1 Introduction to Bayesian Statistics

- "Bayesian Statistics" is another school of thought/theory of drawing statistical inference. The idea originated from Rev. Thomas Bayes, who lived from 1702 to 1761.

- "Frequentist"/classical approach is what you have been studying or using in your previous statistical courses. It relies on the maximum likelihood (ML) estimation.

- Basic set-up of a statistical problem (regardless of your approach): suppose θ (maybe vector valued) is of inferential interest. The data x coming from a parametric model (of known form) summarize information about θ. The model density is denoted by $f(x|\theta)$.

- Bayesian approach/framework is different from the classical approach in several aspects:

	FREQUENTIST	BAYESIAN
INFERENCE	totally objective (depend only on data)	*subjective* + objective (both prior belief and data)
PARAMETER	fixed (i.e., a constant)	random (treated as a random variable)
DATA	random (different from time to time)	fixed (treated as constants once observed)
METHOD	maximum likelihood (ML)	Bayes' Theorem
ESTIMATOR	MLE	*posterior distribution*

- Some terms or notion to be used in Bayesian statistics:

	MEANING	NOTATION
Prior density	random behaviour followed by the unknown parameter	$\pi(\theta)$
likelihood function (model density)	underlying model of an outcome, x, of random behaviour/experiment relating x with θ	$f(x|\theta)$
Posterior density	"updated" random behaviour followed by the unknown parameter in the light of the observed data x	$\pi(\theta|x)$

- Bayesian formulation:

$$\boxed{\text{Posterior} \propto \text{Prior} \times \text{Likelihood}}$$

or in notation,

$$\boxed{\pi(\theta|x) \propto \pi(\theta) \times f(x|\theta)}. \tag{1}$$

2 Conditional Probability and Bayes' Theorem

2.1 Conditional Probability

Let H and A_i's be events of an experiment, Ω be the sample space.

- Axioms of conditional probability:

  1. $0 \leq \Pr(A|H) \leq 1$.
  2. $P(H|H) = 1$.
  3. **Area rule**: if A_1 and A_2 are disjoint,
  $$P(A_1 \cup A_2|H) = P(A_1|H) + P(A_2|H).$$
  4. **Product rule**:
  $$P(A_1 \cap A_2|H) = P(A_1|H)\, P(A_2|A_1 \cap H).$$

- Consequences:

  1. If $H \subseteq A$, $P(A|H) = 1$.
  2. The area rule is valid for events $A_1, \ldots, A_k$ provided that the k events are disjoint.
  $$P(A_1 \cup \cdots \cup A_k|H) = P(A_1|H) + \cdots + P(A_k|H).$$
  3. The product rule is valid for k events.
  $$P(A_1 \cap \cdots \cap A_k|H) = P(A_1|H)\, P(A_2|A_1 \cap H) \cdots P(A_k|A_1 \cap \cdots \cap A_{k-1} \cap H).$$

2.2 Bayes' Theorem

- **A partition of the sample space**: $A_1, \ldots, A_k$ form a partition of the sample space Ω if they are (i) mutually exclusive (i.e., disjoint) and (ii) exhaustive.

- **An expansion with respect to a partition**: Given a partition, $A_1, \ldots, A_k$, for any event B,

$$P(B) = P(A_1)P(B|A_1) + \cdots + P(A_k)P(B|A_k). \tag{2}$$

Example 2.1 *Suppose I need to assign final grades for this course, what is a partition of the sample space?*

Example 2.2 *A box has six numbered tickets {1,2,2,3,3,3}. Draw twice from the box. What is the chance that the second draw is a "2" if the drawings are done without replacement?*

Theorem 2.1 (Bayes' Theorem) *Suppose* $\{A_1, \ldots, A_k\}$ *is a partition with known probabilities* $P(A_i)$*'s. Then, for any event* B,

$$P(A_j|B) \propto P(A_j) P(B|A_j) \qquad j = 1, \ldots, k. \tag{3}$$

By the area rule, the normalization constant is necessarily $1/P(B)$ *where* $P(B)$ *is given by (2).*

- It is customarily to call $P(A_i)$'s prior probabilities of the partition, and $P(A_i|B)$'s posterior probabilities of the partition.

Example 2.3 *Three bags of apples. Bag I has 10% bad apples, Bag II has 20% bad apples, and Bag III has 40% bad apples. (i) Suppose you select one of the three bags (at random), and pick an apple from the selected bag. The apple turned out to be bad, what is the chance that it is from Bag I? (ii) Suppose an apple picked is good, what is the chance that it is from Bag I?*

2.3 Conditional Random Variables

Suppose X and Y are two random variables.

- One defines the conditional random variable of Y given $X = x$, by the **conditional distribution function**

  $$F(y|x) \equiv P(Y \leq y|X = x).$$

  If given $X = x$,

  (i) Y is discrete, then the **(conditional) density / probability mass function (p.m.f)** of Y given $X = x$ is denoted by $\pi(y|x) \equiv P(Y = y|X = x)$.

(ii) Y is continuous, then the **(conditional) density / probability density function (p.d.f.)** of Y given $X = x$ is denoted by $\pi(y|x) \equiv \frac{d}{dy}F(y|x)$.

- The conditional density $\pi(y|x)$ is different from $\pi(y)$ in general. When they are equal, we say that X and Y are **independent**.

Example 2.4 *Suppose the height of the population in a country has a mean of 66" and an SD of 5". Among children less than 12 years old, the height has probably a (conditional) mean of 42" and a (conditional) SD of 4".*

- One can evaluate conditional probabilities of Y given $X = x$ using the conditional density based on the same principles.

- Given $X = x$, the **conditional mean and variance** of Y are

$$E\left(Y|X=x\right) = \begin{cases} \sum_{\text{all } y} y \times \pi(y|x) & \text{if } Y|X=x \text{ is discrete} \\ \int y\pi(y|x)dy & \text{if } Y|X=x \text{ is continuous} \end{cases},$$

and

$$Var\left(Y|X=x\right) = \begin{cases} \sum_{\text{all } y} \left[y - E\left(Y|X=x\right)\right]^2 \pi(y|x) & \text{if } Y|X=x \text{ is discrete} \\ \int \left[y - E\left(Y|X=x\right)\right]^2 \pi(y|x)dy & \text{if } Y|X=x \text{ is continuous} \end{cases},$$

respectively. These are just the usual formulae of $E\left(Y\right)$ and $Var\left(Y\right)$ using the conditional density as basis of computations.

- Suppose the density for X is $\pi(x)$, the **product rule** gives the **joint density** of X and Y to be

$$\pi(x,y) \equiv \pi(x)\pi(y|x), \tag{4}$$

or similarly, $\pi(y)$ is the density for Y,

$$\pi(x,y) \equiv \pi(y)\pi(x|y). \tag{5}$$

Theorem 2.2 (Bayes' Theorem for two random variables) *Suppose $\pi(x)$ and $\pi(y|x)$ are known. According to (4) and (5), the conditional density of X given $Y = y$ is*

$$\pi(x|y) \propto \pi(x)\pi(y|x), \tag{6}$$

where the normalization constant is given by the ***marginal density*** *of y,*

$$\pi\left(y\right) = \begin{cases} \displaystyle\int \pi(x)\pi(y|x)dx & \textit{if } X \textit{ is continuous} \\ \sum_{\textit{all } x} \pi(x)\pi(y|x) & \textit{if } X \textit{ is discrete} \end{cases}.$$

- The normalization constant depends on y such that $\pi(x|y)$ is a proper density.

- Combining with (4), one has

$$\pi(x|y) \propto \pi(x,y), \tag{7}$$

  that is, one is able to get the **kernel** of the conditional density of $\pi(x|y)$ by treating $\pi(x,y)$ as a density function in x. Once getting the kernel, it remains to evaluate the normalization constant to identify the conditional density $\pi(x|y)$.

- The two conditional densities determine the marginal density (up to a normalization constant) by inspection since

$$\pi(y) \propto \frac{\pi(y|x)}{\pi(x|y)} \tag{8}$$

  is a density in y.

Example 2.5 *(i) Suppose X is $U(0,1)$. $Y|X$ has a Bin(n,X) distribution (n is known). Then, the conditional density of X given $Y=y$ is*

(ii) Suppose X has a Gamma$(\alpha,1/\beta)$ distribution. $Y|X$ has a Poisson(X) distribution. Then,

Example 2.6 *Suppose τ is a Gamma$(\alpha,1/\beta)$ random variable and given τ, μ is a $N(m,1/(\tau t))$ random variable where t is known. Find (i) the conditional distribution of $\tau|\mu$ and (ii) the marginal distribution of μ.*

2.4 Extensions to Several Random Variables

Suppose there are $k+1$ random variables, $Y, X_1, \ldots, X_k$.

