ST 337 / ST 405: Bayesian Forecasting and Intervention

Dr Bärbel Finkenstädt Rand[1], Department of Statistics, University of Warwick

[1] B.F.Finkenstadt@warwick.ac.uk.

CONTENTS

CHAPTER 1

BAYESIAN INFERENCE

1.1 Dynamic models

Assume that we observe a sequence of realisations, $\{Y_1, Y_2, \ldots, Y_t, \ldots\}$, of a random quantity of interest at different times (e.g. sales of a particular product, a stock price index, amount of rainfall in a determined place, etc). Our main interest will be to forecast the future value or behaviour of this quantity on the basis of the observed past, say $\{y_1, \ldots, y_t\}$. This is a complicated, highly dimensional problem, so we need to find an intelligent way of modelling it.

Modelling must entertain changing conditions through time, yet remain of feasible complexity. This is achieved through a meaningful parameterisation, $\boldsymbol{\theta}_t$, which captures the key contextual features and summarises all the relevant information to forecast an observable of interest, say Y_t. Thus, at any time, the state of the system, $\boldsymbol{\theta}_t$, is described by a probability distribution $f\left(\boldsymbol{\theta}_t \mid K_t\right)$, where K_t denotes the set of relevant information.

The key feature of parameterisation in Dynamic Linear Models (DLM) is *conditional independence*. At any time t, given the value of the parameter $\boldsymbol{\theta}_t$, past $\{\ldots, t-1\}$, present t, and future $\{t+1, \ldots\}$ observations are **mutually independent**. All the relevant information to forecast the future at time t is summarized in $f\left(\boldsymbol{\theta}_t \mid K_t\right)$.

One fundamental feature of such a parameterisation is that $\boldsymbol{\theta}_t$ is typically of fixed (and *small* dimension). So, how should this parameter be updated as new information becomes available?

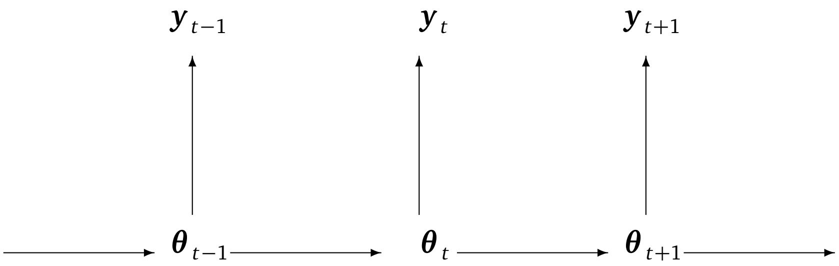


Figure 1.1. *Conditional independence. At any point t, given the value of the state parameter at the present time, $\boldsymbol{\theta}_t$, future observations are independent of present and past observations.*

1.2 Bayesian information updating

Statistics is an information science. Bayesian statistics is a paradigm regarding the way information should be dealt with. A Bayesian combines information from different sources in a *coherent* way. This is done through Bayes rule.

Theorem 1.1 (Bayes rule).
Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$

$$\mathrm{P}[A_i \mid B] = \frac{\mathrm{P}[B \mid A_i]\ \mathrm{P}[A_i]}{\sum_{j=1}^{\infty} \mathrm{P}[B \mid A_j]\ \mathrm{P}[A_j]}$$

Recalling the law of total probability, the denominator, $\mathrm{P}[B] = \sum_{j=1}^{\infty} \mathrm{P}[B \mid A_j]\ \mathrm{P}[A_j]$, is a constant (does not depend on A_i), so we may write

$$\mathrm{P}[A_i \mid B] \propto \mathrm{P}[B \mid A_i]\ \mathrm{P}[A_i]$$

where $\propto$ stands for ‘proportional to’. Bayes rule tells us that we can *update* our initial beliefs about A_i, expressed through the prior $\mathrm{P}[A_i]$, by combining them with the beliefs (information) we gather about some event B, related to A_i. Our final opinion is described by the posterior, $\mathrm{P}[A_i \mid B]$.

Remark 1.1 (Probability is the name of the game). *Note that the entire Bayesian framework is built upon the laws of probability. As we are uncertain about the sets A_i, we **must express any uncertainties through probability**.*

Remark 1.2 (Factorisation of joint probability). *From Theorem 1.1 it is immediate to see that*

$$\mathrm{P}[A_i \mid B]\ \mathrm{P}[B] = \mathrm{P}[B \mid A_i]\ \mathrm{P}[A_i]\,,$$

which is the familiar result regarding factorisation of a joint probability into a conditional and a marginal in two different ways.

1.2.1 Prior-to-posterior

Skipping a bit ahead, assume that the model $\mathcal{M} \equiv \{f(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ describes our beliefs about the behaviour of the *observable* quantity $\boldsymbol{x}$, conditional on a specific *state of the world* captured by the *parameter* $\boldsymbol{\theta}$. This is, if we could know the state of the world, we would be able to determine the value of the parameter. As, of course, we are not able to do so, we must express our (prior) beliefs about it through a probability distribution, $f(\boldsymbol{\theta} \mid K)$; where K stands for our initial knowledge about the state of the world.

Resorting to a continuous version of Bayes rule:

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}, K) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta})\, f(\boldsymbol{\theta} \mid K),$$

i.e. after observing $\boldsymbol{x}$, we update our initial uncertainty about $\boldsymbol{\theta}$ according to Bayes rule, and describe our updated beliefs through the *posterior distribution*, $f(\boldsymbol{\theta} \mid \boldsymbol{x}, K)$.

> We will drop K from the notation, but it must be kept in mind that ALL probabilities are conditional on this initial information.

Example 1 (The lady tasting tea).
There is a lady who claims that she can tell if milk was poured into the cup before or after the tea. To verify this assertion, $n = 10$ cups of tea are prepared with the milk poured before tea randomly (but we know which is which) and shuffled. Then, she tastes each cup and decides whether milk was poured in before tea or not. The number of correct answers, $r \in \{0, 1, \ldots, 10\}$, is recorded.

