &\propto \exp\{-\frac{1}{2}a(\theta - b/a)^2\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{\theta - b/a}{1/\sqrt{a}}\right)^2\right\}. \end{aligned}$$

- This function has exactly the same shape as a normal density curve, with $1/\sqrt{a}$ playing the role of the standard deviation and b/a playing the role of the mean

- Since probability distributions are determined by their shape, this means that $p(\theta|\sigma^2, y_1, \ldots, y_n)$ is indeed a normal density
- We refer to the mean and variance of this density as μ_n and τ_n^2, where

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

Combining Information

- The (conditional) posterior parameters μ_n and τ_n^2 combine the prior parameters μ_0 and τ_0^2 with terms from the data
- Posterior variance and precision: The formula for $1/\tau_n^2$ is

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- The prior inverse variance is combined with the inverse of the data variance

- Inverse variance is often referred to as the precision
- For the normal model let,
  - $\tilde{\sigma}^2 = 1/\sigma^2$: sampling precision, i.e. how close y_i's are to θ
  - $\tilde{\tau_0}^2 = 1/\tau_0^2$: prior precision
  - $\tilde{\tau_n}^2 = 1/\tau_n^2$: posterior precision
- It is convenient to think about precision as the quantity of information on an additive scale
- For the normal model, the posterior variance formula implies

$$\tilde{\tau_n}^2 = \tilde{\tau_0}^2 + n\tilde{\sigma}^2$$

- posterior information $=$ prior information $+$ data information

- Posterior mean: Notice that

$$\mu_n = \frac{\tilde{\tau_0}^2}{\tilde{\tau_0}^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau_0}^2 + n\tilde{\sigma}^2}\bar{y}$$

- The posterior mean is a weighted average of the prior mean and the sample mean
- The weight on the sample mean is n/σ^2, the sampling precision of the sample mean
- The weight on the prior mean is $1/\tau_0^2$ the prior precision

- Suppose the prior mean were based on κ_0 prior observations from the same sample $(Y_1, \ldots, Y_n)$
- We might want to set $\tau_0^2 = \sigma^2/\kappa_0$, the variance of the mean of the prior observations
- In this case, the formula for the posterior mean reduces to

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

Prediction

- Consider predicting a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \ldots, Y_n = y_n)$
- Posterior variance and precision: The formula for $1/\tau_n^2$ is

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- The prior inverse variance is combined with the inverse of the data variance

- Consider predicting a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \ldots, Y_n = y_n)$

- Find the predictive distribution, let's use the following fact:

$$\{\tilde{Y}|\theta,\sigma^2\} \sim \text{ normal}(\theta,\sigma^2) \Leftrightarrow \tilde{Y} = \theta + \tilde{\epsilon}, \quad \{\tilde{\epsilon}|\theta,\sigma^2\} \sim \text{ normal}(0,\sigma^2)$$

- Using this result, let's first compute the posterior mean of $\tilde{Y}$

$$\begin{aligned} \mathrm{E}[\tilde{Y}|y_1,\ldots,y_n,\sigma^2] &= \mathrm{E}[\theta + \tilde{\epsilon}|y_1,\ldots,y_n,\sigma^2] \\ &= \mathrm{E}[\theta|y_1,\ldots,y_n,\sigma^2] + \mathrm{E}[\tilde{\epsilon}|y_1,\ldots,y_n,\sigma^2] \\ &= \mu_n + 0 = \mu_n \end{aligned}$$

- For the variance of $\tilde{Y}$

$$\begin{aligned}\mathrm{Var}[\tilde{Y}|y_1,\ldots,y_n,\sigma^2] &= \mathrm{Var}[\theta+\tilde{\epsilon}|y_1,\ldots,y_n,\sigma^2] \\ &= \mathrm{Var}[\theta|y_1,\ldots,y_n,\sigma^2] + \mathrm{Var}[\tilde{\epsilon}|y_1,\ldots,y_n,\sigma^2] \\ &= \tau_n^2+\sigma^2\end{aligned}$$

