3 Beyond binomial models

3.1 Poisson-distribution

Poisson-distribution is one of the most commonly used models in e.g. reliability research and epidemiology. It is used for describing number of 'rare events'. Poisson distribution can be derived as a limiting case of binomial distribution $\mathrm{Bin}(N_k, r_k)$ when $N_k \to \infty$ and $r_k \to 0$ so that the product $N_k r_k \to \lambda$, when $k \to \infty$. Then, the (Poisson) distribution of a single observation $X \in \{0, 1, 2, 3, \dots\}$ is

$$P(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}.$$

The Poisson distribution also emerges from Poisson process (a special case of stochastic process) with constant intensity λ. If, e.g. accidents occur with constant intensity λ per time unit, then the expected number of accidents in a time unit is λ and the number of them (per time unit) follows Poisson distribution with parameter λ, which is both the mean and the variance of Poisson distribution. Due to additivity of Poisson variables, if $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$, then $X + Y \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$. Likewise, the number of events during time T has Poisson distribution $\mathrm{Poisson}(\lambda T)$. In a Poisson process with constant intensity λ, the waiting time until next event is exponentially distributed with mean $1/\lambda$, regardless of the past history, (if λ given).

As a conjugate distribution, the prior of λ is $\mathrm{Gamma}(\alpha, \beta)$-density

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which leads to the posterior:

$$\pi(\lambda \mid X) \propto \frac{\lambda^X}{X!} e^{-\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which is, up to a normalizing constant, the same as

$$\lambda^{x+\alpha-1} e^{-(1+\beta)\lambda}.$$

In other words: $\mathrm{Gamma}(X + \alpha, 1 + \beta)$-density. The posterior mean is thus

$$E(\lambda \mid X, \alpha, \beta) = \frac{X + \alpha}{1 + \beta} = \frac{1}{1 + \beta} X + \frac{\beta}{1 + \beta} \frac{\alpha}{\beta}$$

which is a weighted average of prior mean α/β and X. If we have a series of observations $X_1, \dots, X_n$, an analogous result can be derived.

3.1.1 Example: Asthma mortality

Epidemiological Example from Gelman [4]: Poisson model parameterized in terms of rate and exposure:

$$X_i \sim \mathrm{Poisson}(E_i \theta)$$

where X_i is the number of e.g. disease cases in a group with exposure E_i and θ is the unknown parameter of interest, the 'underlying rate'. The probability of the data $X = (X_1, \ldots, X_N)$ is

$$\pi(X \mid \theta) \propto \theta^{\sum_{i=1}^{N} X_i} \exp((-\sum_{i=1}^{N} X_i)\theta)$$

With the conjugate prior Gamma(α, β), the posterior is

$$\pi(\theta \mid X) = \text{Gamma}(\alpha + \sum_{i=1}^{N} X_i, \beta + \sum_{i=1}^{N} E_i)$$

Assume there were $X = 3$ deaths due to asthma in a city during a year, out of a population of 200000. Hence the crude estimate per 100000 per year would be 1.5 cases. The model for the observed count could be

$$X \sim \text{Poisson}(2\theta) = \text{Poisson}(E\theta)$$

where θ represents the 'underlying mortality rate' per 100000 per year, and E 'exposure'. To compute the posterior $\pi(\theta \mid X)$, we choose a conjugate prior $\pi(\theta) = \text{Gamma}(\alpha, \beta)$ by choosing (α, β) so that the prior represents reasonable background information. According to literature, the typical asthma mortality rate in Western countries would be around 0.6 per 100000. It is also known that values above 1.5 are rare. Hence, Gamma$(3, 5)$ prior has mean 0.6, standard deviation 0.35, and this prior also has $P(\theta < 1.44) = 97.5\%$. All this seems to fulfill both prior specifications. (The prior parameters can be chosen by trial and error). The posterior distribution is then Gamma$(6, 7)$, which has mean 0.86. That is substantial shrinkage towards prior distribution.

3.2 Exponential distribution

Assume a single observation $X \in \mathbb{R}^+$ (typical example: waiting times, time of next event) for which the conditional distribution is exponential:

$$\pi(X \mid \theta) = \theta \exp(-X\theta).$$

As a conjugate prior of θ, we choose Gamma(α, β), so that the posterior $\pi(\theta \mid X)$ becomes Gamma$(\alpha + 1, \beta + X)$. The posterior mean is

$$E(\theta \mid X, \alpha, \beta) = \frac{\alpha + 1}{\beta + X}$$

With a set of observations $X_1, \ldots, X_n$ (mean $\bar{X} = \sum_{i=1}^{n} X_i / n$) we get

$$\pi(X \mid \theta) = \theta^n \exp(-n\bar{X}\theta)$$

which leads to the posterior Gamma$(\alpha + n, \beta + n\bar{X})$, so that the Gamma$(\alpha, \beta)$ prior can be thought as equivalent of $\alpha - 1$ prior observations $X_1^0, \ldots, X_{\alpha-1}^0$ for which the sum $\sum X_i^0$ equals to β.