Three different analysts are confronted with the experiment. All of them agree that the model governing the random behaviour of each of the lady's answers is adequately described by a bernoulli distribution, with probability function

$$\mathrm{Ber}(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad x = 0, 1 \quad \theta \in (0, 1)$$

where θ is the probability of a correct answer, and that each answer can be considered independent of the rest. Thus, the joint distribution of the ten answers (the likelihood) is given by

$$\mathrm{L}(\theta \mid r) = \theta^r (1-\theta)^{n-r}$$

The three analysts agree that a Beta distribution, $\text{Be}\left(\theta \mid a_i, b_i\right)$, $i = A, B, C$, with pdf

$$f(\theta \mid a, b) = \frac{1}{B(a,b)}\,\theta^{a-1}(1-\theta)^{b-1}, \qquad 0 < \theta < 1 \quad a, b > 0,$$

is adequate to describe their prior beliefs; but that is as far as the agreement goes: Analyst A thinks the lady is bluffing, analyst B really believes that the lady can tell, and analyst C has not got a clue.

Analyst A 'I think she is just guessing, and that if she answers correctly it is by pure chance. Thus, as each cup of tea has the same probability of the milk being poured before or after, I would say that $\text{E}[\theta] = 1/2$ and I am very confident in this. So, my prior parameters are $\{a, b\}=\{3, 3\}$'

Analyst B 'I have a hunch! Not sure really why. I guess is that she looks trustworthy, but I am not very sure. My prior specification is $\{a, b\}=\{3, 2\}$'

Analyst C 'Don't know, anything can happen. Let's see. Thus for me, $\{a, b\}=\{1, 1\}$'

Figure 1.2 (a) depicts the prior distribution that each analyst has specified for this problem, based on their prior beliefs.

When the experiment is conducted and the relevant information gathered, each analyst updates her beliefs according to Bayes rule

$$\begin{aligned} p_i(\theta \mid r) &\propto \text{L}(\theta \mid r)\,\text{Be}\left(\theta \mid a_i, b_i\right) \\ &\propto \left[\theta^r (1-\theta)^{n-r}\right] \times \left[\theta^{a_i-1}(1-\theta)^{b_i-1}\right] \\ &\propto \theta^{r+a_i-1}(1-\theta)^{n-r+b-1}\,. \end{aligned}$$

This is the kernel of a Beta distribution and, as the posterior is a probability distribution, it has to integrate to one. Thus, each of them ends up with a Beta posterior density, $\text{Be}\left(\theta \mid a_i', b_i'\right)$, with parameters $a_i' = a_i + r$ and $b_i' = b_i + n - r$.

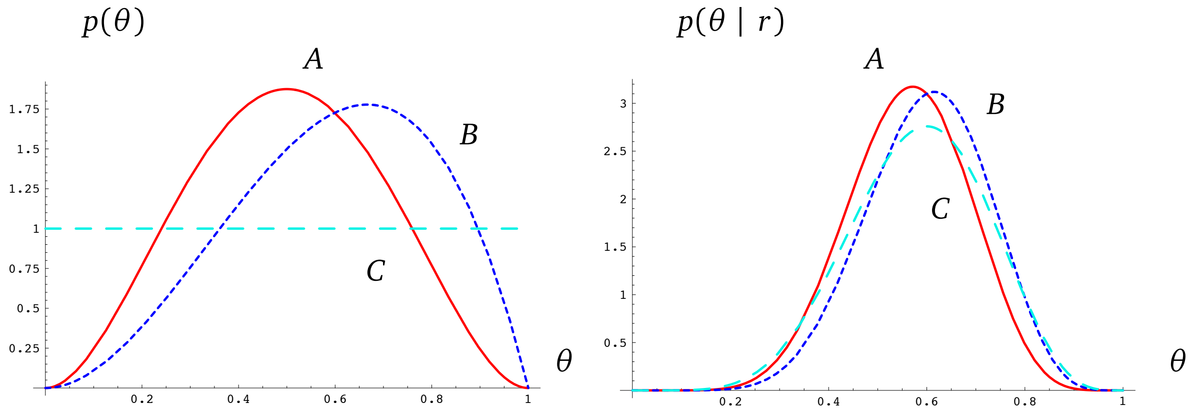
Suppose $n = 10$, and $r = 6$ is observed. The updated beliefs are described by their respective posterior distributions, depicted in Figure 1.2 (b).

Note that after gathering enough relevant information, differences in prior opinions seem to disappear. One way of investigate this phenomenon is to compare the highest posterior density (HPD) regions of size α.

Definition 1 (Highest density regions).
A region $\mathscr{R}_\alpha \subset \Theta$ is said to be a highest density region for θ of size α with respect to $p(\theta)$ if

i. $\text{P}[\theta \in \mathscr{R}_\alpha] = \alpha$

ii. $p(\theta_1) \geq p(\theta_2)$ *for all* $\theta_1 \in \mathscr{R}_\alpha$ *and* $\theta_2 \notin R_\alpha$



(a) *Prior distributions.* **(b)** *Posterior distributions.*

Figure 1.2. *Prior and posterior distributions for each tea-tasting-lady analysts.*

If $p(\theta)$ is a (prior/posterior/predictive) density, we refer to highest (prior/posterior/predictive) density regions.

For this particular problem, these regions are described in Table 1.1 and the posterior HPD are depicted in Figure 1.3.

Table 1.1. *Highest prior and posterior density regions of size 0.95 for each of the three analysts*

	prior	posterior
A	(0.15,0.85)	(0.33, 0.79)
B	(0.23, 0.96)	(0.36,0.83)
C	(0.025, 0.975)	(0.32, 0.84)

There we can see how data processing through Bayes rule aids in updating opinions. If data collecting continued we would end up with almost indistinguishable posterior distributions and, thus, the prior information would finally be ‘swamped’ by the data.