- Recall from the beginning of the chapter that the sum of independent normal random variables is also normal
- Therefore, since both θ and $\tilde{\epsilon}$ conditional on $y_1,\ldots,y_n$ and σ^2 are normally distributed ($\tilde{Y}=\theta+\tilde{\epsilon}$)
- Therefore, since bot The predictive distribution is therefore

$$\tilde{Y}|\sigma^2,y_1,\ldots,y_n \sim N(\mu_n,\tau_n^2+\sigma^2)$$

- It is worthwhile to have some intuition about the form of the variance of $\tilde{Y}$
- Our uncertainty about a new sample $\tilde{Y}$ is a function of our uncertainty about the precision of the population (τ^2) as well as how variable the population is (σ^2)
- As $n \to \infty$ we become more certain about θ, where τ_n^2 goes to zero
- But this certainty does not reduce the sampling variability σ^2
- Hence, our uncertainty about $\tilde{Y}$ never goes below σ^2

5.3. Joint Inference for Mean and Variance

- Bayesian inference for two or more unknown parameters is not conceptually different from the one-parameter case
- For any joint prior distribution $p(\theta, \sigma^2)$ for θ and σ^2 posterior inference proceeds using Bayes' rule

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta, \sigma^2)/p(y_1, \ldots, y_n)$$

- Joint probability can be expressed as the product of a conditional probability and a marginal probability

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$$

- In the last section, we discussed a conjugate prior for θ when σ^2 were known
- Let's consider the particular case in which $\tau_0^2 = \sigma^2/\kappa_0$

$$\begin{aligned} p(\theta, \sigma^2) &= p(\theta|\sigma^2)p(\sigma^2) \\ &= N(\theta, \mu_0, \tau_0^2) \times p(\sigma^2) \end{aligned}$$

- In this case, the parameters μ_0 and κ_0 can be interpreted as the mean and sample size from a set of prior observations

- For σ^2 we need a family of prior distributions that has support on $(0, \infty)$
- One such family of distributions is the gamma family, as we used for the Poisson sampling model
- Unfortunately, this family is not conjugate for the normal variance
- However, the gamma family does turn out to be a conjugate class of densities for $1/\sigma^2$ (precision)
- When using such a prior distribution, we say that σ^2 has an inverse-gamma distribution:
  - precision $= 1/\sigma^2 \sim G(a, b)$
  - variance $= \sigma^2 \sim IG(a, b)$

- For interpretability later on, instead of using a and b we will parameterize this prior distribution as

$$1/\sigma^2 \sim G(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

- Under this parameterization
  - $E[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$
  - $mode[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$
  - $Var[\sigma^2]$ is decreasing in ν_0

Posterior Inference

- Suppose our prior distributions and sampling model
  - $1/\sigma^2 \sim G(\nu_0/2, \nu_0\sigma_0^2/2)$
  - $\theta|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$
  - $Y_1, \ldots, Y_n|\theta, \sigma^2 \sim N(\theta, \sigma^2)$
- Just as the prior distribution, the posterior distribution can be similarly decomposed

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(\theta|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n)$$

- The conditional distribution of θ given the data and σ^2 can be obtained using the results of the previous section

$$\{\theta|y_1,\ldots,y_n,\sigma^2\} \sim \text{normal}(\mu_n, \sigma^2/\kappa_n), \text{ where}$$
$$\kappa_n = \kappa_0 + n \text{ and } \mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}.$$

- The posterior distribution of σ^2 can be obtained b

$$\begin{aligned} p(\sigma^2|y_1,\ldots,y_n) &\propto p(\sigma^2)p(y_1,\ldots,y_n|\sigma^2) \\ &= p(\sigma^2)\int p(y_1,\ldots,y_n|\theta,\sigma^2)p(\theta|\sigma^2)\,d\theta\,. \end{aligned}$$

- The result is that

$$\{1/\sigma^2|y_1,\ldots,y_n\} \sim \text{gamma}(\nu_n/2, \nu_n\sigma_n^2/2), \text{ where}$$
$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2]$$

- Recall that $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$
- We can think of $\nu_0\sigma_0^2$ and $\nu_n\sigma_n^2$ as prior and posterior sums of squares
- Multiplying both sides of the last equation by ν_n almost gives us "posterior sum of squares equals prior sum of squares plus data sum of squares"
- However, the third term in the last equation is a bit harder to understand. A large value of $(\bar{y}-\mu_0)^2$ increases the posterior probability of a large σ^2. This makes sense for our particular joint prior distribution for θ and σ^2