3.2.1 Censored data

In survival analysis and reliability applications, it is common that the 'failure times' (times of death, infections, illness, etc.) are exactly known for only some individuals. For others, the time can be censored, which means that we only know that the event has not happened before some known time point. (This is also information!). Often, the censoring time can be the ending time of the follow-up period, or ending time of the study, T. The probability for such event is the *survival probability*: $P(X_i > T \mid \theta) = 1 - P(X_i < T \mid \theta) = 1 - F(T \mid \theta) = \exp(-\theta T) = S(T \mid \theta)$. The conditional probability of the whole data is then of the form

$$P(X \mid \theta) = \prod_{i=1}^{k} \theta \exp(-\theta X_i) \times S(T \mid \theta)^{n-k} = \theta^k \exp(-\theta[\sum_{i=1}^{k} X_i + (n-k)T]).$$

The posterior is then Gamma$(\alpha + k, \beta + \sum_{i=1}^{k} X_i + (n-k)T)$. More generally, we may know that for some individuals the event occurred before some given time, or between two given times. In each case, this information should be included by writing the corresponding conditional probability. (This is sometimes called as the 'full likelihood'). For example, if some events are only known to have been before time T_1 and some are known to be after time T_2, and for the rest we know the exact time, then the full likelihood would be of this form

$$P(X \mid \theta) = [F(T_1 \mid \theta)]^{\sum 1_{\{i \in E_1\}}} \times [S(T_2 \mid \theta)]^{\sum 1_{\{i \in E_2\}}} \times \prod_{i \in E_3} \theta \exp(-\theta X_i).$$

Note: by using the cumulative probability function F, probability expressions for all different situations of censoring might be written.

Note: when the event time is known, the conditional probability of this observation is $P(X_i \mid \theta) = \theta \exp(-\theta X_i)$, but when the censoring time is known, the observation can be interpreted as a Bernoulli variable (indicator variable!) that was one:

$$Y_i = \begin{cases} 0 & \text{if } X_i < T \\ 1 & \text{if } X_i > T \end{cases}$$

so that $P(Y_i = 1 \mid \theta) = S(T \mid \theta)$.

3.3 Normal-distribution

The normal, or Gaussian, distribution is the most widely used model and has connections to many other models and their asymptotic approximations. For an example of bayesian inference, consider that a measurement, e.g. temperature, is measured from N items, resulting to $X_1, \ldots, X_N$ as the observed temperatures. These are assumed to be normally distributed with mean μ and variance σ^2, representing the infinite population from which the items are drawn. We then have two unknown parameters in our model. Consider first estimating one of them, assuming the other as 'known', and finally estimating both.

3.3.1 Estimating the mean

Assume that variance σ^2 is known, but mean μ unknown. We would like to estimate the mean, representing the average temperature in an 'infinite' population of items. Consider first a single observation. The conditional density is

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

where $\tau = 1/\sigma^2$ is the *precision*. (Gaussian model is parameterized using precision in WinBUGS and often in bayesian notation. Be careful to note which notation is used!!). Before calculating posterior of μ, we need to choose the prior. As we know from physics, there is absolute minimum temperature, but for this example we assume that our measurements are well beyond absolute minimum. Therefore, for all practical purposes it is acceptable to consider the whole set $\mathbb{R}$ of real numbers as the range of possible measurement values. It is convenient 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, 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 [5] BSM p. 62). The posterior mean can be written as

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

where the weight is $w = \tau_0/(\tau_0 + \tau)$. The probability of the whole data set can be written using the average $\bar{X} = \sum X_i/N$:

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

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:

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

Improper prior. When the prior precision approaches zero, the prior density becomes flat and improper density, $\pi(\mu) \propto 1$, but 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 *statistics*, 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:

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/N) \qquad\qquad \mu \sim \mathrm{N}(\bar{X}, \sigma^2/N)$$

3.3.2 Estimating the variance

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:

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

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

where we have used the notation:

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

Since τ is unknown we must choose a prior for it. Following the presentation in Gelman et al [4], a convenient choice for the prior $\pi(\sigma^2)$ is a Scaled Inverse χ^2 distribution. The density function is:

$$\pi(\sigma^2 \mid \nu_0, \sigma_0^2) = \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \sigma_0^{\nu_0} (\sigma^2)^{-(\nu_0/2+1)} \exp(-\nu_0 \sigma_0^2 / 2\sigma^2)$$