◁

Bayes rule is a natural information updating tool. This is a fundamental feature for our dynamic context in the rest of this module. Assume that the whole information $\{n, r\}$ is obtained in two stages. First we observe $\{n_1, r_1\}$, and in the second stage $\{n_2, r_2\}$, with $n_1 + n_2 = n$ and $r_1 + r_2 = r$. Then, after the first set of observations is available one has

$$\begin{aligned} f\left(\theta \mid r_1\right) &\propto \theta^{r_1}(1-\theta)^{n_1-r_1}\,\mathrm{Be}(\theta \mid a, b) \\ &\propto \theta^{r_1+a-1}(1-\theta)^{n_1-r_1+b-1} \end{aligned}$$

Thus, $f(\theta|r_1) = \mathrm{Be}\left(\theta \mid r_1 + a, n_1 - r_1 + b\right)$ and this distribution describes our beliefs after observing $\{n_1, r_1\}$, right before the second stage. Thus, this first posterior becomes our prior

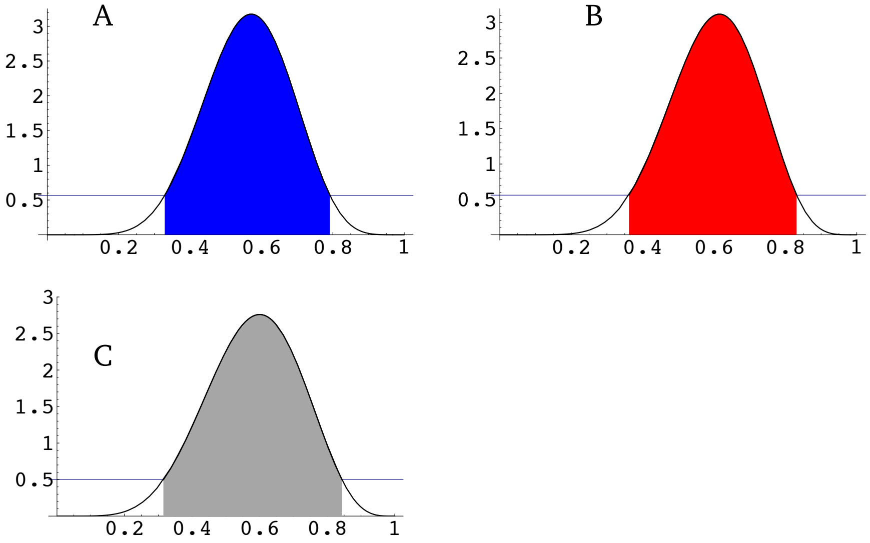


Figure 1.3. *Highest posterior density regions of size 0.95 for the three analysts.*

for the second stage. So, if no additional relevant information is supplied, our posterior distribution after also observing $\{n_2, r_2\}$ is

$$\begin{aligned} f\left(\theta \mid r_1, r_2\right) &\propto \theta^{r_2}(1-\theta)^{n_2-r_2}\,\mathrm{Be}\left(\theta \mid r_1+a, n_1-r_1+b\right) \\ &\propto \theta^{r_1+r_2+a-1}(1-\theta)^{n_1+n_2-r_1-r_2+b-1} \\ f\left(\theta \mid r_1, r_2\right) &= \mathrm{Be}(\theta \mid r+a, n-r+b) = f(\theta \mid r)\ , \end{aligned}$$

which is exactly the same as the distribution resulting from observing all the data at once.

Conjugate priors

According to Bayes rule,

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{x} \mid \boldsymbol{\theta})\ \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}\ ,$$

so, in order to fully know the posterior we must be able to perform the integral appearing in the denominator. Sometimes, though, this may prove unfeasible. Nevertheless, in some occasions, this step can be avoided, as we did in the previous example. One way to do so is to ensure that $f(\boldsymbol{\theta} \mid \boldsymbol{x})$ and $\pi(\boldsymbol{\theta})$ have the same functional form, i.e. that they belong to the same family. Formally,

Definition 2 (Conjugate family).
A family $\mathscr{F}$ of probability distributions on Θ is said to be conjugate (or closed under sampling) for a likelihood function $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ if, for every prior $\pi \in \mathscr{F}$, the posterior distribution, $f(\boldsymbol{\theta} \mid \boldsymbol{x})$, also belongs to $\mathscr{F}$.

Example 2.

Assume that we observe the random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, from a Poisson distribution with pdf

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \qquad \theta > 0, \quad x = 0, 1, \ldots .$$

The likelihood is then given by

$$\mathrm{L}(\theta \mid \boldsymbol{x}) = \Big(\prod_{i=1}^{n} x_i!\Big)^{-1} \theta^{\sum x_i} \exp[-n\,\theta]\,;$$

if we consider as a prior distribution a Gamma, $\mathrm{Ga}(\theta \mid a, b)$, with pdf

$$\pi(\theta \mid a, b) = \frac{b^a}{\Gamma[a]} \theta^{a-1} \exp[-b\,\theta], \qquad \theta > 0, \quad a, b > 0$$

it is readily seen that the posterior distribution

$$\begin{aligned} f(\theta \mid a, b, \boldsymbol{x}) &\propto \mathrm{L}(\theta \mid \boldsymbol{x})\, \pi(\theta \mid a, b) \\ &\propto \theta^{n\bar{x}+a-1} \exp\big[-(n+b)\,\theta\big]\,. \end{aligned}$$

The final expression is the kernel of a Gamma distribution, $\mathrm{Ga}(\theta \mid a', b')$, with parameters $a' = n\,\bar{x} + a$ and $b' = b + n$ and, as the posterior is also a distribution and must integrate to one,

$$f(\theta \mid a, b, \boldsymbol{x}) = \frac{b'^{a'}}{\Gamma[a']} \theta^{a'} \exp[-b'\,\theta]\,,$$

i.e. the posterior distribution belongs to the same family as the prior and, thus, a Gamma prior is the conjugate prior for a Poisson likelihood. ◁

Recall that the main goal is to be able to forecast a future realisation of the random quantity of interest, given the relevant information already gathered up. We now turn our attention to predictive distributions.

1.2.2 Predictive distributions

Typically, parameters are mere tools to aid in modelling. More often, one is interested in the actual observables. The Bayesian framework easily accommodates this need, one just has to derive the *predictive distribution*

$$f(\boldsymbol{y} \mid \boldsymbol{x}) = \int f\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\right) f(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}.$$

Notice that $f\left(\boldsymbol{y} \mid \boldsymbol{x}\right)$ **depends only on the observable $\boldsymbol{x}$.** Also, $\boldsymbol{y}$ can be any quantity whose distribution is indexed by $\boldsymbol{\theta}$, for instance: the next observation, x_{n+1}; or the mean of the next m observations, $\bar{x} = m^{-1}\sum_1^m x_{n+i}$, etc.