- One can always define the conditional distribution of Y given $X_1 = x_1, \ldots, X_k = x_k$ by

  $$F(y|x_1, \ldots, x_k) \equiv P(Y \leq y | X_1 = x_1, \ldots, X_k = x_k),$$

  and also the conditional density $\pi(y|x_1, \ldots, x_k)$.

- The product rule gives the joint density of $Y, X_1, \ldots, X_k$ as

  $$\pi(y, x_1, \ldots, x_k) = \pi(y)\,\pi(x_1|y) \cdots \pi(x_k|y, x_1, \ldots, x_{k-1}). \tag{9}$$

Example 2.7 *Let $\alpha_1, \ldots, \alpha_k > 0$. Define $\alpha_{i+} = \alpha_i + \cdots + \alpha_k$. Suppose $X_1 \sim Beta(\alpha_1, \alpha_{2+})$, $\left(\frac{X_2}{1-X_1}|X_1\right) \sim Beta(\alpha_2, \alpha_{3+}), \ldots, \left(\frac{X_{k-1}}{1-X_1-\cdots-X_{k-2}}|X_1, \ldots, X_{k-2}\right) \sim Beta(\alpha_{k-1}, \alpha_{k+})$ and $\left(\frac{X_k}{1-X_1-\cdots-X_{k-1}}|X_1, \ldots, X_{k-1}\right)$ is always equal to 1. What is the joint distribution of $(X_1, ..., X_k)$?*

- Suppose we are interested in a random variable Y with a density $\pi(y)$. Random variables, say, $X_1, \ldots, X_k$, that convey information about Y comes in one at time (and in the order written). Then the conditional distribution of Y given $X_1 = x_1, \ldots, X_k = x_k$ is summarized as follows:

Theorem 2.3 (Bayes' Theorem for several random variables) *Suppose* $Y, X_1, \ldots, X_k$ *are* $k+1$ *random variables. Then, the conditional density of* Y *given* $X_1 = x_1, \ldots, X_k = x_k$ *is given by*

$$\pi\left(y|x_1, \ldots, x_k\right) \propto \pi\left(y, x_1, \ldots, x_k\right).$$

- A streamline proof:

  Step 1 The Bayes' Theorem for two random variables combines $\pi\left(y\right)$ and $\pi\left(x_1|y\right)$ to give

  $$\pi\left(y|x_1\right) \propto \pi\left(y\right)\pi\left(x_1|y\right).$$

  Step 2 Given $X_1 = x_1$, the Bayes' Theorem combines $\pi\left(y|x_1\right)$ and $\pi\left(x_2|y, x_1\right)$ to give

  $$\pi\left(y|x_1, x_2\right) \propto \pi\left(y|x_1\right)\pi\left(x_2|y, x_1\right).$$

  $\vdots$

  Step k Given $X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1}$, the Bayes' Theorem combines $\pi\left(y|x_1, \ldots, x_{k-1}\right)$ and $\pi\left(x_k|y, x_1, \ldots, x_{k-1}\right)$ to give

  $$\pi\left(y|x_1, \ldots, x_k\right) \propto \pi\left(y|x_1, \ldots, x_{k-1}\right)\pi\left(x_k|y, x_1, \ldots, x_{k-1}\right).$$

- The above kind of sequential updating procedure of the density of the interested random variable Y is referred to as **Bayesian sequential updating**.

Example 2.8 *Suppose* $(X_1, ..., X_k)$ *has a Multinomial distribution with* n *number of trials and nonnegative probabilities of success* $p_1, ..., p_k$*, where* $\sum_{i=1}^{k} p_i = 1$*. Given* $X_1, \ldots, X_{k-2}$*, what is the distribution of* X_{k-1}*?*

3 Overview of Bayesian Inference

We will first get an overview of the difference between Bayesian approach and the frequentist approach. Suppose we are interested in a random quantity X which has a model density denoted by $\pi(X|\theta)$ parametrized by θ. Suppose we observe n iid observations from X, denoted by $X_1, \ldots, X_n$. To make inference on θ, we have

The frequentist approach:

- By the maximum likelihood (ML) principle, one uses MLE, denoted by $\widehat{\theta} = f(X_1, \ldots, X_n)$, which is the value of θ that maximizes the likelihood function $\prod_{i=1}^{n} \pi(X_i|\theta)$, to estimate θ.
- The variability of $\widehat{\theta}$ depends on the variance of $\widehat{\theta} = f(X_1, \ldots, X_n)$ through the probability distribution of $f(X_1, \ldots, X_n)$.
- A $100(1-\alpha)\,\%$ confidence interval tells us that if we repeatedly get realizations of n observations from $\pi(X|\theta)$ a large number of time and calculate the corresponding $100(1-\alpha)\,\%$ confidence intervals by the same formula, then approximately $100(1-\alpha)\,\%$ of the intervals will include the true parameter θ.

The Bayesian Approach:

- Assume the unknown parameter has a **"prior" density** $\pi(\theta)$ before any data come in.
- Update the "prior" to get the **"posterior" density** by the Bayes' Theorem,

$$\pi(\theta|X_1, \ldots, X_n) \propto \pi(\theta) \times \prod_{i=1}^{n} \pi(X_i|\theta).$$

- Estimate θ by the **posterior mean** $E[\theta|X_1, \ldots, X_n]$.
- The **posterior variance** $Var[\theta|X_1, \ldots, X_n]$, which is usually readily obtained, measures how variable θ is.
- One can obtain a $100(1-\alpha)\,\%$ **credible set** / **highest density region (HDR)** easily from the posterior distribution of θ. It conveys the information that there is a $100(1-\alpha)\,\%$ chance that this interval contains the true parameter θ.

Remark 3.1 *Bayesian inference often contains the results of frequentist inference as a special case by the action of "**deflating**" the prior via choosing an "**improper**" density as the prior. Improper density here means that it is NOT necessarily integrated out to 1. It is interesting that **improper** or **flat** prior often gives a posterior distribution which is well-defined.*

4 Bayesian Inference for a Normal Population

Suppose we are interested in a random quantity X which follows $N\left(\mu, \sigma^2\right)$. Reparametrize σ^2 by τ such that $\tau = \frac{1}{\sigma^2}$. We say τ is a precision parameter of a Normal random variable. In this course, we will follow this parametrization of Normal random variables.

4.1 Normal Population with Known Variance

Assume $\tau = r$ is known. Then, it suffices to estimate the unknown mean μ. Suppose we observe n iid observations from X, denoted by $\mathbf{x} = (x_1, \ldots, x_n)$.

The frequentist results:

Proposition 4.1 (Normal population with known variance) *Suppose* μ *has a* $N\left(m, \frac{1}{t}\right)$ *prior. Given iid observations* $\mathbf{x} = (x_1, \ldots, x_n)$ *from* $X \sim N\left(\mu, \frac{1}{r}\right)$ *where* r *is known. Then,*

$$(\mu|\mathbf{x}) \sim N\left(m_n, \frac{1}{t_n}\right), \tag{10}$$

where

$$t_n = t + nr \qquad \textit{and} \qquad m_n = \left(\frac{t}{nr+t}\right) m + \left(\frac{nr}{nr+t}\right) \overline{x}. \tag{11}$$

- The posterior distribution of μ depends on the data through their average. This function of the observations (statistic) is the same as the **sufficient statistic** for μ.
- The population mean μ is estimated by the posterior mean of μ,

  $$E\left[\mu|\mathbf{x}\right] = m_n = w_n m + (1 - w_n)\overline{x},$$

  which is a **weighted average** of the prior mean m and the MLE, sample mean of the data $\overline{x}$. The weighting factor $w_n = \dfrac{t}{nr+t}$ is determined by the relative strength of information between the prior and the data. If nr is large relative to t, then w_n is small and the posterior mean is close to $\overline{x}$.
- While deflating the prior (i.e., letting $t \to 0$), μ is Normal with variance ∞, then
  - the posterior mean m_n converges to the MLE, sample average $\overline{x}$, and
  - the posterior variance $\dfrac{1}{t_n}$ converges to $\dfrac{1}{nr}$ which is the standard error of the MLE.
- If we are getting more and more data such that $nr >> t$,

  $$(\mu|\mathbf{x}) \sim N\left(\overline{x}, \frac{1}{nr}\right)$$

  in which the prior has no effect.

