COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

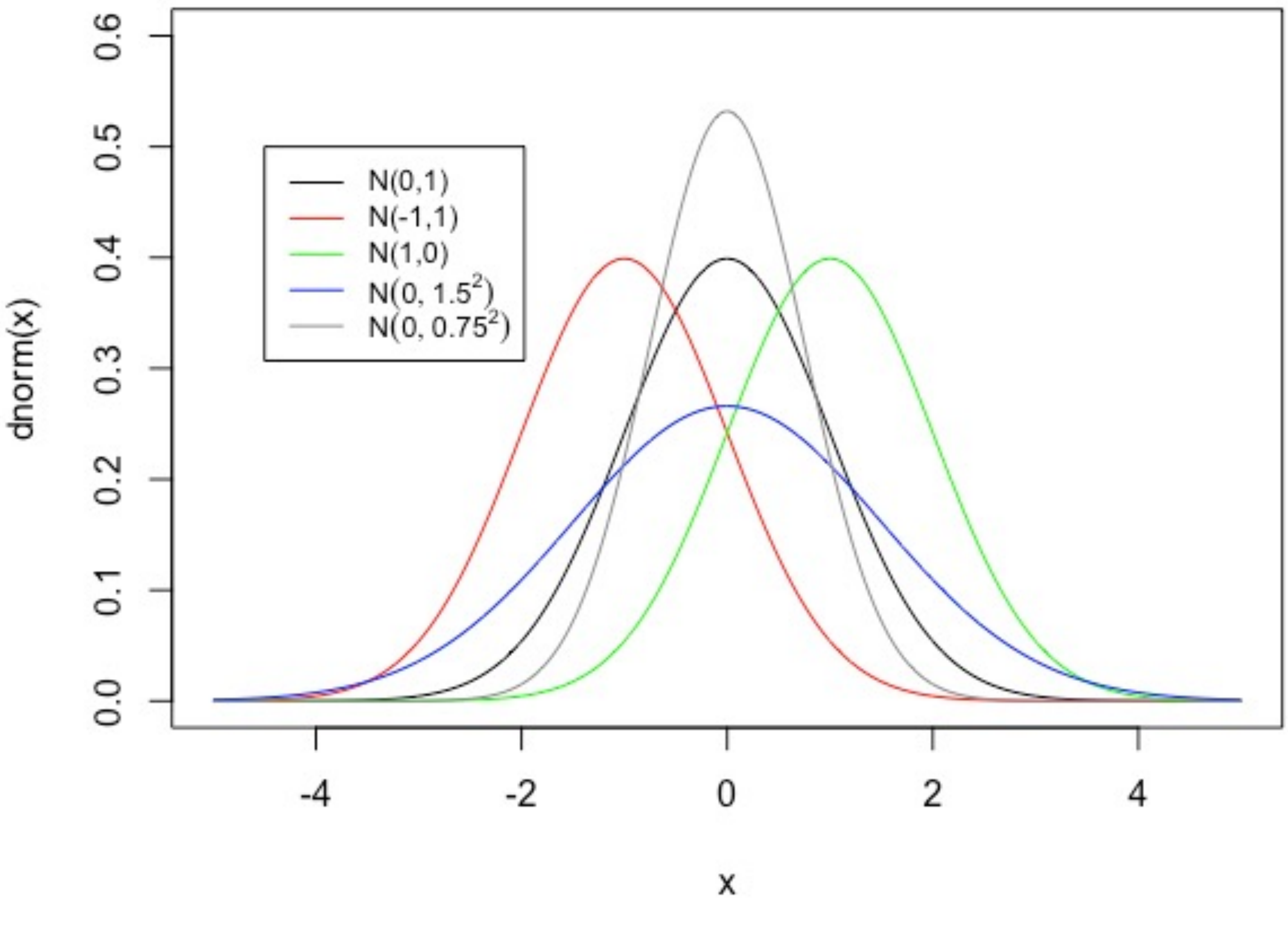
Normal Distribution

- Bell-shaped distribution with tendency for individuals to clump around the group median/mean
- Used to model many biological phenomena
- Many *estimators* have approximately normal sampling distributions (Central Limit Theorem)
- Notation: $X \sim \mathrm{N}(\theta, \sigma^2)$ where θ is mean and σ^2 is the variance

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

- The distribution 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);

Normal Distribution – Density Functions (pdf)



Example 1: Human Body Temperature

Distribution of temperature measures (^{o}F) for each cohort: UAVCW (1860–1940), NHANES I (1971–1975) and STRIDE (2007–2017).