Remark 1.3. *As $f\left(\boldsymbol{y} \mid \boldsymbol{x}\right)$ involves only observables, we should concentrate on predictive distributions as often as possible.*

Example 1 (Continued).
After n trials, and before the lady tastes the next, $n+1$, cup, the analysts are asked their opinions about whether she will answer correctly. Defining $y=1$ if she answers correctly and $y=0$ if she misses, analysts' predictive distributions are derived as

$$\begin{aligned} f\left(y \mid r, a_i, b_i\right) &= \int_0^1 f\left(\theta \mid r\right) f\left(y \mid \theta\right)\, \mathrm{d}\theta \\ &= \int_0^1 \mathrm{Be}\left(\theta \mid r+a_i, n-r+b_i\right) \mathrm{Ber}\left(y \mid \theta\right)\, \mathrm{d}\theta \\ &= \frac{1}{\mathrm{B}(a_i+r, b_i+n-r)} \int_0^1 \left[\theta^{r+a_i-1}(1-\theta)^{n-r+b_i-1}\right] \times \left[\theta^{y}(1-\theta)^{1-y}\right] \mathrm{d}\theta \\ &= \frac{1}{\mathrm{B}(a_i+r, b_i+n-r)} \int_0^1 \theta^{y+r+a_i-1}(1-\theta)^{n-r+b-y}\, \mathrm{d}\theta \end{aligned}$$

Since the expression within the integral of the last line is the kernel of a Beta distribution with parameters $\{y+r+a_i, n-r+b+1-y\}$, and this must integrate to one over the parametric space $\Theta = (0,1)$, then

$$f\left(y \mid r, a_i, b_i\right) = \frac{\mathrm{B}(a_i+r+y, b_i+n-r+1-y)}{\mathrm{B}(a_i+r, b_i+n-r)}$$

with $\mathrm{B}(\cdot,\cdot)$ the Beta function. Thus,

$$\begin{aligned} \mathrm{P}[Y=1 \mid r, a_i, b_i] &= \frac{\mathrm{B}(a_i+r+1, b_i+n-r)}{\mathrm{B}(a_i+r, b_i+n-r)} \\ &= \frac{a_i+r}{a_i+b_i+n} \end{aligned}$$

and

$$\begin{aligned} \mathrm{P}[Y=0 \mid r, a_i, b_i] &= \frac{\mathrm{B}(a_i+r, b_i+n-r+1)}{\mathrm{B}(a_i+r, b_i+n-r)} \\ &= \frac{b_i+n-r}{a_i+b_i+n} \end{aligned}$$

Hence, the predictive distribution $f\left(y \mid r\right)$ is a Bernoulli probability, $\mathrm{Ber}\left(y \mid a_i''\right)$, with parameter $a_i'' = (a_i+r)/(a_i+b_i+n)$. For our analysts, we thus have

	A	B	C
$\mathrm{P}[Y=1 \mid r, a_i, b_i]$	$9/16$	$3/5$	$7/12$
	(0.563)	(0.600)	$(0.58\bar{3})$

$\lhd$

We will now analyse in detail one further model, which is of paramount interest in the DLM framework.

1.3 Gaussian model

We say that a real random quantity, x, follows a Gaussian (Normal) distribution with mean μ and variance σ^2, $\mathrm{N}\left(x \mid \mu, \sigma^2\right)$, if its pdf is given by

$$f\left(x \mid \mu, \sigma^2\right) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right], \qquad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \sigma^2 > 0\,.$$

It is usual to work with either the precision $\lambda = 1/\sigma^2$, or the standard deviation σ. In this section we will find it most useful to parameterise in terms of the precision.

This distribution has a long history, dating back to its introduction by Gauss around 1810 (and therefore its name) and since it has been extensively used as an 'error' distribution in Physics and Biology. It is also frequently used in econometric modelling, mainly due to its role in the Central Limit Theorem.

As illustrated in Figure 1.4, if a quantity x is distributed according to a Normal law, its mass is symmetrically distributed around the mean, which is also the mode of the distribution; the length of the highest probability regions is determined by the scale parameter: the larger (smaller) the precision (variance) the shorter the region.

In the next example, we will illustrate that the Normal-mean prior is closed under sampling.

1.3.1 Normal mean, known precision

Suppose that we are interested in making inference about some unknown quantity μ (the mean height of a population, the mean amount of radiation at a given site, the mean expenditure of a family in consumption goods, etc.). To this end, we collect measurements of the variable of interest (we select a random sample of the population and then measure the quantity of interest), and denote them by $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Assume that given our measurement system we know the precision $\lambda = \lambda_0$, so that given the unknown quantity μ, each x_i

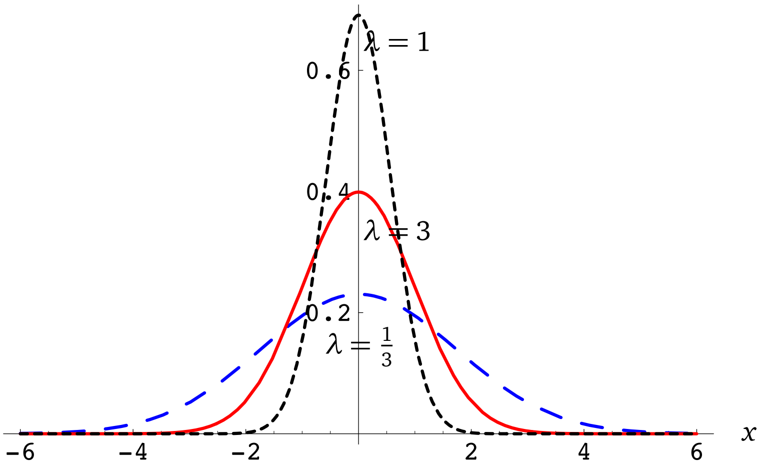


Figure 1.4. *The Normal density, centred at 0, with precision λ.*

is independently distributed according to a Normal, $\mathrm{N}\left(x_i \mid \mu, \lambda_0^{-1}\right)$. Thus,

$$l(\mu \mid \boldsymbol{x}, \lambda_0) = \prod_{i=1}^{n} \mathrm{N}\left(x_i \mid \mu, \lambda_0^{-1}\right)$$

$$= \left(2\pi\lambda_0^{-1}\right)^{-\frac{n}{2}} \exp\left[-\frac{\lambda_0}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

working out the exponent

$$= \left(2\pi\lambda_0^{-1}\right)^{-\frac{n}{2}} \exp\left[-\frac{\lambda_0}{2}\sum_{i=1}^{n}\left[(x_i-\bar{x})-(\mu-\bar{x})\right]^2\right]$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$

$$= \left(2\pi\lambda_0^{-1}\right)^{-\frac{n}{2}} \exp\left[-\frac{n\,\lambda_0}{2}\left[s^2+(\mu-\bar{x})^2\right]\right]$$

with $s^2 = n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Dropping the terms not involving μ we finally have

$$l(\mu \mid \bar{x}, \lambda_0) \propto \exp\left[-\frac{n\,\lambda_0}{2}(\mu-\bar{x})^2\right]$$

