6.1 Normal model

As an example of a 2-dimensional problem, consider normal model for X, $\pi(X) = \mathrm{N}(\mu, \sigma^2)$, with unknown parameters μ and σ. Sometimes, another parametrization is used by defining *precision* $\tau = 1/\sigma^2$ instead of variance. (This is used in WinBUGS and OpenBUGS). **Note: notations get easily mixed! Below $\mathrm{N}(\mu, \sigma^2)$ can be casually written as $\mathrm{N}(\mu, \tau)$ which should not be understood as if τ was in the place of variance. Remember:** $\tau = 1/\sigma^2$.

Before the 2-dimensional problem, take a look at the one-dimensional problems where one of the parameters is assumed to be 'known'.

6.1.1 Unknown mean, known variance

Assume that <u>variance σ^2 is known</u>, but mean μ unknown. We would like to estimate the mean. Consider first a single observation X_i only. The conditional density of the observation is

$$\pi(X_i \mid \mu, \sigma) = \mathrm{N}(X_i \mid \mu, \sigma^2) = \underbrace{\mathrm{N}(X_i \mid \mu, \tau)}_{\text{notation with } \tau} \propto \exp(-0.5\tau(X_i - \mu)^2).$$

where $\tau = 1/\sigma^2$ is the *precision.* As always, before calculating posterior of μ, we need to choose the prior. Assume that, for all practical purposes it is acceptable to consider the whole set $\mathbb{R}$ of real numbers as the range of possible values. It is possible to use a conjugate prior density, $\mathrm{N}(\mu_0, \tau_0)$:

$$\pi(\mu) \propto \exp(-0.5\tau_0(\mu - \mu_0)^2).$$

With the single measurement X_i, the posterior density would be of the form

$$\pi(\mu \mid X_i, \tau, \mu_0, \tau_0) \propto \exp(-0.5(\tau_0(\mu - \mu_0)^2 + \tau(X_i - \mu)^2)),$$

and this is the same as

$$\mathrm{N}\Big(\frac{n_0\mu_0 + X_i}{n_0 + 1}, \frac{\sigma^2}{n_0 + 1}\Big),$$

where $n_0 = \tau_0/\tau$ can be interpreted as *a priori* sample size. The normal density is obtained from the bayes formula by using the technique of completing a square. (See e.g. [9] BSM p. 62). The posterior mean can be written as weighted average

$$w\mu_0 + (1 - w)X_i,$$

where the weight is $w = \tau_0/(\tau_0 + \tau)$.

Next, assume the data has several values $X_1, \ldots, X_N$. The probability of the whole data set can be written using the average $\bar{X} = \sum X_i/N$ (which is the sufficient statistic):

$$\pi(\bar{X} \mid \mu, \sigma) = \mathrm{N}(\bar{X} \mid \mu, \sigma^2/N) = \mathrm{N}(\bar{X} \mid \mu, N\tau).$$

By using bayes formula, this leads to the posterior

$$\mathrm{N}\Big(\frac{n_0\mu_0 + \bar{X}}{n_0+1}, \frac{\sigma^2/N}{n_0+1}\Big),$$

with $n_0 = \tau_0/(N\tau)$. The posterior mean and variance can also be written in this form:

$$E(\mu \mid X) = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{X}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} \qquad V(\mu \mid X) = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}.$$

Improper prior. When the prior precision τ_0 approaches zero, the prior density becomes flat and approaches zero everywhere. To describe a 'distribution' that is flat everywhere, we define an improper uniform density, $\pi(\mu) \propto 1$. The posterior density still exists, becoming $\mathrm{N}(\bar{X}, \sigma^2/N)$. The posterior mean then equals sample mean, and posterior variance equals the variance of the sample average. This is a perfect mirror image of the non-bayesian approach where a sampling distribution is derived for a *statistic*, such as sample mean, whereas the unknown population mean μ is considered constant. In bayesian inference μ is unknown, therefore random, but the data $\bar{X}$ is known, therefore constant. The roles of $\bar{X}$ and μ are reversed in the two paradigms:

frequentist says: $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/N)$ bayesian says: $\mu \sim \mathrm{N}(\bar{X}, \sigma^2/N)$