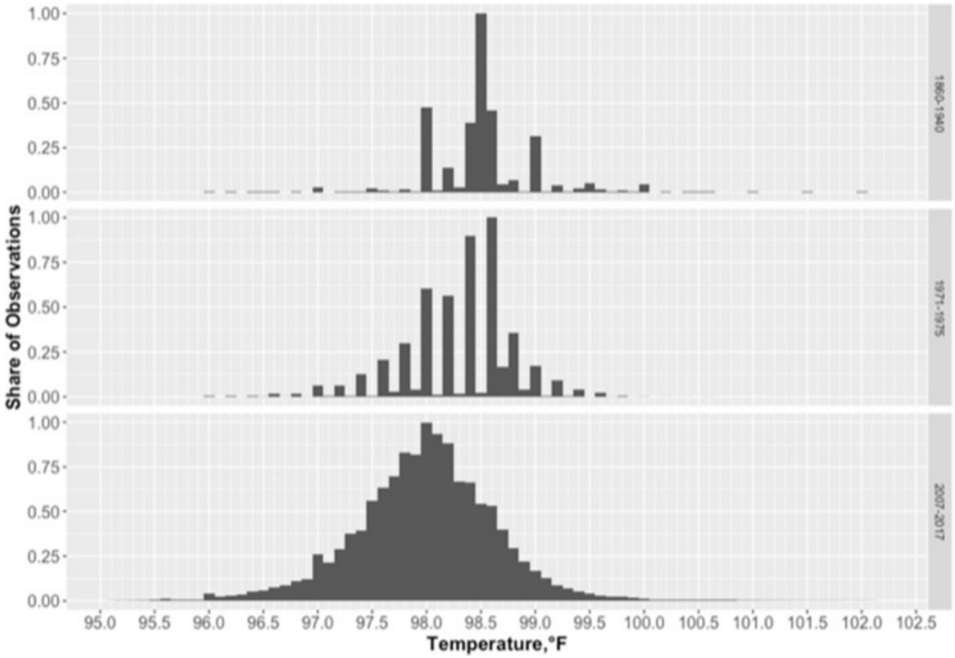
Source:

eLife. 2020; 9: e49555.

Published online 2020 Jan 7. doi: 10.7554/eLife.49555

Decreasing human body temperature in the United States since the Industrial Revolution

Myroslava Protsiv,[1] Catherine Ley,[1] Joanna Lankester,[2] Trevor Hastie,[3,4] and Julie Parsonnet[1,]



Likelihood

Suppose that

$$X_1, \dots, X_n | \theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2)$$

Then:

$$f(x_1, \dots, x_n | \theta, \sigma^2) = \prod_{i=1}^{n} f(x_i | \theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \theta}{\sigma}\right)^2}$$

Notice that

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

This shows that the pair $\left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i\right)$ is a two-dimensional sufficient statistic.

Lemma

Show that

$$\sum_{i=1}^{n}(x_i-\theta)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\theta-\bar{x})^2$$

Proof:

$$\sum_{i=1}^{n}(x_i-\theta)^2=\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\theta)^2$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2+2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\theta)+\sum_{i=1}^{n}(\bar{x}-\theta)^2$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-\theta)\sum_{i=1}^{n}(x_i-\bar{x})+n(\theta-\bar{x})^2$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-\theta)\times 0+n(\theta-\bar{x})^2$$

Case 1: σ^2 is known

Suppose we want to find a conjugate prior distribution for $\pi(\theta|\sigma^2)$. We have that

$$f(\theta|x_1,\dots,x_n,\sigma^2) \propto \pi(\theta|\sigma^2)\times e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}$$
$$\propto \pi(\theta|\sigma^2)\times e^{-\frac{1}{2\sigma^2}(\theta-\bar{x})^2}$$

This means that for $\pi(\theta|\sigma^2)$ to be conjugate it must include quadratic terms like $e^{c_1(\theta-c_2)^2}$. The simplest such distribution is the normal. That is, let

$$\theta \mid \sigma^2 \sim N(\mu_{,0}\ \tau_0^2)$$

Then

$$f(\theta|x_1,\dots,x_n,\sigma^2) = \frac{\pi(\theta|\sigma^2)\times f(x_1,\dots,x_n|\theta,\sigma^2)}{f(x_1,\dots,x_n|\sigma^2)}$$
$$\propto e^{-\frac{1}{2\,\tau_0^2}(\theta-\mu_0)^2}\times e^{-\frac{1}{2\sigma^2}(\theta-\bar{x})^2}$$