Now, assume that our prior uncertainty about μ can be adequately described by a Normal distribution, $\mathrm{N}\left(\mu \mid m, p^{-1}\right)$, so that its pdf is

$$f\left(\mu \mid m, p\right) = \sqrt{\frac{p}{2\pi}}\,\exp\left[-\frac{p}{2}(\mu-m)^2\right].$$

Updating our prior beliefs through Bayes rule, we get

$$
\begin{aligned}
f\left(\mu \mid \bar{x}, \lambda_0, m, p\right) &\propto l(\mu \mid \bar{x}, \lambda_0) f\left(\mu \mid m, p\right) \\
&\propto \exp\left[-\frac{n\lambda_0}{2}(\mu-\bar{x})^2\right] \exp\left[-\frac{p}{2}(\mu-m)^2\right]
\end{aligned}
$$

Again, working out the exponents

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\left[n\lambda_0(\mu-\bar{x})^2+p(\mu-m)^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[n\lambda_0(\mu^2-2\bar{x}\mu+\bar{x}^2)+p(\mu^2-2\mu m+m^2)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(n\lambda_0+p)\left(\mu^2-2\mu\frac{n\lambda_0\bar{x}+p\,m}{n\lambda_0+p}\right)+\bar{x}^2 n\lambda_0+m^2 p\right]\right\}
\end{aligned}
$$

completing the quadratic form and dropping those terms not depending on μ,

$$
\propto \exp\left[-\frac{n\lambda_0+p}{2}\left(\mu-\frac{n\lambda_0\bar{x}+p\,m}{n\lambda_0+p}\right)^2\right].
$$

We can recognise this last expression as the kernel of a Normal distribution for μ with mean

$$
m' = \frac{n\lambda_0}{n\lambda_0+p}\,\bar{x}+\frac{p}{n\lambda_0+p}\,m \qquad \text{and precision} \qquad \lambda' = n\lambda_0+p
$$

It is interesting to note that the posterior mean is a *weighted average* of the prior and the sample means, with weights depending on the measurement and prior precisions. Also, the posterior precision is the sum of the likelihood and the prior precisions.

Note that for a sample size, n, large enough, the posterior mean will be dominated by the sample mean and the posterior precision will hardly be influenced by the prior precision. So, again, when enough relevant information is accumulated, the prior influence in the posterior is swamped by that contained in the sample. However, this is not necessarily the case when n is relatively small.

Example 3.

Depicted in Figure 1.5 is the prior-to-posterior updating process for a Normal mean with known precision and four different sample sizes, $n = 3, 10, 30, 60$. We are assuming that a priori $\mu \sim \mathrm{N}\left(\mu \mid -4, 1/2\right)$ and that $\lambda_0 = 1$ and $\bar{x} = 0$ for all n.

We can see how for relatively small sample sizes, prior and sample information are quite different and thus the posterior lies somewhere in between both distributions; and also, how, as sample size grows, the posterior distribution resembles more and more the likelihood, regardless of the prior.

◁

We have already shown that the Normal-mean process with known variance is closed un-

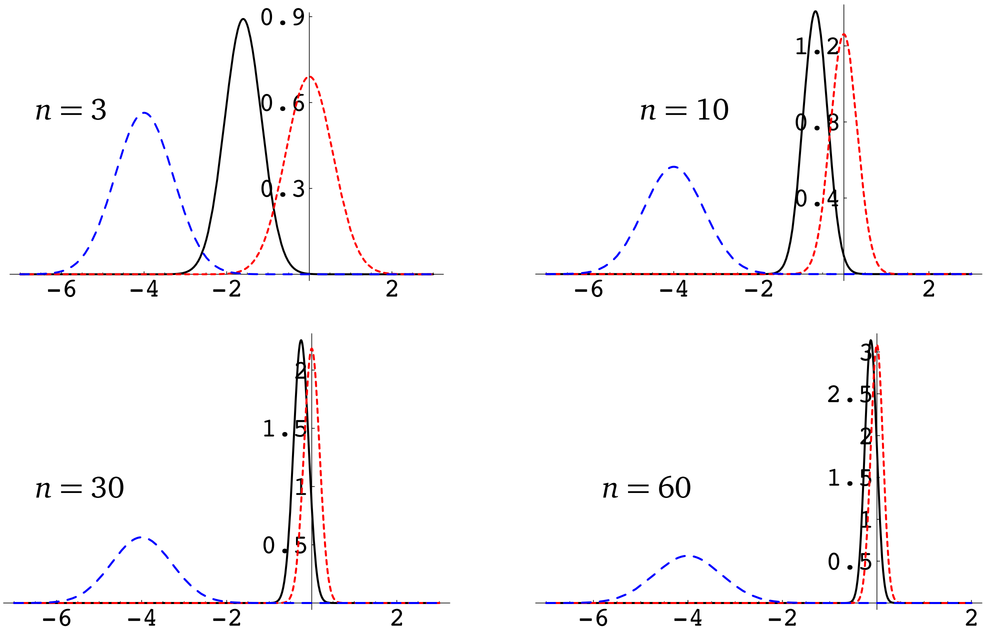


Figure 1.5. *Prior (broken line) to posterior (solid line) process in the Normal mean model with conjugate prior and different sample sizes. The (normalised) likelihood is the dotted line.*

der sampling. Now, recall that we will be often interested in forecasting, so let us investigate the predictive distribution associated with this model.