- A streamline proof by the identity,

$$r(\mu - a)^2 + t(\mu - b)^2 = (\mu - \overline{m})^2(t + r) + (a - b)^2 \frac{1}{t^{-1} + r^{-1}} \tag{12}$$

where $\overline{m} = \frac{r}{t+r} a + \frac{t}{t+r} b$.

Example 4.1 *Suppose the average height, θ, of a Normal population is of interest. It is known that variance of heights of this population is 1. Observe 10 observations: 14.5, 15.1, 15.3, 15.5, 16.3, 16.5, 17.3, 17.3, 17.6, 18.0. Find the posterior distribution of θ and the posterior probability of $\theta \geq 15$ if (i) θ is $N(0, 1)$; (ii) θ is exponential with mean 16.*

4.2 Normal Population with Known Mean

Assume $\mu = h$ is known. Then, it suffices to estimate the unknown precision τ. Suppose we observe n iid observations from X, denoted by $\mathbf{x} = (x_1, \ldots, x_n)$.

The frequentist results:

Proposition 4.2 (Normal population with known mean) *Suppose τ has a Gamma $\left(\alpha, \frac{1}{\beta}\right)$ prior. Given iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from $X \sim N\left(h, \frac{1}{\tau}\right)$ where h is known. Then,*

$$(\tau|\mathbf{x}) \sim Gamma\left(\alpha_n, \frac{1}{\beta_n}\right) \tag{13}$$

where

$$\alpha_n = \alpha + \frac{n}{2} \qquad \textit{and} \qquad \beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - h)^2. \tag{14}$$

- The posterior distribution of τ depends on the data through the sufficient statistic $\sum_{i=1}^{n}(x_i - h)^2$.
- The population precision τ is estimated by the posterior mean of τ,

  $$E[\tau|\mathbf{x}] = \frac{\alpha_n}{\beta_n} = \frac{\alpha + \frac{n}{2}}{\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - h)^2},$$

  which is a weighted average of the prior mean and the MLE,

  $$w_n \times \frac{\alpha}{\beta} + (1 - w_n) \times \frac{n}{\sum_{i=1}^{n}(x_i - h)^2},$$

  with a weighting factor $w_n = \beta / \left[\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - h)^2\right]$.
- When deflating the prior (i.e., letting $\alpha, \beta \to 0$), the posterior mean converges to $\frac{n}{\sum_{i=1}^{n}(x_i - h)^2}$, that is, the population variance is estimated by the MLE, $\frac{\sum_{i=1}^{n}(x_i - h)^2}{n}$, as a result of an application of the **Taylor's series expansion**:

  A Taylor's series expansion of $g(X)$ (a given function of any random variable X) about a (any constant) is given by

  $$g(X) = g(a) + g'(a)(X - a) + \frac{g''(a)(X - a)^2}{2!} + \cdots,$$

  by assuming the existences of all derivatives of order r of the function $g(x)$, denoted by $g^r(x) = \frac{d^r}{dx^r}g(x)$. Hence, if $g(X) = 1/X$ and $a = E[X]$, we have

  $$E\left[\frac{1}{X}\right] \approx \frac{1}{E[X]}$$

  by taking expectation at both sides in the above expansion and assuming that the higher-order terms tend to zero.

4.3 Normal Population with both Mean and Variance Unknown

Suppose we observe n iid observations, denoted by $x_1, \ldots, x_n$, from $X \sim N\left(\mu, \frac{1}{\tau}\right)$ in which both μ and τ are unknown. Suppose (μ, τ) is

$$Gamma - Normal\left(\alpha, \frac{1}{\beta}; m, \frac{1}{t}\right) \tag{15}$$

such that $\pi(\mu, \tau) = \pi(\tau)\, \pi(\mu|\tau)$ where

$$\tau \text{ is } Gamma\left(\alpha, \frac{1}{\beta}\right) \qquad \text{and} \qquad (\mu|\tau) \text{ is } N\left(m, \frac{1}{\tau t}\right). \tag{16}$$

Consider only one observation x_1, the joint posterior density of (μ, τ) based on x_1 is

$$\begin{aligned}\pi(\mu, \tau|x_1) &\propto \tau^{1/2} \exp\left[-\frac{\tau}{2}(\mu - x_1)^2\right] \pi(\mu, \tau) \\ &\propto \tau^{1/2} \exp\left[-\frac{\tau}{2}(\mu - x_1)^2\right] \times \tau^{\alpha-1} \exp(-\beta\tau) \times \tau^{1/2} \exp\left[-\frac{\tau t}{2}(\mu - m)^2\right] \mathbf{I}_{\{\tau>0, \mu\in\mathbb{R}\}} \\ &= \tau^{\alpha} \exp\left\{-\tau\left[\beta + \frac{1}{2}(\mu - x_1)^2 + \frac{t}{2}(\mu - m)^2\right]\right\} \mathbf{I}_{\{\tau>0, \mu\in\mathbb{R}\}}.\end{aligned}$$

An application of the identity (12) yields $(\mu, \tau|x_1)$ is $Gamma - Normal\left(\alpha_1, \frac{1}{\beta_1}; m_1, \frac{1}{t_1}\right)$ where

$$\alpha_1 = \alpha + \frac{1}{2}, \quad \beta_1 = \beta + \frac{1}{2}\frac{(m - x_1)^2}{(t^{-1} + 1)}, \quad t_1 = t + 1, \text{ and } m_1 = \left(\frac{t}{t+1}\right) m + \left(\frac{1}{t+1}\right) x_1.$$

Following the essence of the Bayesian sequential updating technique, the posterior distribution of $(\mu, \tau|x_1, x_2)$ can be obtained symmetrically by treating $\pi(\mu, \tau|x_1)$ as the prior $\pi(\mu, \tau)$ (i.e., replacing the **hyperparameters** (α, β, m, t) by $(\alpha_1, \beta_1, m_1, t_1)$ in the above arguments). That is, $(\mu, \tau|x_1, x_2)$ is $Gamma - Normal\left(\alpha_2, \frac{1}{\beta_2}; m_2, \frac{1}{t_2}\right)$ where

$$\alpha_2 = \alpha_1 + \frac{1}{2}, \quad \beta_2 = \beta_1 + \frac{1}{2}\frac{(m_1 - x_2)^2}{\left(t_1^{-1} + 1\right)}, \quad t_2 = t_1 + 1, \text{ and } m_2 = \left(\frac{t_1}{t_1+1}\right) m_1 + \left(\frac{1}{t_1+1}\right) x_2.$$

Applications of this recursive argument of another $n - 2$ times yield the first result in the coming proposition.

Proposition 4.3 (Normal population with both mean and variance unknown) *Suppose* (μ, τ) *is* $Gamma - Normal\left(\alpha, \frac{1}{\beta}; m, \frac{1}{t}\right)$. *Given* n *iid observations* $\mathbf{x} = (x_1, \ldots, x_n)$ *from a population* $X \sim N\left(\mu, \frac{1}{\tau}\right)$, *Then,*

(i) $(\mu, \tau|\mathbf{x})$ *is* $Gamma - Normal\left(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n}\right)$, *where*

$$\begin{aligned}\alpha_n &= \alpha + \frac{n}{2}, \qquad \beta_n = \beta + \frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{(m - \overline{x})^2}{(t^{-1} + n^{-1})}\right], \\ t_n &= t + n, \qquad \textit{and} \qquad m_n = \left(\frac{t}{t+n}\right) m + \left(\frac{n}{t+n}\right) \overline{x}.\end{aligned} \tag{17}$$

(ii) $(\mu|\mathbf{x})$ *has a t-distribution with* $2\alpha_n$ *degrees of freedom, location parameter* m_n *and precision parameter* $\frac{\alpha_n}{\beta_n}t_n$*, denoted by* $t_{2\alpha_n}\left(m_n, \left(\frac{\alpha_n}{\beta_n}t_n\right)^{-1}\right)$*, that is,*

$$\pi(\mu|\mathbf{x}) \propto \left[1 + \frac{1}{2\alpha_n}\left(\frac{\alpha_n}{\beta_n}t_n\right)(\mu - m_n)^2\right]^{-(2\alpha_n+1)/2} \mathbf{I}_{\{\mu\in\mathbb{R}\}} \tag{18}$$

with the proportionality constant $\frac{\Gamma\left(\frac{2\alpha_n+1}{2}\right)}{\Gamma\left(\frac{2\alpha_n}{2}\right)}\sqrt{\frac{1}{2\alpha_n\pi}\left(\frac{\alpha_n}{\beta_n}t_n\right)}$.