Posterior Distribution of $\theta \mid x_1, \dots, x_n, \sigma^2$

Recall HW 1, Q1:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2+\frac{1}{\tau_0^2}(\theta-\mu_0)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\tau_0^2}\mu_0^2-\frac{1}{\tau_n^2}\mu_n^2+\frac{1}{\tau_n^2}(\theta-\mu_n)^2$$

where $\tau_n^2=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}$ and $\mu_n=\tau_n^2\left(\frac{n}{\sigma^2}\bar{x}+\frac{1}{\tau_0^2}\mu_0\right)$

This shows that

$$f(\theta|x_1,\dots,x_n,\sigma^2)\propto e^{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2}$$

which proves that

$$\theta|x_1,\dots,x_n,\sigma^2\sim N(\mu_n,\tau_n^2)$$

Note: see pp. 70 – 71 of the textbook for a direct derivation of the result using “complete the square method”.

Posterior Analysis

- Posterior variance and precision:

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

Inverse variance is often referred to as the *precision,* so

posterior precision = prior precision + data precision

Note that this is not an identity true in general. The only thing we can says always holds is:

$$V(\theta) = E[V(\theta \mid X)] + V[E(\theta \mid X)]$$

This means that the *expected* posterior variance is always smaller than the *prior* variance.

Posterior Analysis

- Posterior mean:

$$\mu_n = \tau_n^2\left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau_0^2}\mu_0\right) = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\bar{x}$$

so, the posterior mean is a weighted average of the prior mean and the sample mean.

It will be closer to the sample mean when *n* is large and closer to the prior mean if *n* is small (or τ_0^2 is small).

Exercise 1

Find values for *n, x* and the Beta hyperparameters *a* and *b* in the binomial model such that the posterior variance is *larger* than the prior variance.

Prediction

Consider predicting a new observation X_{new} from the population after having observed ($X_1 = x_1$, . . . , $X_n = x_n$).

$$f(x_{new}|x_1, \dots, x_n) = \int f(x_{new}|\theta, x_1, \dots, x_n)\, f(\theta|x_1, \dots, x_n) d\theta$$

$$= \int f(x_{new}|\theta)\, f(\theta|x_1, \dots, x_n) d\theta$$

$$\propto \int e^{-\frac{1}{2\sigma^2}(x_{new}-\theta)^2} e^{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2}\, d\theta$$

The only bivariate density $f(x_{new}, \theta|x_1, \dots, x_n)$ that has the above form is the bivariate normal, which means that $X_{new}|x_1, \dots, x_n$ must be univariate normal and all we have to find is its mean and variance.

Prediction (continued)

Recall the laws of iterated expectation and variance:

$$E(X) \ = E[E(X \mid \theta)]$$
$$V(X) = E[V(X \mid \theta)] \ + V[E(X \mid \theta)]$$

This means that

$$E(X_{new}|x_1, \dots, x_n, \sigma^2)$$
$$= E[E(X_{new} \,|\theta, x_1, \dots, x_n, \sigma^2)|x_1, \dots, x_n, \sigma^2]$$
$$= E[\theta|x_1, \dots, x_n, \sigma^2] = \mu_n$$

$$Var(X_{new}|x_1, \dots, x_n, \sigma^2)$$
$$= E[V(X_{new} \,|\theta, x_1, \dots, x_n, \sigma^2)|x_1, \dots, x_n, \sigma^2]$$
$$+V[E(X_{new} \,|\theta, x_1, \dots, x_n, \sigma^2)|x_1, \dots, x_n, \sigma^2]$$
$$= E[\sigma^2] + Var[\theta|x_1, \dots, x_n, \sigma^2] = \sigma^2 + \tau_n^2$$

That is,

$$X_{new}|x_1, \dots, x_n \sim N(\mu_n, \tau_n^2 + \sigma^2)$$

Example 1 (from Lecture 1)

Ten subjects had both their heights measured and heir heights (in cm) where as follows:

169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3

which results in $\bar{x} = 165.52$. We will assume the population variance is known to be $\sigma^2 = 50$.

An experts gave the following prior distribution for the mean height: $\pi(\theta) =$ N(165, 2^2), which means $\mu_0 = 165, \tau_0^2 = 4$

The posterior mean and variance are

$$\mu_n = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \bar{x} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{5}} 165 + \frac{\frac{1}{5}}{\frac{1}{4} + \frac{1}{5}} 165.52$$

$$= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{5}} 165 + \frac{\frac{1}{5}}{\frac{1}{4} + \frac{1}{5}} 165.52 = 0.56(165) + 0.44(165.52) = 165.2311$$

$$\tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} = \frac{1}{\frac{1}{5} + \frac{1}{4}} = 2.22$$

Exercise 2

Compute a 95% posterior CI for θ

Answer: (162.31, 168.15)