It has two parameters, ν_0, σ_0^2, and it has some connections to other densities, which provide alternative ways in constructing the prior:

$$\sigma^2 \sim \text{Scaled Inv-}\chi^2(\nu_0, \sigma_0^2) = \text{Inv-}\Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}) \iff \tau = \frac{1}{\sigma^2} \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$$

$$\theta \sim \chi^2_{\nu_0} \iff \frac{\nu_0 \sigma_0^2}{\theta} \sim \text{Scaled Inv-}\chi^2(\nu_0, \sigma_0^2)$$

Using the Scaled Inverse-χ^2 prior, the posterior is of the form:

$$\pi(\sigma^2 \mid X) \propto \pi(\sigma^2) \pi(X \mid \sigma^2)$$

$$\propto (\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}) \times (\sigma^2)^{-N/2} \exp(-\frac{N\nu}{2\sigma^2})$$

$$= (\sigma^2)^{-((\nu_0+N)/2+1)} \exp(-\frac{\nu_0 \sigma_0^2 + N\nu}{2\sigma^2})$$

Which is the Scaled Inverse-χ^2 density:

$$\pi(\sigma^2 \mid X, \mu) = \text{Scaled Inv-}\chi^2(\nu_0 + N, \frac{\nu_0 \sigma_0^2 + N\nu}{\nu_0 + N}).$$

Note that the prior can be thought of as ν_0 observations with average squared deviation σ_0^2.

Improper prior. When the ν_0 parameter of the prior is set to zero, we obtain an improper prior

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

which does not integrate to one. Nevertheless, the posterior density still exists, and it is Scaled Inv-$\chi^2(N, \nu)$ where $\nu = \frac{1}{n} \sum (X_i - \mu)^2$. The prior is equivalent to the uniform improper prior $\pi(\log(\sigma)) \propto 1$.

3.4 Multiparameter models

In nearly all inference problems there is more than one unknown quantity. Often, only one of them is of interest and the others are *nuisance parameters*. Assume there are two unknown parameters θ_1, θ_2 (both can be vectors) and some set of data X. The posterior density is

$$\pi(\theta_1, \theta_2 \mid X) \propto \pi(X \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2),$$

and the marginal density of θ_1 is

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1, \theta_2 \mid X)\mathbf{d}\theta_2,$$

which can also be calculated as

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1 \mid \theta_2, X)\pi(\theta_2 \mid X)\mathbf{d}\theta_2.$$

This integral is usually not computed directly, but it shows an important structure that is used when hierarchical models are constructed, and also when MCMC algorithms are implemented.

Note: the unknown parameters θ can be 'unknown model parameters', or missing data variables, or variables to be predicted, or unobservable latent (hidden) variables. They are all simply unknown, and in bayesian inference they are all treated as unknown quantities, so that we aim to compute the posterior:

$$P(\text{'all unknowns'} \mid \text{'all known things'})$$

Note: it is difficult to visualize a posterior density for three or more unknown quantities. Therefore, we often plot one-dimensional marginal distributions, or two-dimensional marginal distributions for selected quantities of interest. This is always based on the full posterior density that can be multidimensional.

3.4.1 Multinomial model, unknown $r_1, \ldots, r_k$

Binomial model can be generalized to multinomial model by considering outcomes of several types instead of two types. For example, in a large bag there are balls of k different colours. The proportions of these are $r = r_1, \ldots, r_k$. A sample of N balls is drawn, and we observe the number of balls of each colour $X_1, \ldots, X_k$. The goal is now to solve the posterior density:

$$\pi(r_1, \ldots, r_k \mid X_1, \ldots, X_k).$$

Note that the unknown proportions have to sum to one: $\sum r_i = 1$. The conditional distribution of the data is now

$$P(X_1, \ldots, X_k \mid r_1, \ldots, r_k, N) = \binom{N}{X_1, \ldots, X_k} r_1^{X_1} \times \cdots \times r_k^{X_k}.$$

The conjugate prior density is $\text{Dir}(\alpha) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$:

$$\pi(r_1, \ldots, r_k) = \frac{\Gamma(\alpha_1, \ldots, \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} r_1^{\alpha_1 - 1} \times \cdots \times r_k^{\alpha_k - 1},$$

so that the posterior density will also be Dirichlet, with parameters $(\alpha_1 + X_1, \ldots, \alpha_k + X_k)$:

$$\propto r_1^{\alpha_1 + X_1 - 1} \times \cdots \times r_k^{\alpha_k + X_k - 1}.$$