Result in part (ii) is implied by example 2.6; first of all, notice that the set-up there fits in this situation perfectly. Here we are given

$$(\tau|\mathbf{x}) \text{ is } Gamma\left(\alpha_n, \frac{1}{\beta_n}\right) \qquad \text{and} \qquad (\mu|\tau, \mathbf{x}) \text{ is } N\left(m_n, \frac{1}{\tau t_n}\right).$$

Then, we should have

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto \left[\beta_n + \frac{t_n}{2}(\mu - m_n)^2\right]^{-\alpha_n+1/2} \mathbf{I}_{\{\mu\in\mathbb{R}\}} \\ &\propto \left[1 + \frac{1}{2\alpha_n}\left(\frac{\alpha_n}{\beta_n}t_n\right)(\mu - m_n)^2\right]^{-(2\alpha_n+1)/2} \mathbf{I}_{\{\mu\in\mathbb{R}\}}.\end{aligned}$$

If one defines $\upsilon = \sqrt{\frac{\alpha_n}{\beta_n}t_n}(\mu - m_n)$, that is, $\mu = \sqrt{\frac{\beta_n}{\alpha_n t_n}}\upsilon + m_n$, then it has a posterior density

$$\pi(\upsilon|\mathbf{x}) \propto \left[1 + \frac{\upsilon^2}{2\alpha_n}\right]^{-(2\alpha_n+1)/2} \mathbf{I}_{\{\upsilon\in\mathbb{R}\}}, \tag{19}$$

and hence, υ is called a Student's t distribution on $2\alpha_n$ degrees of freedom (usually denoted by $t_{2\alpha_n}$ in standard textbooks), with

$$E[\upsilon|\mathbf{x}] = 0 \qquad \text{and} \qquad Var[\upsilon|\mathbf{x}] = \frac{2\alpha_n}{2\alpha_n - 2}. \tag{20}$$

The proportionality constant in (19) is given by $\frac{\Gamma\left(\frac{2\alpha_n+1}{2}\right)}{\Gamma\left(\frac{2\alpha_n}{2}\right)}\sqrt{\frac{1}{2\alpha_n\pi}}$ (mean, variance and the proportional constant can be checked out in standard textbooks). Standard transformation method gives the proportionality constant of (18) as

$$\frac{\Gamma\left(\frac{2\alpha_n+1}{2}\right)}{\Gamma\left(\frac{2\alpha_n}{2}\right)}\sqrt{\frac{1}{2\alpha_n\pi}\left(\frac{\alpha_n}{\beta_n}t_n\right)}.$$

Furthermore, the posterior mean and posterior variance of μ are

$$E[\mu|\mathbf{x}] = m_n \qquad \text{and} \qquad Var[\mu|\mathbf{x}] = \left(\frac{\alpha_n}{\beta_n}t_n\right)^{-1}\left(\frac{\alpha_n}{\alpha_n - 1}\right) = \frac{\beta_n}{t_n(\alpha_n - 1)}. \tag{21}$$

Remark 4.1 *The precision parameter above is NOT equal to the reciprocal of the variance as in Normal case.*

- The posterior distribution of (μ, τ) depends on the data through the sufficient statistic $\left(\overline{x}, \sum_{i=1}^{n}(x_i - \overline{x})^2\right)$.

- The population precision τ is estimated by the posterior mean of τ,

$$E[\tau|\mathbf{x}] = \frac{\alpha_n}{\beta_n} = \frac{\alpha + \frac{n}{2}}{\beta + \frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{(m-\overline{x})^2}{(t^{-1}+n^{-1})}\right]} \overset{\alpha,\beta,t\to 0}{\longrightarrow} \frac{n}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

  Hence when deflating the prior (letting $\alpha, \beta, t \to 0$), the population variance is estimated by the MLE, $\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ due to the Taylor's series expansion.

- The population mean μ is estimated by the posterior mean of μ,

$$E[\mu|\mathbf{x}] = m_n = \left(\frac{t}{t+n}\right)m + \left(\frac{n}{t+n}\right)\overline{x} \overset{t\to 0}{\longrightarrow} \overline{x},$$

  where sample average $\overline{x}$ is the MLE.

- The Bayes estimates for (μ, τ) coincide with the MLEs when the sample size n is large.

Remark 4.2 *One can work out the following assumed prior properties on μ:*

$$E[\mu] = m \qquad \textit{and} \qquad Var[\mu] = \frac{\beta}{t(\alpha - 1)}.$$

Hence, it is possible to reflect one's own belief on μ via adjusting the hyperparameters.

Example 4.2 *Suppose that the results of a certain test are known to be approximately $N(\mu, 1/\tau)$. Suppose further that your prior belief about (μ, τ) is $Gamma - Normal\,(1, 0.5; 74, 2/3)$. Next, 36 observations are obtained from the population with sample mean 82 and sample variance $s^2 = 27$. Find the posterior distribution of (μ, τ). Find 90% prior and posterior intervals for μ.*

5 Conjugate Prior Distributions

After studying the Bayesian inference for the Normal models, it seems that the major obstacle or difficulty is to updating the hyperparameters; the resulted posterior distributions are of known and standard parametric families, Normal or Gamma, and hence any probabilistic statement or statistical inference can be made based on techniques learnt in elementary courses. Nevertheless,

Nice posterior densities are NOT to be taken for granted!

One might end up with some "not-so-familar", or even intractable, probability distributions if s/he blindly chose her/his prior distribution for the parameter.

Example 5.1 *Consider the Normal model with known variance, if one assume a Student's t distribution with* 40 *degrees of freedom for the prior of* μ, *one will end up with a posterior density having a kernel which is so complicated that no one can recognize the posterior density to be from any known parametric family.*

To avoid this from happening, the notion or concept of **conjugate family** is developed.

Definition 5.1 (conjugate family) *Let* $\{f(x|\theta) : \theta \in \Theta\}$ *be a family of distributions of interest. A class* Π *of probability distributions is said to form a conjugate family if the posterior density*

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$$

is in the class Π *for all* x *whenever* $\pi(\theta)$ *is in* Π.

Remark 5.1 *It is sensible to have elements/probability distributions in* Π *being defined on the parameter space* Θ, *or at least including the whole parameter space. Due to the Bayes' Theorem,*

$$\pi(\theta|x) \propto \pi(\theta, x) = \pi(\theta) f(x|\theta),$$

if $\pi(\theta)$ *has zero probability in subset* $\Theta_1 \subset \Theta$, *the posterior distribution can NEVER have positive probabilities at points* $x \in \Theta_1$ *even though* $f(x|\theta)$ *is positive.*

Example 5.2 *(i) According to proposition* 4.1, the **Normal family** of distributions is a conjugate family of prior distributions for the mean of a Normal population with known variance.

(ii) According to proposition 4.2, the **Gamma family** of distributions is a conjugate family of prior distributions for the precision of a Normal population with known mean.

(iii) According to example 4.1, the Exponential family of distributions is NOT a conjugate family of prior distributions for the mean of a Normal population with known variance since the resulting posterior distribution is a truncated Normal distribution but not an exponential distribution.

A few more examples of conjugate families for different populations are discussed below. Throughout we assume n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ are observed from the interested population X.

5.1 Bernoulli distributions

Suppose we are interested in a binary outcome of a Bernoulli experiment, in which $0 < \theta < 1$ is the probability of success. To get the conjugate family for θ as a **Beta family**, one observe that the model density,

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1,$$

looks like the kernel of a Beta density of a random variable θ by viewing it as a density in θ. Hence, if we choose $\pi(\theta)$ to be a $Beta(a,b)$ density, the posterior density of θ will be another Beta density:

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \pi(\theta) \prod_{i=1}^{n} \left[\theta^{x_i} (1-\theta)^{1-x_i}\right] \\
&\propto \theta^{a-1} (1-\theta)^{b-1} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \mathbf{I}_{\{0<\theta<1\}} \\
&= \theta^{\sum_{i=1}^n x_i + a - 1} (1-\theta)^{n-\sum_{i=1}^n x_i + b - 1} \mathbf{I}_{\{0<\theta<1\}}.
\end{aligned}$$

That is, $(\theta|\mathbf{x})$ is $Beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)$.