Case 2: σ^2 is unknown

The posterior distribution is:

$$f(\theta, \sigma^2 | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta, \sigma^2)\pi(\theta, \sigma^2)}{f(x_1, \dots, x_n)}$$

We will find some simple conjugate prior. Specifying a joint prior distribution on parameters with different parameter spaces is very challenging, so usually it is done sequentially:

$$\pi(\theta, \sigma^2) = \pi(\theta | \sigma^2) \times \pi(\sigma^2)$$

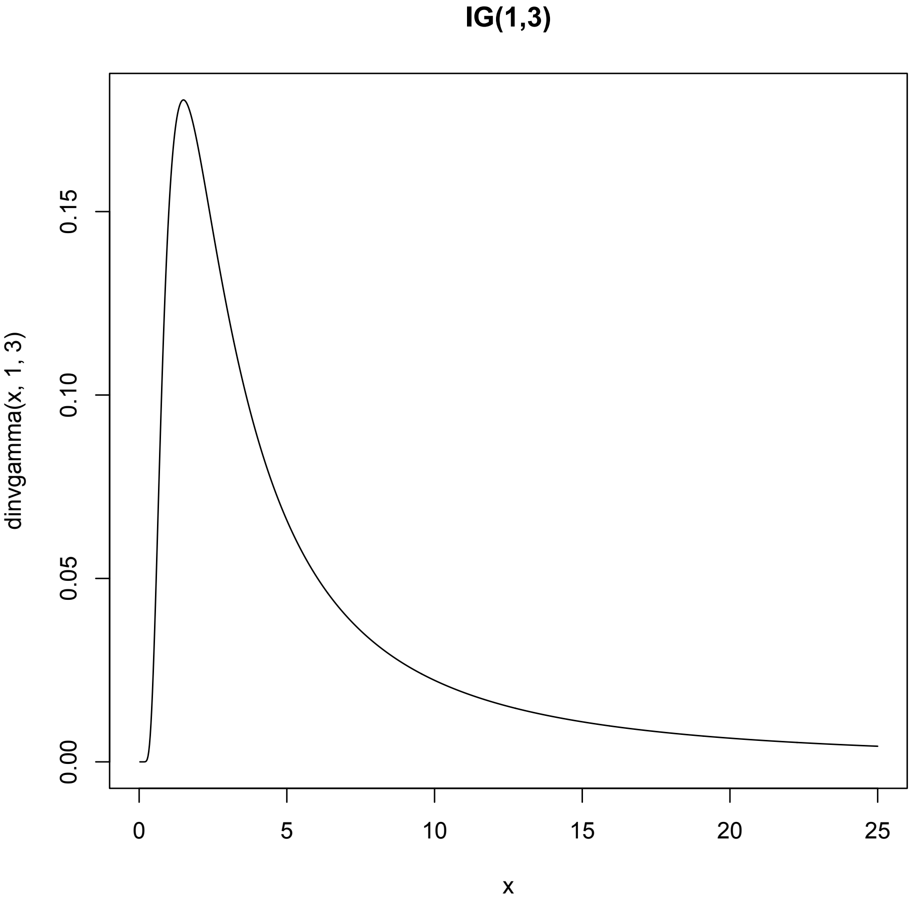
For $\pi(\theta | \sigma^2)$ we can choose a normal prior with variance $\frac{\sigma^2}{\kappa_0}$

For σ^2 we need a family of prior distributions that has support on $(0, \infty)$. One obvious choice 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 the precision $1/\sigma^2$.

Aside: Inverse Gamma Distribution

- Let $X \sim \text{IG}(a, b)$
- Then $\frac{1}{X} \sim \Gamma(a, b)$
- $E(X) = \frac{b}{a-1}$, for $a > 1$
- $Var(X) = \frac{b^2}{(a-1)^2(a-2)}$, for $a > 2$
- $f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}}, x > 0$

- It is available in R with the dinvgamma function from the package invgamma.



Posterior Inference

Suppose our model is:

$$X_1, \dots, X_n | \theta, \sigma^2 \sim N(\theta, \sigma^2)$$

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

The prior parameters (σ_0^2, ν_0) can be interpreted as the sample variance and sample size of prior observations.