First, suppose that, given a sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, we want to describe our uncertainty about a future observation x_{n+1}. In Section 1.2.2 we stated that

$$f\left(x_{n+1} \mid \boldsymbol{x}\right) = \int_{\mathbb{R}} f\left(\mu \mid \boldsymbol{x}\right) f\left(x_{n+1} \mid \mu, \boldsymbol{x}\right) \, \mathrm{d}\mu$$

As we have assumed that $x_i \mid \mu$ are i.i.d. following a Normal distribution, we may then write

$$f\left(x_{n+1} \mid \boldsymbol{x}\right) = \int_{\mathbb{R}} f\left(\mu \mid \boldsymbol{x}\right) f\left(x_{n+1} \mid \mu\right) \, \mathrm{d}\mu$$

so that

$$f\left(x_{n+1} \mid \boldsymbol{x}\right) = \int_{\mathbb{R}} \mathrm{N}\left(\mu \mid m', \lambda'^{-1}\right) \mathrm{N}\left(x_{n+1} \mid \mu, \lambda_0^{-1}\right) \, \mathrm{d}\mu$$

Performing similar algebraic manipulations as those in the derivation of the posterior (expanding the quadratic forms, gathering similar terms and the completing the remaining quadratic form), we can see that

$$f\left(x_{n+1} \mid \boldsymbol{x}\right) \propto \int_{\mathbb{R}} \exp\left[-\frac{\lambda_0+\lambda'}{2}\left(\mu-\frac{\lambda_0 x_{n+1}+\lambda' m'}{\lambda_0+\lambda'}\right)^2-\frac{\lambda_0\lambda'}{2(\lambda_0+\lambda')}\left(x_{n+1}-m'\right)^2\right]\,\mathrm{d}\mu$$
$$\propto \exp\left[-\frac{\lambda_0\lambda'}{2(\lambda_0+\lambda')}\left(x_{n+1}-m'\right)^2\right],$$

since the integral in μ is simply the integral of a Normal kernel which integrates to $\sqrt{\dfrac{2\pi}{\lambda_0+\lambda'}}$, a quantity that does not involve x_{n+1}.

Again, surprisingly, this is the kernel of a Normal distribution. So we have just proved that

$$x_{n+1} \mid \boldsymbol{x} \sim \mathrm{N}\left(x_{n+1} \mid m', \lambda''^{-1}\right) \qquad \text{with} \qquad \lambda'' = \frac{\lambda_0\lambda'}{\lambda_0+\lambda'}$$

viz. the next observation follows a Normal distribution centred at the posterior mean and with variance equal to the sum of the posterior and the observational variances ($\lambda''^{-1} = \lambda_0^{-1}+\lambda'^{-1}$).

Example 3 (Continued).
Figure 1.6 illustrates the prior, posterior and predictive distributions for the four different sample sizes.

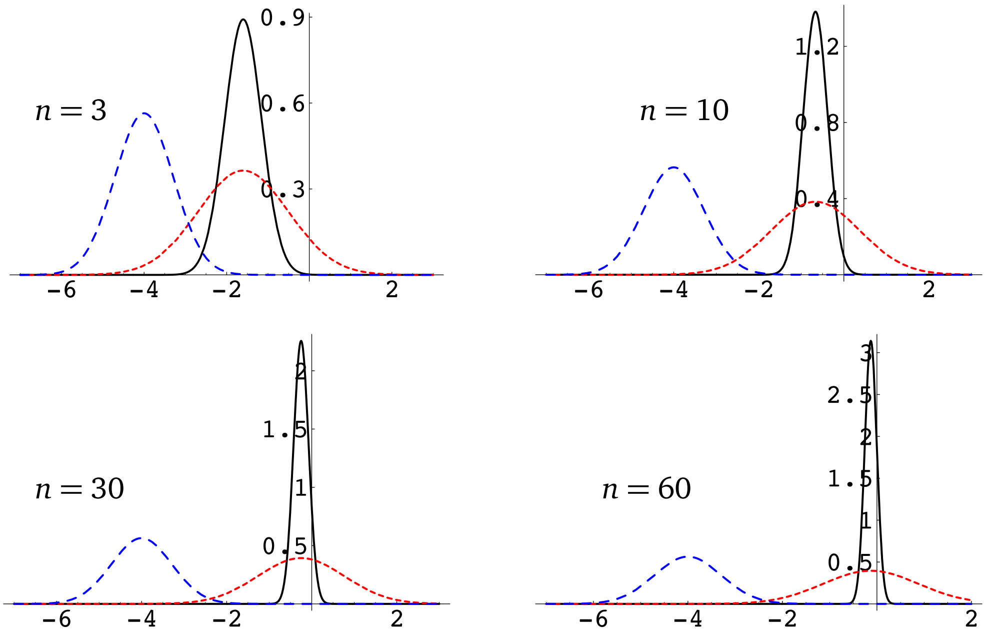


Figure 1.6. *Prior (broken line), posterior (solid line) and predictive (dotted) distributions in the Normal mean model with conjugate prior and different sample sizes.*

In the figure above it is apparent how the predictive distribution is centred at the same value as the posterior, but its uncertainty (variance) is increased due to the uncertainty embedded in forecasting x_{n+1} given μ. ◁

This result can be easily extended to some more involved functions of the observables. Suppose, for instance, that instead of the next observation, we are interested in making inference about the mean of the next m observations, $\bar{x}_m = \frac{1}{m}\sum_1^m x_{n+i}$. As $\bar{x}_m \sim \mathrm{N}\left(\bar{x}_m \mid \mu, (m\lambda_0)^{-1}\right)$, it is immediate to verify that

$$\begin{aligned} f\left(\bar{x}_m \mid \boldsymbol{x}\right) &= \int_{\mathbb{R}} f\left(\bar{x}_m \mid \mu, \lambda_0\right) f\left(\mu \mid \bar{x}\right)\, \mathrm{d}\mu \\ &= \mathrm{N}\left(\bar{x}_m \mid m', \lambda'''^{-1}\right) \end{aligned}$$

with $\lambda''' = \dfrac{m\lambda_0\lambda'}{m\lambda_0 + \lambda'}$.