Example 5.3 *We would like to estimate the probability of getting a head, θ, in a toss of a given coin. By inspection, we summarize the prior opinion of θ to be a Beta random variable with mean 0.55 and an SD of 0.04. Suppose we flip the coin 100 times and observe 52 heads (and 48 tails). In view of this new evidence, what is the posterior distribution of θ, and what is the posterior probability that θ is between 0.50 ± 0.05? (A standardized Beta random variable is roughly $N(0,1)$ by the Central Limit Theorem.)*

5.2 Poisson distributions

Suppose we are interested in a Poisson population, in which $\lambda > 0$ is the mean or intensity. To get the conjugate family for λ as a **Gamma family**, one observe that the model density,

$$f\left(x|\lambda\right) = \lambda^x \exp\left(-\lambda\right)/x!, \qquad x = 0, 1, \ldots,$$

looks like the kernel of a Gamma density of a random variable λ by viewing it as a density in λ. Hence, if we choose $\pi\left(\lambda\right)$ to be a *Gamma* $\left(\alpha, \frac{1}{\beta}\right)$ density, the posterior density of λ will be another Gamma density:

$$\begin{aligned}\pi\left(\lambda|\mathbf{x}\right) &\propto \pi\left(\lambda\right)\prod_{i=1}^{n}\left[\lambda^{x_i}\exp\left(-\lambda\right)\right]\\ &\propto \lambda^{\alpha-1}\exp\left(-\beta\lambda\right) \times \lambda^{\sum_{i=1}^{n} x_i}\exp\left(-n\lambda\right)\mathbf{I}_{\{\lambda>0\}}\\ &= \lambda^{\sum_{i=1}^{n} x_i+\alpha-1}\exp\left[-\left(n+\beta\right)\lambda\right]\mathbf{I}_{\{\lambda>0\}}.\end{aligned}$$

That is, $\left(\lambda|\mathbf{x}\right)$ is *Gamma* $\left(\sum_{i=1}^{n} x_i + \alpha, \dfrac{1}{n+\beta}\right)$.

Example 5.4 *Traffic accidents per week at a certain dangerous intersection has a Poisson*(λ) *distribution. Based on past data for the last ten years, the prior distribution of λ is Gamma with mean 4.4 and an SD of 0.4. In a period of 52 weeks, there were 257 accidents. In view of these data, what is the posterior density of λ, and what is the posterior probability that λ is greater than 5? (A standardized Gamma random variable is roughly $N(0,1)$.)*

5.3 Exponential distributions

Suppose we are interested in an Exponential lifetime, in which $1/\lambda > 0$ is the mean. To get the conjugate family for λ as a **Gamma family**, one observe that the model density,

$$f\left(x|\lambda\right) = \lambda\exp\left(-\lambda x\right), \qquad x > 0,$$

looks like the kernel of a Gamma density of a random variable λ by viewing it as a density in λ. Hence, if we choose $\pi\left(\lambda\right)$ to be a *Gamma* $\left(\alpha, \frac{1}{\beta}\right)$ density, the posterior density of λ will be another

Gamma density:

$$\begin{aligned}\pi(\lambda|\mathbf{x}) &\propto \pi(\lambda)\prod_{i=1}^{n}[\lambda\exp(-\lambda x_i)]\\ &\propto \lambda^{\alpha-1}\exp(-\beta\lambda)\times\lambda^{n}\exp\left(-\lambda\sum_{i=1}^{n}x_i\right)\mathbf{I}_{\{\lambda>0\}}\\ &= \lambda^{n+\alpha-1}\exp\left[-\left(\sum_{i=1}^{n}x_i+\beta\right)\lambda\right]\mathbf{I}_{\{\lambda>0\}}.\end{aligned}$$

That is, $(\lambda|\mathbf{x})$ is *Gamma* $\left(n+\alpha, \dfrac{1}{\sum_{i=1}^{n}x_i+\beta}\right)$.

Example 5.5 *The lifetime (in hours) of a 1000-hour-rated or 1-unit-rated (i.e., 1 unit = 1000 hours) light bulb has an Exponential distribution with mean $1/\lambda$. In the past, for several batches of light bulbs, a histogram of the mean lifetimes has an average of 0.95 units and an standard deviation of 0.021 units. Suppose the prior distribution for λ is Gamma$(\alpha, 1/\beta)$. (i) Find α and β. In an experiment conducted by the manufacturer, 50 such light bulbs were found to burn for a total of 46,000 hours. (ii) What is the posterior probability that the average lifetime of the bulb is less than 0.925 units.*

5.4 Uniform distributions

Suppose we are interested in $X \sim U(0,\theta)$. Viewing the model density

$$f(x|\theta)=\frac{1}{\theta}\mathbf{I}_{\{0<x<\theta\}}$$

as a density in θ suggests that the conjugate family for θ is a **Pareto family** with densities

$$\pi(\theta)=\frac{am^{a}}{\theta^{a+1}}\mathbf{I}_{\{\theta>m\}},$$

since $f(x|\theta)$ looks like the kernel of a Pareto density of a random variable θ. Hence, if we choose $\pi(\theta)$ to be a $Pareto\,(m,a)$ density, the posterior density of θ will be another Pareto density:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \frac{am^a}{\theta^{a+1}}\mathbf{I}_{\{\theta>m\}}\prod_{i=1}^{n}\left[\frac{1}{\theta}\mathbf{I}_{\{0<x_i<\theta\}}\right]\\ &\propto \frac{1}{\theta^{a+1}}\mathbf{I}_{\{\theta>m\}}\times\frac{1}{\theta^n}\mathbf{I}_{\{0<\max x_i<\theta\}}\\ &= \frac{1}{\theta^{a+n+1}}\mathbf{I}_{\{\theta>\max(m,\max x_i)\}}.\end{aligned}$$

That is, $(\theta|\mathbf{x})$ is $Pareto\,(\max(m,\max x_i),a+n)$.

Remark 5.2 (Pareto distribution) *A nonnegative random variable X has a Pareto distribution with parameters $m>0$ and $a>0$, denoted by*

$$X\sim Pareto\,(m,a)\,,$$

if it has density

$$\pi(x)=\frac{am^a}{x^{a+1}}\mathbf{I}_{\{x>m\}}.$$

- *The mean and variance of X are*

$$\begin{aligned}E[X] &= \frac{am}{(a-1)} && \textit{(provided } a>1\textit{)}\\ Var[X] &= \frac{am^2}{(a-1)^2(a-2)} && \textit{(provided } a>2\textit{).}\end{aligned}$$

- *The distribution function is*

$$F(x)=\left[1-\left(\frac{m}{x}\right)^a\right]\mathbf{I}_{\{x>m\}}.$$

- *The mode occurs at m.*

Example 5.6 *Suppose the prior guess of θ in a $U(0,\theta)$ model is summarized by a $Pareto\,(0.01,1.7)$ distribution. According to a sample $\{0.2,0.58,0.1,1.5,2.4,1.77\}$, what is the posterior probability that $\theta>4$?*

5.5 Multinomial distributions

Suppose we draw n times "with replacement" from a population with k different categories/kinds of items (n and k are known), where the proportion of the j-th category is $p_j > 0$, and $\sum_{j=1}^{k} p_j = 1$. If we are interested in the total numbers of items of each individual category, denoted by $X_1, \ldots, X_k$, they can be modelled by a Multinomial distribution $(X_1, \ldots, X_k) \sim \mathit{Multinomial}\,(n; p_1, \ldots, p_k)$, where n is called the number of trials and p_j's are the nonnegative probabilities of successes of drawing an j-th category such that $\sum_{j=1}^{k} p_j = 1$. Note that a multinomial distribution is a multivariate generalization of a Binomial distribution. The density of $(X_1, \ldots, X_k)$ is

$$f\left(x_1, \ldots, x_k | p_1, \ldots, p_k\right) = \frac{n!}{x_1! \cdots x_k!}\, p_1^{x_1} \ldots p_k^{x_k}\, \mathbf{I}_{\{\sum_{j=1}^{k} x_j = n\}}.$$

Viewing it as a density function in $(p_1, \ldots, p_k)$, one can see that a **Dirichlet family** is a conjugate family for $(p_1, \ldots, p_k)$.