6.1.2 Unknown variance, known mean

It is next assumed that the mean μ is known, and we would like to estimate the unknown variance σ^2, (or precision τ). It is not sensible to estimate variance unless there are several (at least more than one) observations. Therefore, we assume that we have some number of observations $X = X_1, \ldots, X_N$. We can start again with the conditional density of all observations:

$$\pi(X \mid \mu, \sigma) \propto \sigma^{-N} \exp(-\frac{1}{2\sigma^2}\sum_i^N (X_i - \mu)^2).$$

$$= (\sigma^2)^{-N/2} \exp(-\frac{N}{2\sigma^2} s_0^2) = \tau^{N/2} \exp(-\frac{N\tau}{2} s_0^2)$$

where we have used the notation:

$$s_0^2 = \frac{1}{N}\sum_i^N (X_i - \mu)^2.$$

Since τ is unknown we must choose a prior for it. Alternatively, we could work out using $\sigma^2 = 1/\tau$, but let's use τ, because that is actually the parametrization used in WinBUGS and OpenBUGS. A conjugate choice would be Gamma(α, β) -distribution The posterior is then proportional to

$$\tau^{N/2} \exp(-\frac{N\tau}{2} s_0^2) \times \tau^{\alpha-1} \exp(-\beta\tau) = \tau^{N/2+\alpha-1} \exp(-(\frac{N}{2} s_0^2 + \beta)\tau)$$

Which can be recognized as Gamma$(N/2 + \alpha, \frac{N}{2} s_0^2 + \beta)$. An uninformative Gamma-prior is again obtained by setting α, β 'nearly zero'. So, in the limit the posterior would be Gamma$(\frac{N}{2}, \frac{N}{2} s_0^2)$ which has mean $1/s_0^2$. Setting $\alpha = \beta = 0$ in the Gamma-prior density gives $\pi(\tau) \propto \tau^{-1}$. By making the transformation $\theta = 1/\tau$, we get $\pi(\theta) \propto \theta \mid \frac{\mathbf{d}\theta^{-1}}{\mathbf{d}\theta} \mid = 1/\theta$. Hence, the corresponding improper prior for $\sigma^2 = 1/\tau$ is $\pi(\sigma^2) \propto 1/\sigma^2$.

6.1.3 Unknown mean and unknown variance

The previous solutions provided conditional distributions $\pi(\mu \mid \tau, \text{data})$ and $\pi(\tau \mid \mu, \text{data})$. These are called *full conditional distributions* which are obtained from the joint posterior density, based on the data and the two (independent) priors $\pi(\mu)$ and $\pi(\tau)$. These could be used for drawing random samples of μ and τ (one after another) from these full conditionals, which finally produces samples from the joint posterior distribution. (Gibbs sampling).

(1) Conjugate prior for the 2D-problem can be formulated as

$$\pi(\mu, \tau) = \pi(\mu \mid \tau)\pi(\tau) \quad \text{or } \pi(\mu, \sigma^2) = \pi(\mu \mid \sigma^2)\pi(\sigma^2)$$

In this case, joint posterior density can still be solved as a known distribution. A common choice is to use normal-inverse gamma prior for (μ, σ^2) so that an inverse gamma prior is applied for σ^2 and a conditional normal density for μ: $\mathrm{N}(\mu_0, c\sigma^2)$. In other words, the prior for (μ, τ) is then normal-gamma, with density

$$\pi(\mu, \tau) = (2\pi c)^{-0.5}\tau^{-0.5}\exp(-\frac{\tau}{2c}(\mu - \mu_0)^2) \times \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau)$$

The resulting posterior for (μ, σ^2) is then normal-inverse gamma. For practical data analysis purposes, this 2D-prior specification is slightly problematic because it requires to specify the prior distribution of μ conditionally on τ. This can be difficult to get e.g. from expert opinions, or any judgements of the application context. It seems more natural to specify priors for μ and τ separately. This leads to independent priors:

(2) Independent priors can be chosen as

$$\pi(\mu, \tau) = \pi(\mu)\pi(\tau) \quad \text{or } \pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2)$$

In this case, it is not possible to choose the distributions $\pi(\mu)$ and $\pi(\tau)$ so that the joint posterior density could be solved in any familiar form. In this case, we are forced to numerical calculations instead of analytical solutions.

(3) Finally, improper prior would be $\pi(\mu, \tau) \propto 1/\tau$, or $\pi(\mu, \sigma^2) \propto 1/\sigma^2$.