1.3.2 Both parameters unknown

It is often the case that the observational precision is not known. In this case we must specify a prior distribution over the unknown quantities $\{\mu, \lambda\}$ and combine it with the likelihood, which in this case is

$$\begin{aligned} l(\mu, \lambda \mid \boldsymbol{x}) &\propto \lambda^{\frac{n}{2}} \exp\left[-\frac{n\lambda}{2}\left(s^2 + (\mu - \bar{x})^2\right)\right] \\ &\propto \left[\exp\left[-\frac{n\lambda}{2}(\mu - \bar{x})^2\right]\right]\left[\lambda^{\frac{n}{2}} \exp\left[-\frac{ns^2}{2}\lambda\right]\right] \\ &\propto l(\mu \mid \lambda, \boldsymbol{x})\, l(\lambda \mid \boldsymbol{x}) \end{aligned}$$

This factorization suggests that we may specify the prior as $f(\mu, \lambda) = f(\mu \mid \lambda) f(\lambda)$. Further, it also hints at the form that the two components should have in order to be closed under sampling.

According to the first term we can propose

$$f\left(\mu \mid \lambda\right) = \mathrm{N}\left(\mu \mid m, (k\lambda)^{-1}\right), \qquad \text{with known } k > 0.$$

Here, the parameter k can be interpreted roughly as the number of independent observations our prior information (about the mean) is worth –remember that the precision increases linearly with the number of observations in a Normal model.

Now, looking at the second term we can propose a Gamma distribution for the precision, i.e.

$$f\left(\lambda \mid a, b\right) = \frac{b^a}{\Gamma[a]} \lambda^{a-1} \exp[-b\lambda], \qquad a, b > 0\,.$$

It is readily seen that $\mathrm{E}[\lambda \mid a,b] = a/b$ and $\mathrm{Var}[\lambda \mid a,b] = a/b^2$; further, if $a > 1$ the distribution has a unique mode at $\mathrm{Mode}[\lambda \mid a,b] = (a-1)/b$. So, we can set the mean (or mode), $\lambda^* = a/b$, to be our best a priori guess about the value of the observational precision and then use the variance equality to set the remaining parameter, according to the confidence we wish to assign to our guess.

All together, we then have

$$\begin{aligned} f\left(\mu,\lambda \mid m,k,a,b\right) &= \mathrm{N}\left(\mu \mid m,(k\lambda)^{-1}\right)\,\mathrm{Ga}(\lambda \mid a,b) \\ &= \sqrt{\frac{k\lambda}{2\pi}}\,\exp\left[-\frac{k\lambda}{2}(\mu-m)^2\right]\,\frac{b^a}{\Gamma[a]}\,\lambda^{a-1}\,\exp[-b\,\lambda]\,, \end{aligned}$$

where the so-called hyperparameters $\{m,k,a,b\}$ have to be chosen to best reflect our prior beliefs.

Updating our prior beliefs according to Bayes rule (and dropping from the notation the dependence of the posterior on the prior hyperparameters), we get

$$\begin{aligned} f\left(\mu,\lambda \mid \bar{x},s^2\right) &\propto l(\mu,\lambda \mid \boldsymbol{x})\,f\left(\mu,\lambda \mid m,k,a,b\right) \\ &\propto \begin{array}{c} \left[\exp\left[-\frac{n\lambda}{2}(\mu-\bar{x})^2\right]\right]\left[\lambda^{\frac{n}{2}}\exp\left[-\frac{ns^2}{2}\lambda\right]\right] \quad \times \\ \sqrt{\frac{k\lambda}{2\pi}}\,\exp\left[-\frac{k\lambda}{2}(\mu-m)^2\right]\,\frac{b^a}{\Gamma[a]}\,\lambda^{a-1}\,\exp[-b\,\lambda] \end{array} \end{aligned}$$

arranging terms as before we have

$$\propto \lambda^{\frac{1}{2}}\exp\left[-\frac{k'\lambda}{2}(\mu-m')^2\right]\,\lambda^{a'-1}\exp[-b'\lambda]$$

where $k' = k+n$,

$$m' = \frac{k\,m+n\,\bar{x}}{k+n} \qquad a' = \frac{n}{2}+a \quad \text{and} \quad b' = \frac{1}{2}\left(n\,s^2+2\,b+\frac{n\,k}{k+n}(m-\bar{x})^2\right)$$

We have just proved that the Normal model with both parameters unknown is closed under sampling with a *Normal-Gamma* prior distribution; i.e., the joint posterior is also a Normal-Gamma.

Moreover, using the well known equalities for conditional expectations,

$$\mathrm{E}[X] = \mathrm{E}_Y\left[\mathrm{E}_{X\mid Y}[X \mid Y]\right] \quad \text{and} \quad \mathrm{Var}[X] = \mathrm{Var}_Y\left[\mathrm{E}_{X\mid Y}[X \mid Y]\right] + \mathrm{E}_Y\left[\mathrm{Var}_{X\mid Y}[X \mid Y]\right]$$

we can calculate

$$\mathrm{E}[\mu \mid \boldsymbol{x}] = \mathrm{E}_\lambda[m'] = m' = \frac{k\,m + n\,\bar{x}}{k'} \qquad \mathrm{Var}[\mu \mid \boldsymbol{x}] = \mathrm{Var}_\lambda[m'] + \mathrm{E}_\lambda\big[\frac{1}{\lambda\,k'}\big] = \frac{1}{k'}\frac{b'}{a'-1}$$

$$\mathrm{E}[\lambda \mid \boldsymbol{x}] = \frac{a'}{b'} \qquad \mathrm{Var}[\lambda \mid \boldsymbol{x}] = \frac{a'}{b'^2}\,.$$

So, as $n \to \infty$, $m' \to \bar{x}$ and $\mathrm{Var}[\mu \mid \boldsymbol{x}] \to 0$ as in the one parameter case. Also $\mathrm{E}[\lambda \mid \boldsymbol{x}] \to 1/s^2$ (the MLE) and $\mathrm{Var}[\lambda \mid \boldsymbol{x}] \to 0$. Like in the previous case, the influence of the prior in the posterior is gradually attenuated as the sample size increases.