Suppose we assume $(p_1, \ldots, p_k) \sim \mathit{Dirichlet}\,(\alpha_1, \ldots, \alpha_k)$, $\alpha_j > 0$, that is,

$$\pi\left(p_1, \ldots, p_k\right) = \frac{\Gamma\left(\alpha_1 + \cdots + \alpha_k\right)}{\Gamma\left(\alpha_1\right) \cdots \Gamma\left(\alpha_k\right)}\, p_1^{\alpha_1 - 1} \ldots p_k^{\alpha_k - 1}\, \mathbf{I}_{\left\{0 < p_j < 1 : \sum_{j=1}^{k} p_j = 1\right\}}. \tag{22}$$

Given M iid vectors of observations, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, M$, where x_{ij} records how many items among the n drawn items are of the j-th category for the i-th vector of observations. Define the total number of items of the j-th category in the M iid realizations of Multinomial vectors to be

$$\#\left(j\right) = \sum_{i=1}^{M} x_{ij}, \tag{23}$$

which implies $\sum_{j=1}^{k} \#\left(j\right) = Mn$. Then, the posterior density of $(p_1, \ldots, p_k)$ is given by

$$\begin{aligned}
\pi\left(p_1, \ldots, p_k | \mathbf{x}_1, \ldots, \mathbf{x}_M\right) &\propto \pi\left(p_1, \ldots, p_k\right) \prod_{i=1}^{M} f\left(x_{i1}, \ldots, x_{ik} | p_1, \ldots, p_k\right) \\
&\propto p_1^{\alpha_1 - 1} \ldots p_k^{\alpha_k - 1}\, \mathbf{I}_{\left\{0 < p_j < 1 : \sum_{j=1}^{k} p_j = 1\right\}} \times p_1^{\#(1)} \ldots p_k^{\#(k)} \\
&= p_1^{\alpha_1 + \#(1) - 1} \ldots p_k^{\alpha_k + \#(k) - 1}\, \mathbf{I}_{\left\{0 < p_j < 1 : \sum_{j=1}^{k} p_j = 1\right\}}.
\end{aligned}$$

That is, $(p_1, \ldots, p_k | \mathbf{x}_1, \ldots, \mathbf{x}_M)$ is $\mathit{Dirichlet}\,(\alpha_1 + \#\left(1\right), \ldots, \alpha_k + \#\left(k\right))$.

6 Predictive Distribution

In usual practice of Statistics, sometimes we might not only be interested in estimating the unknown parameters in the model (inference) but also want to make a forecast of any new observations (prediction). Suppose $x_1, \ldots, x_n$ are iid from $f(x|\theta)$ where θ is the unknown parameter of interest. Let $\pi(\theta)$ be the prior density. A posterior distribution of θ given the data can be obtained by techniques in the previous two sections. However, after observing n data points, one might want to know what is the probability that a new observation X_{n+1} from f or F is in between a given interval, that is,

$$P\left[a < X_{n+1} < b | X_1 = x_1, \ldots, X_n = x_n\right].$$

Such probability depends on the conditional density of $(X_{n+1}|X_1 = x_1, \ldots, X_n = x_n)$. It can be obtained from the posterior distribution of θ due to the following proposition. Denote the event, $\{X_1 = x_1, \ldots, X_n = x_n\}$ by $\{\mathbf{X} = \mathbf{x}\}$.

Proposition 6.1 *The* ***predictive distribution*** *of X_{n+1} based on the first n data $x_1, \ldots, x_n$ is*

$$P\left[X_{n+1} \leq x | \mathbf{X} = \mathbf{x}\right] = E\left[F(x|\theta)\,|\mathbf{X} = \mathbf{x}\right]. \tag{24}$$

This is a consequence of the double expectation formula $E[Y] = E[E[Y|Z]]$ (or $E[Y|X] = E[E[Y|Z, X]\,|X]$),

$$\begin{aligned} P\left[X_{n+1} \leq x | \mathbf{X} = \mathbf{x}\right] &= E\left[\mathbf{I}_{\{X_{n+1} \leq x\}} | \mathbf{X} = \mathbf{x}\right] \\ &= E\left\{E\left[\mathbf{I}_{\{X_{n+1} \leq x\}} | \theta, \mathbf{X} = \mathbf{x}\right] | \mathbf{X} = \mathbf{x}\right\} \\ &= E\left\{E\left[\mathbf{I}_{\{X_{n+1} \leq x\}} | \theta\right] | \mathbf{X} = \mathbf{x}\right\} \\ &= E\left\{F(x|\theta)\,|\mathbf{X} = \mathbf{x}\right\}. \end{aligned}$$

Similarly, We can also define the **predictive density** by $\pi_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) = E\left[f(x|\theta)\,|\mathbf{X} = \mathbf{x}\right]$.

6.1 Bernoulli distributions

Here $f(x|\theta) = \theta^x (1-\theta)^{1-x}$. Assume a $Beta(a, b)$ prior for θ, then the predictive density based on $\mathbf{X} = \mathbf{x}$ can be described by

$$\begin{aligned} P\left[X_{n+1} = 1 | \mathbf{X} = \mathbf{x}\right] &= E\left[\theta | \mathbf{X} = \mathbf{x}\right] \\ &= \frac{a + \sum_{i=1}^n x_i}{a + \sum_{i=1}^n x_i + b + n - \sum_{i=1}^n x_i} = \frac{a + \sum_{i=1}^n x_i}{a + b + n}, \end{aligned}$$

and

$$\begin{aligned} P\left[X_{n+1} = 0 | \mathbf{X} = \mathbf{x}\right] &= E\left[1 - \theta | \mathbf{X} = \mathbf{x}\right] \\ &= \frac{n - \sum_{i=1}^n x_i + b}{a + \sum_{i=1}^n x_i + b + n - \sum_{i=1}^n x_i} = \frac{b + n - \sum_{i=1}^n x_i}{a + b + n}, \end{aligned}$$

since $(\theta|\mathbf{x})$ is $Beta\left(a+\sum_{i=1}^{n} x_i, b+n-\sum_{i=1}^{n} x_i\right)$ according to section 5.1. Hence, the predictive distribution of X_{n+1} given $\{X_1 = x_1, \ldots, X_n = x_n\}$ is a Bernoulli random variable with probability of success $\dfrac{a+\sum_{i=1}^{n} x_i}{a+b+n}$. Hence, the predictive variance $Var\left[X_{n+1}|\mathbf{X}=\mathbf{x}\right]$ of the next observation is given by

$$\begin{aligned}
\frac{a+\sum_{i=1}^{n} x_i}{a+b+n}\left(1-\frac{a+\sum_{i=1}^{n} x_i}{a+b+n}\right) &= \frac{\left(a+\sum_{i=1}^{n} x_i\right)\left(b+n-\sum_{i=1}^{n} x_i\right)(a+b+n+1)}{(a+b+n)^2(a+b+n+1)} \\
&= (a+b+n+1)\, Var\left[\theta|\mathbf{X}=\mathbf{x}\right] \\
&= (a+b+n)\, Var\left[\theta|\mathbf{X}=\mathbf{x}\right] + Var\left[\theta|\mathbf{X}=\mathbf{x}\right].
\end{aligned}$$

Remark 6.1 *In general, the predictive distribution has two components in the variance since our prediction is subject to two sources of uncertainty / error:*

- *Error in estimating the parameter (from the posterior distribution), and*
- *Uncertainty due to the randomness of any future value (from the model).*

Example 6.1 *Suppose $X_i = 1$ if the sun rises on the i-th day. After observing sunrises on 500 days, are you certain that the sun will rise tomorrow based on a uniform prior?*

6.2 Exponential distributions

Here $f(x|\lambda) = \lambda \exp(-\lambda x)$. Assume a $Gamma(\alpha, 1/\beta)$ prior for λ, then the predictive density based on $\{\mathbf{X}=\mathbf{x}\}$ is given by

$$\begin{aligned}
E\left[\lambda \exp(-\lambda x)\,|\mathbf{X}=\mathbf{x}\right] &= \int_0^\infty \lambda \exp(-\lambda x)\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\lambda^{\alpha_n-1}\exp(-\beta_n\lambda)\,d\lambda \\
&= \frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\int_0^\infty \lambda^{\alpha_n+1-1}\exp\left[-(x+\beta_n)\lambda\right]d\lambda \\
&= \frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\frac{\Gamma(\alpha_n+1)}{(x+\beta_n)^{\alpha_n+1}} \\
&= \frac{\alpha_n(\beta_n)^{\alpha_n}}{(x+\beta_n)^{\alpha_n+1}}
\end{aligned}$$

since $(\lambda|\mathbf{x})$ is $Gamma(\alpha_n, 1/\beta_n)$ where $\alpha_n = \alpha+n$ and $\beta_n = \beta+\sum_{i=1}^{n} x_i$ according to section 5.3.