The posterior distribution can be similarly decomposed:

$$f(\theta, \sigma^2 | x_1, \dots, x_n) = f(\theta | \sigma^2, x_1, \dots, x_n) \times f(\sigma^2 | x_1, \dots, x_n)$$

Recycling the results from Case 1, we can conclude that

$$\theta | \sigma^2, x_1, \dots, x_n \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$

where $\kappa_n = \kappa_0 + n$, $\mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n}$

Posterior Distribution of $\sigma^2 \mid x_1, \dots, x_n$

The posterior distribution of σ^2 can be obtained by performing an integration over the unknown value of θ.

$$f(\sigma^2|x_1,\dots,x_n) \propto \pi(\sigma^2) f(x_1,\dots,x_n|\sigma^2)$$

$$= \pi(\sigma^2)\int f(x_1,\dots,x_n|\theta,\sigma^2)\pi(\theta|\sigma^2)\,d\theta$$

$$= \frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}(\sigma^2)^{-\frac{\nu_0}{2}-1}e^{-\frac{\frac{\nu_0\sigma_0^2}{2}}{\sigma^2}}\int \pi(\theta|\sigma^2)(\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}(\theta-\bar{x})^2}\,d\theta$$

= … (exercise)

We conclude that

$$\sigma^2|x_1,\dots x_n \sim IG\left(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2}\right)$$

where $\nu_n = \nu_0 + n$ and $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i-\bar{x})^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{x}-\mu_0)^2\right]$

Example 1 continued

We will use $\kappa_0 = \nu_0 = 1$ which suggests that our prior distributions are only weakly centered around the values μ_0 = 165 and $\sigma_0^2 = 50.$ The sample mean and variance are:

$$\bar{x} = 165.52, s^2 = 52.453$$

Then

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n} = 165.4727$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{x} - \mu_0)^2\right] = 47.4838$$

See R code for posterior distributions.

Monte Carlo Sampling

What if wanted to calculate E($\theta|x_1, \ldots, x_n$)?

The only thing we know is that

$$\theta|\sigma^2, x_1, \ldots, x_n \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$

but we don’t know the marginal posterior distribution of θ (without conditioning on σ^2).

Instead of trying to derive it with calculus, we can resort to the Monte Carlo method to generate:

$$\sigma^{2(1)}|x_1, \ldots x_n \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right), then\ \theta^{(1)} \sim N\left(\mu_n, \frac{\sigma^{2(1)}}{\kappa_n}\right)$$

…

$$\sigma^{2(S)}|x_1, \ldots x_n \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right), then\ \theta^{(S)} \sim N\left(\mu_n, \frac{\sigma^{2(S)}}{\kappa_n}\right)$$

Exercise 3

Derive the distribution

$$\theta | x_1, \ldots, x_n$$

Answer:
The distribution is a location-scale *t* with

$$df = n + \nu_0$$

$$location\ parameter = \mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n}$$

$$scale\ parameter = \frac{\sigma_n}{\sqrt{\kappa_n}}$$

Improper Priors

What if we want to make the priors as less informative as possible?

In the Beta-Binomial model this was easily done by choosing Beta(1, 1) prior (which is the same as U(0, 1)). But what do we do in the normal model setup?

We might want to choose $\kappa_0 = \nu_0 = 0$ as they represent the prior "sample size". However, this is technically impossible. But we can notice that the posterior mean and variance converge to the sample mean and variance if we let $\kappa_0 \to 0$ & $\nu_0 \to 0$

More formally, we can use an *improper* prior

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Notice this is not a pdf since $\iint \pi(\theta, \sigma^2) d\theta\, d\sigma^2 = \infty$. However, it can be shown that both posterior distributions are *proper* with

$$\sigma^2 | x_1, \ldots x_n \sim IG\left(\frac{n-1}{2}, \frac{1}{2} \sum (x_i - \bar{x})^2\right)$$

Exercise 3

In the Binomial model consider the prior

$$\pi(\theta) \propto \frac{1}{\theta}$$

Find the posterior distribution and show that it is proper iff

$$\sum_{i=1}^{n} x_i \geq 1$$

Bias

Recall from frequentist statistics that an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is called *unbiased* if $\mathrm{E}\left(\hat{\theta}\right) = \theta$.

However, this concept is not applicable in Bayesian analysis since the parameter is a random variable.

For example, in the normal model

$$E(\theta|x_1, .., x_n) = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \bar{x} = w\bar{x} + (1 - w)\mu_0$$

The only way to compare this to a “true value” θ is to think that we know the data were generate from a specific θ, then the “bias” of the Bayesian estimator is $w\theta + (1 - w)\mu_0$, so in a sense Bayesian estimators are always “biased”.

MSE

Bias is an overrated property of estimators. What is more desirable is that the *mean squared error* is as small as possible.

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

It can be shown that

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$$

Bayesian estimators have smaller variances than their frequentist counterparts, and often they will also have smaller MSE. Some argue that if you know even just a little bit about the population you are about to sample from, you should be able to find values of the hyperparameters of the prior such that this inequality holds.