We can go a little further. Usually we will concentrate on the mean, μ, itself, without referring to the precision. The natural way to do this within the Bayesian framework is to integrate out the nuisance parameter, λ. To this end, again we can collect similar terms in the joint posterior such that

$$f\left(\mu \mid \bar{x}, s^2\right) \propto \int_0^\infty \lambda^{a'+\frac{1}{2}-1} \exp\Big[-\lambda\big(b' + \frac{k'}{2}(\mu - m')^2\big)\Big]\,\mathrm{d}\lambda$$

after recognising this as the kernel of a Gamma distribution is easy to verify that the marginal posterior of μ is

$$\propto \Big[2\,b' + k'\,(\mu - m')^2\Big]^{-\frac{2a'+1}{2}}$$

rearranging terms again yields

$$\propto \Big[1 + \frac{1}{2a'}\frac{k'\,a'}{b'}(\mu - m')^2\Big]^{-\frac{2a'+1}{2}}\,.$$

We can now recognise this last expression as the kernel of a Student distribution, $\mathrm{St}\left(\mu \mid \eta, \tau, \nu\right)$, with location parameter $\eta = m'$, scale parameter $\tau = b'/(k'\,a')$ and $\nu = 2a'$ degrees of freedom. Further, if $\nu > 1$, then $\mathrm{E}[\mu \mid \boldsymbol{x}] = m'$ and if $\nu > 2$, then $\mathrm{Var}[\mu \mid \boldsymbol{x}] = \tau\,\nu(\nu - 2)^{-1}$.

Example 3 (Continued).
Assume that our best guess about the location of the precision is 2, but we are not too sure about it, so we specify $a = 1$ and $b = 1/2$, and that this information is roughly based in one observation, i.e. $k = 1$. We believe a priori that $m = -2$.

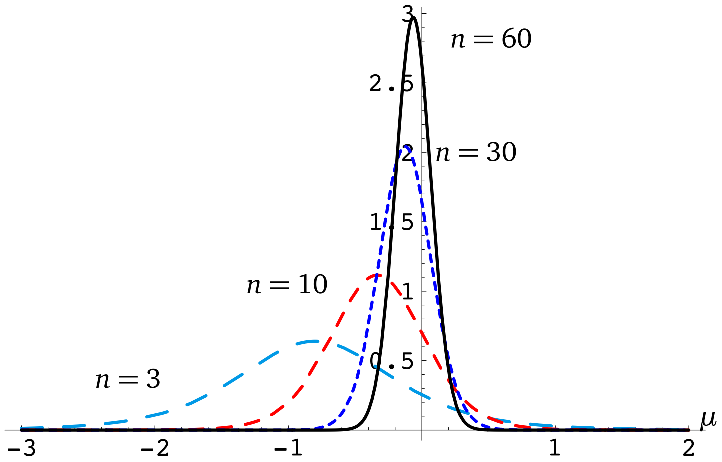


Figure 1.7. *Marginal posterior pdf's of the mean for different sample sizes.*

Remember that we have obtained $\bar{x} = 0$ for all sample sizes and, now we calculated the other sufficient statistic from the sample as $s^2 = 1$ for all sample sizes. Figure 1.7 depicts the four posterior distributions corresponding to $n = 3, 10, 30, 60$. Again we can check how the influence of the prior on the posterior gradually fades out as more sample information becomes available. $\triangleleft$

APPENDIX A

PROBABILITY DENSITY AND MASS FUNCTIONS

With slight abuse of notation, we shall use the same notation for distributions and their density or mass functions; e.g. if the observable x follows a Normal distribution with mean μ and precision λ, its density is denoted by $\mathrm{N}\left(x \mid \mu, \lambda^{-1}\right)$ and we shall also state that $x \sim \mathrm{N}\left(x \mid \mu, \lambda^{-1}\right)$. Table A.1 presents the density and mass functions used throughout the notes.

Table A.1. *Density functions utilised throughout the notes.*

Bernoulli	$\{\mathrm{Ber}(x \mid \theta), \quad x = 0, 1, \quad 0 < \theta < 1\}$	$\mathrm{Ber}(x \mid \theta) = \theta^x (1-\theta)^{1-x}$
Beta	$\{\mathrm{Be}\left(x \mid \alpha, \beta\right), \quad 0 < x < 1, \quad \alpha > 0, \beta > 0\}$	$\mathrm{Be}\left(x \mid \alpha, \beta\right) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Binomial	$\{\mathrm{Bi}(x \mid n, \theta), \quad x \in \{0, \ldots, n\}, n \in \mathbb{N}, \quad 0 < \theta < 1\}$	$\mathrm{Bi}(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$
Exponential	$\{\mathrm{Ex}(x \mid \theta), \quad x > 0, \quad \theta > 0\}$	$\mathrm{Ex}(x \mid \theta) = \theta\, e^{-\theta x}$
Gamma	$\{\mathrm{Ga}\left(x \mid \alpha, \beta\right), \quad x > 0, \quad \alpha > 0, \beta > 0\}$	$\mathrm{Ga}\left(x \mid \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

Table A.1. *Continued*

Normal	$\{\mathrm{N}\left(x \mid \mu, \sigma^2\right) \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \sigma^2 > 0\}$	$\mathrm{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right]$
Multivariate Normal	$\left\{\mathrm{N}_p\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma\right), \quad \boldsymbol{x} \in \mathbb{R}^p, \quad \boldsymbol{\mu} \in \mathbb{R}^p, \Sigma \text{ symmetric, positive definite}\right\}$	$\mathrm{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma\right) = \lvert\Sigma\rvert^{-1/2}(2\pi)^{-p/2} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]$
Poisson	$\{\mathrm{Po}\left(x \mid \lambda\right), \quad x \in \mathbb{N}, \quad \lambda > 0\}$	$\mathrm{Po}\left(x \mid \lambda\right) = e^{-\lambda}\,\frac{\lambda^x}{x!}$
Uniform	$\{\mathrm{Un}\left(x \mid \alpha, \beta\right), \quad \alpha \le x \le \beta, \quad \alpha < \beta \in \mathbb{R}\}$	$\mathrm{Un}\left(x \mid \alpha, \beta\right) = \frac{1}{\beta-\alpha}$
Student	$\{\mathrm{St}\left(x \mid \mu, \lambda, \alpha\right), \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \lambda > 0, \alpha > 0\}$	$\mathrm{St}\left(x \mid \mu, \lambda, \alpha\right) = \frac{\Gamma[(\alpha+1)/2]}{\Gamma(\alpha/2)}\left(\frac{1}{\lambda\alpha\pi}\right)^{\frac{1}{2}}\left[1+\frac{1}{\lambda\alpha}\left(x-\mu\right)^2\right]^{-\frac{\alpha+1}{2}}$