Example 6.2 *Consider an Exponential model with mean* $1/\lambda$*. Suppose* λ *is Gamma* $(1, 1)$*. (i) What is the predictive probability that a new observation is less than 8 without any data? (ii) How about after collecting 10 observations with sum equal to 98?*

6.3 $N(\mu, 1/r)$ with r known

Assume a $N(m, 1/t)$ prior for μ. According to proposition 4.1, given $\mathbf{x} = (x_1, \ldots, x_n) \overset{iid}{\sim} N(\mu, 1/r)$, the posterior distribution of μ is $N(m_n, 1/t_n)$. Then, the predictive density is given by

$$\begin{aligned}
E[f(x|\mu)|\mathbf{X} = \mathbf{x}] &= \int_{-\infty}^{\infty} \sqrt{\frac{r}{2\pi}} \exp\left[-\frac{r}{2}(x-\mu)^2\right] \sqrt{\frac{t_n}{2\pi}} \exp\left[-\frac{t_n}{2}(\mu - m_n)^2\right] d\mu \\
&= \sqrt{\frac{1}{2\pi}} \sqrt{\frac{t_n r}{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\left(t_n^{-1} + r^{-1}\right)}(x - m_n)^2\right] \exp\left[-\frac{t_n + r}{2}(\mu - \overline{m})^2\right] d\mu \\
&= \sqrt{\frac{t_n r}{2\pi(t_n + r)}} \exp\left[-\frac{1}{2\left(t_n^{-1} + r^{-1}\right)}(x - m_n)^2\right],
\end{aligned}$$

where the last equality is a result of integrating out a Normal density with mean $\overline{m} = \frac{xr}{r + t_n} + \frac{m_n t_n}{r + t_n}$ and precision $t_n + r$. That is, the predictive distribution of X_{n+1} given $\{X_1 = x_1, \ldots, X_n = x_n\}$ is $N\left(m_n, t_n^{-1} + r^{-1}\right)$ where $m_n = \left(\frac{nr}{nr + t}\right)\overline{x} + \left(\frac{t}{nr + t}\right) m$ and $t_n = nr + t$.

Remark 6.2 *It is noted that this predictive distribution and the posterior distribution of* μ *share the same mean,* m_n*, but the predictive distribution is more variable than the posterior distribution; the former one has an additional term* r^{-1}*, which is the population variance, in the variance compared with that of the latter one.*

Example 6.3 *Refer to example 4.1(i), (i) what is the predictive distribution of a new observation from the model if we have no data? (ii) Construct 95% intervals for θ and any new observation based on the 10 observed data respectively.*

6.4 $N(\mu, 1/\tau)$ with both μ and τ unknown

Assume a $Gamma-Normal\left(\alpha,\dfrac{1}{\beta};m,\dfrac{1}{t}\right)$ prior for (μ,τ). According to proposition 4.3, given $\mathbf{x}=(x_1,\ldots,x_n)\overset{iid}{\sim}N(\mu,1/\tau)$, the posterior distribution of (μ,τ) is $Gamma{-}Normal\left(\alpha_n,\dfrac{1}{\beta_n};m_n,\dfrac{1}{t_n}\right)$. Then, the predictive density is given by

$$
\begin{aligned}
&E\left[f(x|\mu,\tau)\,|\mathbf{X}=\mathbf{x}\right]\\
&=\int_0^\infty\int_{-\infty}^\infty\sqrt{\frac{\tau}{2\pi}}\exp\left[-\frac{\tau}{2}(x-\mu)^2\right]\sqrt{\frac{\tau t_n}{2\pi}}\exp\left[-\frac{\tau t_n}{2}(\mu-m_n)^2\right]\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\tau^{\alpha_n-1}\exp(-\beta_n\tau)\,d\mu d\tau\\
&=\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\frac{\sqrt{t_n}}{2\pi}\int_0^\infty\left\{\int_{-\infty}^\infty\tau\exp\left[-\frac{1}{2}\left\{\tau(\mu-x)^2+\tau t_n(\mu-m_n)^2\right\}\right]d\mu\right\}\tau^{\alpha_n-1}\exp(-\beta_n\tau)\,d\tau,
\end{aligned}
$$

where the inner integral reduces to $\sqrt{\dfrac{2\pi\tau}{1+t_n}}\exp\left[-\dfrac{1}{2\left(\tau^{-1}+(\tau t_n)^{-1}\right)}(x-m_n)^2\right]$ resulted from an application of the identity (12) followed by an integration of a Normal density. Hence, the predictive density equals

$$
\begin{aligned}
&\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\frac{\sqrt{t_n}}{2\pi}\int_0^\infty\sqrt{\frac{2\pi\tau}{1+t_n}}\exp\left[-\frac{1}{2\left(\tau^{-1}+(\tau t_n)^{-1}\right)}(x-m_n)^2\right]\tau^{\alpha_n-1}\exp(-\beta_n\tau)\,d\tau\\
=&\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\sqrt{\frac{t_n}{2\pi(1+t_n)}}\int_0^\infty\tau^{(\alpha_n+1/2)-1}\exp\left\{-\left[\beta_n+\frac{t_n}{2(1+t_n)}(x-m_n)^2\right]\tau\right\}d\tau\\
=&\frac{(\beta_n)^{\alpha_n}}{\Gamma(\alpha_n)}\sqrt{\frac{t_n}{2\pi(1+t_n)}}\frac{\Gamma(\alpha_n+1/2)}{\left[\beta_n+\frac{t_n}{2(1+t_n)}(x-m_n)^2\right]^{(2\alpha_n+1)/2}}\\
=&\frac{\Gamma((2\alpha_n+1)/2)}{\Gamma(2\alpha_n/2)}\sqrt{\frac{1}{2\alpha_n\pi}\left(\frac{\alpha_n}{\beta_n}\frac{t_n}{1+t_n}\right)}\left[1+\frac{1}{2\alpha_n}\left(\frac{\alpha_n}{\beta_n}\frac{t_n}{1+t_n}\right)(x-m_n)^2\right]^{-(2\alpha_n+1)/2}.
\end{aligned}
$$

That is, the predictive distribution of X_{n+1} given $X_1 = x_1, \ldots, X_n = x_n$ is a t-distribution with $2\alpha_n$ degrees of freedom, location parameter m_n and precision parameter $\left(\frac{\alpha_n}{\beta_n}\frac{t_n}{1+t_n}\right)$, where $\alpha_n = \alpha + \frac{n}{2}$, $\beta_n = \beta + \frac{1}{2}\left[\sum_{i=1}^n (x_i - \overline{x})^2 + \frac{(m-\overline{x})^2}{(t^{-1}+n^{-1})}\right]$, $m_n = \left(\frac{t}{t+n}\right)m + \left(\frac{n}{t+n}\right)\overline{x}$ and $t_n = t+n$.

Remark 6.3 *It is noted that this predictive distribution and the posterior distribution of μ also share the same mean, m_n, but the predictive distribution is more variable than the posterior distribution; the former one has a variance*

$$Var\left[X_{n+1}|\mathbf{X}=\mathbf{x}\right] = \left(\frac{\alpha_n}{\beta_n}\frac{t_n}{1+t_n}\right)^{-1}\left(\frac{2\alpha_n}{2\alpha_n-2}\right) = Var\left[\mu|\mathbf{X}=\mathbf{x}\right] + \frac{\beta_n}{\alpha_n-1},$$

that is, the variance has an additional positive term $\frac{\beta_n}{\alpha_n-1}$ apart from the posterior variance of μ.

Example 6.4 *Refer to example 4.2, another observation is obtained from the population after collection of the 36 observations. (i) What is the predictive probability that the new observation is greater than the sample mean of the previous 36 observations? (ii) What is a 90% prediction interval for the new observation?*

7 Hypothesis Testing: one-sample problem

Hypothesis tests are procedures to make decision about an interest parameter θ on choosing between two hypotheses: $\{\theta \in \Theta_1\}$ or $\{\theta \in \Theta_2\}$, which are disjoint subsets of the parameter space Θ. A Bayesian approach to this problem consists of computing and comparing the posterior probabilities of two events $\{\theta \in \Theta_1\}$ and $\{\theta \in \Theta_2\}$. This approach requires that the prior probabilities for the two choices are both positive; otherwise the corresponding posterior probabilities are also zero.

This section concerns one-sample problems only. Let us examine the simplest case first.

7.1 Test between two parameter values: $\{\theta = \theta_1\}$ or $\{\theta = \theta_2\}$

For a given parametric family $f(x|\theta)$, this is the choice between two known densities $f(x|\theta_1)$ and $f(x|\theta_2)$ based on observing data $\mathbf{x} = (x_1, \ldots, x_n)$. Suppose the prior is specified by a prior $\pi(\theta)$ (restricted on the two points θ_1 and θ_2) summarized by $P[\theta = \theta_1] = p > 0$ and $P[\theta = \theta_2] = 1 - p > 0$. By the Bayes theorem, the posterior probabilities for the two choices are

$$\begin{aligned} P[\theta = \theta_1|\mathbf{x}] &\propto P[\theta = \theta_1]\,\pi(\mathbf{x}|\theta_1) = p\pi(\mathbf{x}|\theta_1), \text{ and} \\ P[\theta = \theta_2|\mathbf{x}] &\propto P[\theta = \theta_2]\,\pi(\mathbf{x}|\theta_2) = (1-p)\,\pi(\mathbf{x}|\theta_2), \end{aligned} \qquad (25)$$

respectively. The normalization constants turn out to be

$$p\pi(\mathbf{x}|\theta_1) + (1-p)\,\pi(\mathbf{x}|\theta_2)$$

such that $\pi(\theta|\mathbf{x})$ is a proper probability distribution restricted on the two points θ_1 and θ_2.