Example

- Studies of other populations suggest that the true mean and standard deviation of our population should not be too far from 1.9 mm and 0.1 mm. ($\mu_0 = 1.9$, $\sigma_0^2 = 0.01$)
- We choose $\kappa_0 = \nu_0 = 1$ so that our prior distributions are only weakly centered around these estimates from other populations.
- The sample mean and variance of our observed data are $\bar{y} = 1.804$ and $s^2 = 0.0169$
- From these values, we compute μ_n and σ_n^2

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n} = \frac{1.9 + 9 \times 1.804}{1 + 9} = 1.814$$

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2]$$

$$= \frac{0.010 + 0.135 + 0.008}{10} = 0.015\,.$$

- These calculations can be done with R

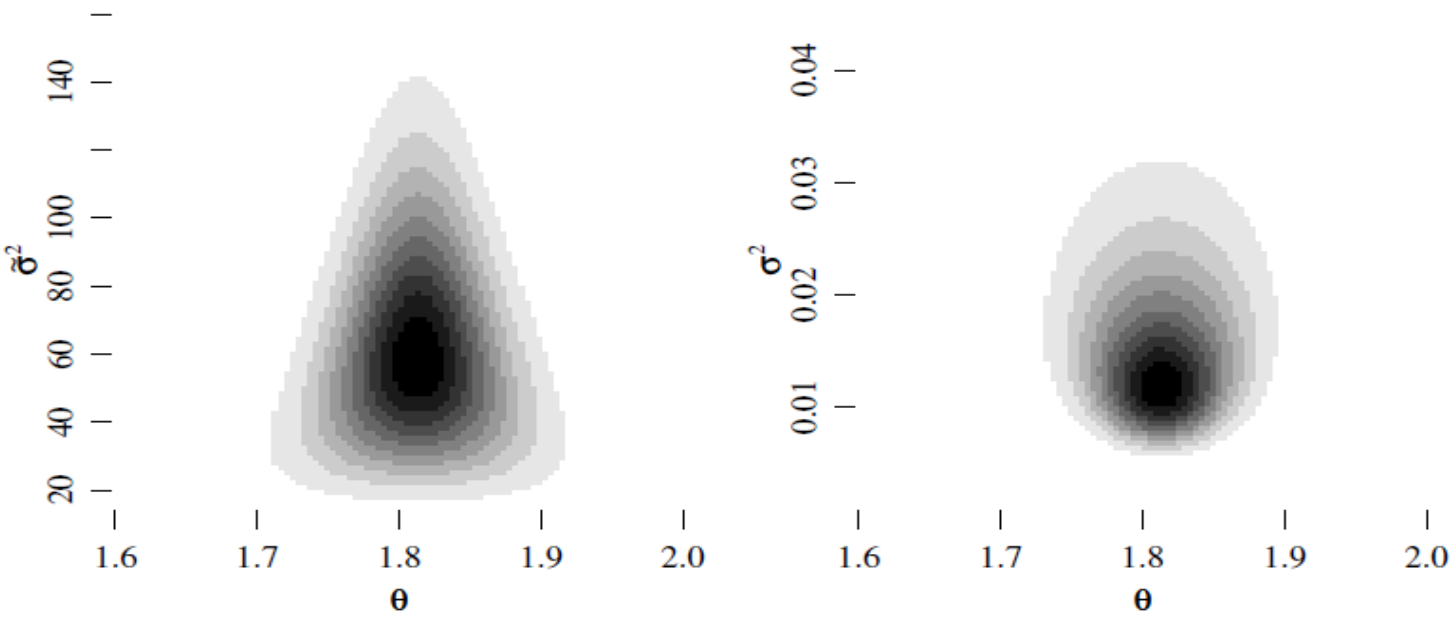
```
# prior
mu0<-1.9   ; k0<-1
s20<-.010  ; nu0<-1

# data
y<-c(1.64,1.70,1.72,1.74,1.82,1.82,1.82,1.90,2.08)
n<-length(y) ; ybar<-mean(y) ; s2<-var(y)

# posterior inference
kn<-k0+n ; nun<-nu0+n
mun<- (k0*mu0 + n*ybar)/kn
s2n<- (nu0*s20 +(n-1)*s2 +k0*n*(ybar-mu0)^2/(kn))/(nun)

> mun
[1] 1.814
> s2n
[1] 0.015324
> sqrt(s2n)
[1] 0.1237901
```

- Our joint posterior distribution is completely determined by $\mu_n = 1.814$, $\kappa_n = 0.015$, and $\nu_n = 10$