Again, prior parameters $\alpha_1, \ldots, \alpha_k$ can be interpreted to represent 'prior data' so that the 'prior sample size' is $\sum \alpha_i$. A usual uninformative prior choice is $\text{Dir}(1, \ldots, 1)$, which is the generalization of $\text{Beta}(1, 1)$. The posterior means can be written as weighted mean of prior and data proportions

$$E(r_i \mid X, \alpha) = \frac{\alpha_i + X_i}{\sum(\alpha_i + X_i)} = \frac{\sum \alpha_i}{\sum(X_i + \alpha_i)} \frac{\alpha_i}{\sum \alpha_i} + \frac{\sum X_i}{\sum(X_i + \alpha_i)} \frac{X_i}{\sum X_i}$$

Note also that if $r \sim \text{Dir}(\alpha)$, then the marginal distribution of each r_j is $\text{Beta}(\alpha_j, \sum_i \alpha_i - \alpha_j)$, with variance $\alpha_j(\sum_i \alpha_i - \alpha_j)/((\sum_i \alpha_i)^2(\sum \alpha_i + 1))$. To simplify notations, write $A = \sum_i \alpha_i$. Then the marginal variance may be written as $\frac{\alpha_j}{A}(1 - \frac{\alpha_j}{A})/(A + 1)$.

If dirichlet distribution is not found in a software, the following result can be useful:

$$Z_i \sim \text{Gamma}(\alpha_i, 1) \quad \Rightarrow \quad (\frac{Z_1}{\sum Z_i}, \ldots, \frac{Z_k}{\sum Z_i}) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

Congdon, BSM, p. 38, shows a possible method for constructing an informative prior based on 'expert opinion'. It starts by picking up two estimates for the parameters. Denote them as $p_1, \ldots, p_k$ and $q_1, \ldots, q_k$. Their differences are $d_i = p_i - q_i$ and means are $\eta_i = (p_i + q_i)/2$. The expected value of the sum of squared differences can be written as

$$
\begin{aligned}
E(\sum(p_i - q_i)^2) &= \sum E(p_i^2) - 2 \sum E(p_i)E(q_i) + \sum E(q_i^2) \\
&= \sum(2E(p_i^2) - 2E(p_i)^2) \\
&= 2 \sum V(p_i) \\
&= 2 \sum \eta_i(1 - \eta_i)/(A + 1) \\
&= 2 \sum(\eta_i - \eta_i^2)/(A + 1) \\
&= 2(1 - \sum \eta_i^2)/(A + 1).
\end{aligned}
$$

This expected value and the prior 'observed' sum of squared differences of the two estimates are marked as equal. Then, using the 'observed' prior estimate η_i, we can solve the prior sample size A. The parameters of the Dir-prior are finally obtained as $(A\eta_1, \ldots, A\eta_k)$.

3.4.2   Normal model, unknown μ and σ^2

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:

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

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

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

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

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 $\pi(\sigma^2 \mid Y)$ is. This can be calculated from the joint density by integrating over μ:

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

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

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

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

In other words: $\pi(\sigma^2 \mid X) = $ Scaled Inv-$\chi^2(n-1, s^2)$.

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, [4]. As a result, the marginal posterior is found to be a t-distribution so that

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

3.5   Comment

The above posterior distributions were obtained using conjugate priors. Conjugate priors are convenient, because (1) the posterior density is among well known standard densities (exact solution exists), (2) the prior can be thought of as some amount of 'prior data'. On the other hand, conjugate priors may not be flexible enough to represent more complicated prior information. But if non-conjugate priors are used, then the posterior does not take the form of any standard distribution and we must use numerical methods for all computations. Conjugate priors can only be used for a limited number of problems. They can be useful as a first approach. Some examples are given in Table (1):

Data distribution	Prior
Binomial(n, p), n known	$p \sim$ Beta
Multinomial$(n, p_1, \ldots, p_k)$, n known	$p_1, \ldots, p_k \sim$ Dirichlet
Poisson(λ)	$\lambda \sim$ Gamma
N(μ, σ^2), σ known	$\mu \sim$ N
N(μ, σ^2), μ known	$\frac{1}{\sigma^2} \sim$ Gamma
MN$((\mu_1, \ldots, \mu_k), \Sigma)$, μ known	$\Sigma^{-1} \sim$ Wishart
Gamma(α, β), α known	$\beta \sim$ Gamma
Beta(α, β), β known	$\alpha \sim$ Gamma

Table 1: Some conjugate models.

References

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

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

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

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

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

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