A *computationally useful* result is always to find out the full conditional distributions of τ and μ (or, in general, with any multidimensional bayesian inference). A numerical method called Gibbs sampler can be constructed from these. This provides a way to draw samples from the joint posterior distribution of τ and μ. With a large enough sample, we can calculate everything we need from the posterior, as a Monte Carlo approximation.

Solution with improper priors:

The goal is to solve the posterior (joint) density $\pi(\mu, \sigma^2 \mid X)$, i.e. both parameters are unknown. The prior density is assumed **improper** and uninformative so that

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

This prior is the same as an improper uniform prior

$$\pi(\mu, \log(\sigma)) \propto 1.$$

First, there's some preliminary math that will be needed when solving the posterior density.

$$\sum_i^n (X_i - \mu)^2 = \sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Proof:

$$\sum_i^n (X_i - \mu)^2 = \sum_i^n (X_i^2 - 2X_i\mu + \mu^2)$$

$$= \sum_i^n (X_i^2 - 2X_i\mu + \mu^2 - \bar{X}^2 + \bar{X}^2 - 2X_i\bar{X} + 2X_i\bar{X})$$

$$= \sum_i^n (X_i - \bar{X})^2 + \sum_i^n (\mu^2 - 2X_i\mu - \bar{X}^2 + 2X_i\bar{X})$$

$$= \sum_i^n (X_i - \bar{X})^2 + n(\mu^2 - 2\bar{X}\mu - \bar{X}^2 + 2\bar{X}\bar{X}) = \sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Then, using this 'trick', the posterior density can be solved as

$$\pi(\mu, \sigma \mid X) \propto \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} \sum_i^n (X_i - \mu)^2)$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[\sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{X} - \mu)^2]),$$

where $s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$.

The posterior density is finally solved by using factorization:

$$\pi(\mu, \sigma^2 \mid X) = \pi(\mu \mid \sigma^2, X)\pi(\sigma^2 \mid X).$$

We already know from earlier results that $\pi(\mu \mid \sigma^2, X) = \mathrm{N}(\bar{X}, \sigma^2/n)$. Therefore, we only need to find out what the marginal density $\pi(\sigma^2 \mid X)$ is. This can be calculated from the joint density by integrating over μ:

$$\pi(\sigma^2 \mid X) \propto \int_{-\infty}^{\infty} \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{X} - \mu)^2])\mathbf{d}\mu$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}(n-1)s^2) \times \int_{-\infty}^{\infty} \exp(-\frac{n}{2\sigma^2}(\bar{X} - \mu)^2)\mathbf{d}\mu$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}(n-1)s^2) \times \sqrt{2\pi\sigma^2/n}$$

$$\propto (\sigma^2)^{-(n+1)/2} \exp(-\frac{(n-1)s^2}{2\sigma^2}).$$

In other words: $\pi(\sigma^2 \mid X) = \text{Scaled Inv-}\chi^2(n-1, s^2)$ or $\pi(\tau \mid X) = \text{Gamma}(\frac{n-1}{2}, \frac{n-1}{2}s^2)$.

Compare this with the earlier result where μ was assumed to be known.

The full joint density can thus be computed as a product of two known densities $\pi(\sigma^2 \mid X)$ and $\pi(\mu \mid \sigma^2, X)$. This is also convenient for Monte Carlo implementations, because we can then simulate both unknown parameters from these known distributions. This example happens to be such that it is also possible to solve the marginal posterior density of the mean $\pi(\mu \mid X)$. This follows from calculating the integral:

$$\pi(\mu \mid X) = \int_0^\infty \pi(\mu, \sigma^2 \mid X) \mathbf{d}\sigma^2.$$

The details are given in Gelman et al, [5]. As a result, the marginal posterior is found to be a t-distribution so that

$$\pi\Big(\frac{\mu - \bar{X}}{s/\sqrt{n}} \mid X\Big) = t_{n-1}.$$

References

[1] McGrayne S B: The theorem that would not die. Yale University Press. 2011.

[2] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[3] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[4] Christensen R, Johnson W, Branscum A, Hanson E: Bayesian Ideas and Data Analysis. CRC Press. 2011.

[5] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[6] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.

[7] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[8] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[9] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[10] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[11] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.

[12] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515-533. 2006.