- They can be expressed as

$$\{\theta|y_1,\ldots,y_n,\sigma^2\} \sim N(1.814,\sigma^2/10)$$
$$\{1/\sigma^2|y_1,\ldots,y_n\} \sim G(10/2,10\times 0.015/2)$$

- Letting $\tilde{\sigma}^2/1/\sigma^2$, contour plots of the bivariate posterior density of $(\theta,\tilde{\sigma}^2)$ and (θ,σ^2) appear in Figure



- Notice that the contours are more peaked as a function of θ for low values of σ^2 than high values

Monte Carlo Sampling

- For many data analyses, interest primarily lies in estimating the population mean θ like $E[g(\theta)|y_1, \ldots, y_n]$
- These quantities are all determined by the marginal posterior distribution of θ given the data
- But we have the conditional distribution of θ given the data and σ^2. σ^2 given the data is inverse-gamma
- If we could generate marginal samples of θ from $p(\theta|y_1, \ldots, y_n)$ then we could use the Monte Carlo method to approximate the above quantities of interest

- Consider simulating parameter values using the following Monte Carlo procedure

$$\begin{array}{ll} \sigma^{2(1)} \sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), & \theta^{(1)} \sim \text{normal}(\mu_n, \sigma^{2(1)}/\kappa_n) \\ \vdots & \vdots \\ \sigma^{2(S)} \sim \text{inverse gamma}(\nu_n/2, \sigma_n^2\nu_n/2), & \theta^{(S)} \sim \text{normal}(\mu_n, \sigma^{2(S)}/\kappa_n) \end{array}$$

- Note that $\theta^{(s)}$ is sampled from $p(\theta|y_1, \ldots, y_n, \sigma^{2(s)})$
- The approximation can be calculated in R

```
s2.postsample <- 1/rgamma(10000, nun/2, s2n*nun/2 )
theta.postsample <- rnorm(10000, mun, sqrt(s2.postsample/kn))
```

- $\{(\theta^{(1)}, \sigma^{2(1)}), \ldots, (\theta^{(S)}, \sigma^{2(S)})\}$ using this are independent samples from the joint posterior dist $p(\theta, \sigma^2 | y_1, \ldots, y_n)$
- Additionally, $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ can be seen as independent samples from the marginal posterior dis of $p(\theta | y_1, \ldots, y_n)$
- Thereby, we use this sequence for Monte Carlo approximations
- Note that $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ are indeed each conditional samples while they are each conditional on different σ^2
- Taken together, they consist of marginal samples of θ

Improper Priors

- What if we want to "be objective" by not using any prior information (not to be Bayesian)
- The smaller κ_0 and ν_0 are, the more objective the estimates will be
- The formula for μ_n and σ_n^2

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$
$$\sigma_n^2 = \frac{1}{\nu_0 + n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2]$$

- As $\kappa_0, \nu_0 \to 0$

$$\mu_n \to \bar{y}$$
$$\sigma_n^2 \to \frac{(n-1)}{n}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

- This has led some to suggest the following posterior dist

$$\{1/\sigma^2|y_1,\ldots,y_n\} \sim G(\frac{n}{2}, \frac{n}{2}\frac{1}{n}\sum(y_i-\bar{y})^2)$$

$$\{\theta|y_1,\ldots,y_n,\sigma^2\} \sim N(\bar{y}, \frac{\sigma^2}{n})$$

- If $\tilde{p}(\theta,\sigma^2) = \sigma^2$ (not a probability density)
- Set $p(\theta,\sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\theta,\sigma^2)\tilde{p}(\theta,\sigma^2)$, we get the same conditional dist for θ and $G(\frac{n-1}{2}, \frac{1}{2}\sum(y_i-\bar{y})^2)$ for $1/\sigma^2$
- We can integrate this latter joint distribution over σ^2 to show that

$$\frac{\theta-\bar{y}}{s/\sqrt{n}}|y_1,\ldots,y_n \sim t_{(n-1)}$$

- Consider the sampling distribution of the t-statistic, conditional on θ but unconditional on the data:

$$\frac{\bar{Y} - \theta}{s/\sqrt{n}} | \theta \sim t_{(n-1)}$$

- This says that, before you sample the data, the uncertainty about the scaled deviation of the sample mean $\bar{Y}$ from the population mean θ is represented with $t_{(n-1)}$ dist.
- The former says that after we sample your data, our uncertainty is still represented with $t_{(n-1)}$ dist
- The difference is that before we sample our data, both $\bar{Y}$ and θ are unknown
- After we sample our data, then $\bar{Y} = \bar{y}$ is known and this provides us with information about θ

5.4. Bias, Variance and Mean Squared Error

- A point estimator of an unknown parameter θ is a function that converts your data into a single element of the parameter space Θ
- In case of a normal sampling model and conjugate prior distribution, the posterior mean estimator of θ is

$$\hat{\theta}_b(y_1, \ldots, y_n) = \mathrm{E}[\theta | y_1, \ldots, y_n] = \frac{n}{\kappa_0 + n}\bar{y} + \frac{\kappa_0}{\kappa_0 + n}\mu_0 = w\bar{y} + (1 - w)\mu_0$$

- For an estimator $\hat{\theta}_b$ and the true value of the population mean θ_0

$$\mathrm{E}[\hat{\theta}_e | \theta = \theta_0] = \theta_0, \quad \text{and we say that } \hat{\theta}_e \text{ is "unbiased,"}$$
$$\mathrm{E}[\hat{\theta}_b | \theta = \theta_0] = w\theta_0 + (1 - w)\mu_0, \quad \text{and if } \mu_0 \neq \theta_0 \text{ we say that } \hat{\theta}_b \text{ is "biased."}$$

- Bias refers to how close the center of mass of the sampling distribution of an estimator is to the true value
- An unbiased estimator is an estimator with zero bias, which sounds desirable. However, bias does not tell us how far away an estimate might be from the true value
- For instance, y_1 is an unbiased estimator of the population mean θ_0, but will generally be farther away from θ_0 than $\bar{y}$
- To evaluate this, we use the mean squared error (MSE)
- Letting $m = E[\hat{\theta}|\theta_0]$

$$\begin{aligned}
\mathrm{MSE}[\hat{\theta}|\theta_0] &= \mathrm{E}[(\hat{\theta} - \theta_0)^2|\theta_0] \\
&= \mathrm{E}[(\hat{\theta} - m + m - \theta_0)^2|\theta_0] \\
&= \mathrm{E}[(\hat{\theta} - m)^2|\theta_0] + 2\mathrm{E}[(\hat{\theta} - m)(m - \theta_0)|\theta_0] + \mathrm{E}[(m - \theta_0)^2|\theta_0]
\end{aligned}$$

- The first term is the variance of $\hat{\theta}$ and the third term is the square of the bias and so

$$\mathrm{MSE}[\hat{\theta}|\theta_0] = \mathrm{Var}[\hat{\theta}|\theta_0] + \mathrm{Bias}^2[\hat{\theta}|\theta_0]$$

- Getting back to our comparison of $\hat{\theta}_b$ to $\hat{\theta}_e = \bar{y}$, the bias of $\bar{y}$ is zero
- But

$$\mathrm{Var}[\hat{\theta}_e|\theta = \theta_0, \sigma^2] = \frac{\sigma^2}{n}, \text{ whereas}$$

$$\mathrm{Var}[\hat{\theta}_b|\theta = \theta_0, \sigma^2] = w^2 \times \frac{\sigma^2}{n} < \frac{\sigma^2}{n},$$

- $\hat{\theta}_b$ has lower variability

- Which one is better in terms of MSE?

$$\text{MSE}[\hat{\theta}_e|\theta_0] = E[(\hat{\theta}_e - \theta_0)^2|\theta_0] = \frac{\sigma^2}{n}$$

$$\begin{aligned}\text{MSE}[\hat{\theta}_b|\theta_0] = E[(\hat{\theta}_b - \theta_0)^2|\theta_0] &= E[\{w(\bar{y} - \theta_0) + (1-w)(\mu_0 - \theta_0)\}^2|\theta_0] \\ &= w^2 \times \frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \theta_0)^2\end{aligned}$$

- We can show that $MSE[\hat{\theta}_b|\theta_0] < MSE[\hat{\theta}_e|\theta_0]$ if

$$\begin{aligned}(\mu_0 - \theta_0)^2 &< \frac{\sigma^2}{n}\frac{1+w}{1-w} \\ &= \sigma^2\left(\frac{1}{n} + \frac{2}{\kappa_0}\right)\end{aligned}$$

- In this case, you can construct a Bayesian estimator that will have a lower average squared distance to the truth than does the sample mean