One can define the **prior odds** and **posterior odds** on θ_1 against θ_2,

$$O = \frac{P[\theta = \theta_1]}{P[\theta = \theta_2]} \qquad \text{and} \qquad O_n = \frac{P[\theta = \theta_1|\mathbf{x}]}{P[\theta = \theta_2|\mathbf{x}]}, \qquad (26)$$

respectively, then it implies that

$$P[\theta = \theta_2|\mathbf{x}] = \frac{1}{1 + O_n}. \qquad (27)$$

Hence, if we have

$$\boxed{P\left[\theta = \theta_2|\mathbf{x}\right] > 0.5 \qquad \Longleftrightarrow \qquad O_n < 1}$$

we are in favour of θ_2.

Remark 7.1 *In general, one can calculate either the posterior probabilities or the posterior odds to conduct a hypothesis test.*

Example 7.1 *A housewife believes that the chance that the stock goes up or down in any given day equally likely. A stock forecaster observes that for the days in the last twenty years, the stock went up in 75% of the days and hence he claims that this probability is 0.75. Is the housewife's claim is more favourable if there are 62 up-days for stocks in current 100 days.*

7.2 Test between two parameter subsets: $\{\theta \in \Theta_1\}$ or $\{\theta \in \Theta_2\}$

This section discusses a choice of θ between two subsets of the parameter space, Θ_1 and Θ_2, where Θ_1 and Θ_2 are disjoint. This includes the last section as a special case. Assume a prior $\pi\left(\theta\right)$ that gives positive probabilities on both Θ_1 and Θ_2. This prior for θ governs the prior probability in belief of Θ_1 and Θ_2. Then the posterior density of θ given the data $\mathbf{x}$ is $\pi\left(\theta|\mathbf{x}\right) \propto \pi\left(\theta\right)\pi\left(\mathbf{x}|\theta\right)$. Suppose we make use of a conjugate prior for θ. Then, the posterior distribution of θ can be obtained nicely to be in the same parametric family. The testing procedure or the posterior probability for each choice depends on the nature of the two choice subsets; it can be divided into two different cases:

Case 1 $-$ $\Theta_1 \cup \Theta_2 = \Theta$: Suppose Θ_1 and Θ_2 span the whole parameter space Θ. The posterior probabilities for the two choices,

$$\begin{aligned} P\left[\theta \in \Theta_1 | \mathbf{x}\right] &= \int_{\Theta_1} \pi\left(\theta | \mathbf{x}\right) d\theta, \text{ and} \\ P\left[\theta \in \Theta_2 | \mathbf{x}\right] &= \int_{\Theta_2} \pi\left(\theta | \mathbf{x}\right) d\theta, \end{aligned} \tag{28}$$

respectively , can be obtained easily in many ways, like looking up in statistical tables or using Normal approximation.

Case 2 – $\Theta_1 \cup \Theta_2 \neq \Theta$: Suppose Θ_1 and Θ_2 altogether do not span the parameter space Θ. Then, the posterior probabilities for the two choices need to be re-normalized; that is,

$$P\left[\theta \in \Theta_1 | \mathbf{x}, \theta \in \Theta_1 \cup \Theta_2\right] = \frac{P\left[\theta \in \Theta_1 | \mathbf{x}\right]}{P\left[\theta \in \Theta_1 | \mathbf{x}\right] + P\left[\theta \in \Theta_2 | \mathbf{x}\right]}$$

and

$$P\left[\theta \in \Theta_2 | \mathbf{x}, \theta \in \Theta_1 \cup \Theta_2\right] = \frac{P\left[\theta \in \Theta_2 | \mathbf{x}\right]}{P\left[\theta \in \Theta_1 | \mathbf{x}\right] + P\left[\theta \in \Theta_2 | \mathbf{x}\right]}.$$

One can define similarly the prior odds and the posterior odds on θ_1 against θ_2 as

$$O = \frac{\int_{\Theta_1} \pi\left(\theta\right) d\theta}{\int_{\Theta_2} \pi\left(\theta\right) d\theta} \qquad \text{and} \qquad O_n = \frac{\int_{\Theta_1} \pi\left(\theta | \mathbf{x}\right) d\theta}{\int_{\Theta_2} \pi\left(\theta | \mathbf{x}\right) d\theta},$$

respectively. The prior odds is defined implicitly by your choice of prior for θ.

Remark 7.2 *The posterior probabilities given in equation (28) can be obtained in a different way via a re-expression. For instance,*

$$\begin{aligned} P\left[\theta \in \Theta_2 | \mathbf{x}\right] &\propto \int_{\Theta_2} \pi\left(\theta\right) \pi\left(\mathbf{x} | \theta\right) d\theta \\ &\propto P\left[\theta \in \Theta_2\right] \int \pi\left(\theta | \theta \in \Theta_2\right) \pi\left(\mathbf{x} | \theta\right) d\theta \end{aligned}$$

*where $\pi\left(\theta\right)$ has a total mass $P\left[\theta \in \Theta_2\right] \equiv \int_{\Theta_2} \pi\left(\theta\right) d\theta$ over Θ_2, and $\pi\left(\theta | \theta \in \Theta_2\right)$ is defined to be a proper density restricted over Θ_2 by a re-normalization as $\dfrac{\pi\left(\theta\right)}{P\left[\theta \in \Theta_2\right]}$. Such a prior distribution is called a **mixture prior**. The domain of integration Θ_2 can be removed since $\pi\left(\theta | \theta \in \Theta_2\right)$ is already defined only on Θ_2. Hence, adopting a mixture prior in a testing problem enables flexibilities in not only prior probabilities to the choices but also the prior behaviour of θ over different choice subsets; on one hand, the prior probabilities are governed by $P\left[\theta \in \Theta_1\right]$ and $P\left[\theta \in \Theta_2\right]$ that can be chosen* *arbitrarily as long as they add up to 1. On the other hand, one need not assume the same prior density over both subsets but different prior densities for θ over the individual choice subsets, and then work out the posterior probabilities by the latest formula separately up to a normalization constant.*

Example 7.2 *Refer to the last example, suppose now the housewife believes that the chance that the stock will go up in any given day is less than 54% and the stock forecaster believes that the chance should be at least 70% of the days. Is the housewife's claim is more favourable if there are 62 up-days for stocks in current 100 days? (i) by assuming a uniform prior, (ii) based on a mixture prior: $P[\theta < 0.54] = P[\theta > 0.7] = 1/2$, and both $\pi(\theta|\theta < 0.54)$ and $\pi(\theta|\theta > 0.7)$ are uniform.*

Example 7.3 *Consider the result of a test which follows $N(\theta, 1)$. Test results of 100 trials give a sample mean of 78.5. Assume a flat prior. (i) Will you favour more that the population mean is greater than 77? (ii) Do you favour $\{\theta < 77\}$ or $\{\theta > 80\}$?*

7.3 Test between a parameter point and a set: $\{\theta = \theta_1\}$ or $\{\theta \in \Theta_2\}$

It is very often that we are interested in whether a parameter is different from a certain pre-specified value or not. That is, the common test of a point null hypothesis in classical statistics

$$H_0 : \theta = \theta_1 \qquad \text{vs} \qquad H_1 : \theta \neq \theta_1.$$

To tackle problems concerning decision between $\{\theta = \theta_1\}$ and $\{\theta \in \Theta_2\}$ from a Bayesian viewpoint, one has to assume a mixture prior for θ; any continuous prior distribution assumes zero probability at a single point, hence if we assume a usual continuous prior for θ, this results in zero prior probability at θ_1, as well as zero posterior probability at θ_1 by the Bayes' Theorem. A mixture prior, which helps circumvent this problem, is defined as such:

$$\begin{cases} \theta = \theta_1 \quad \text{with probability} \quad P[\theta = \theta_1] = p, \\ \theta \in \Theta_2 \quad \text{with probability} \quad P[\theta \in \Theta_2] = 1 - p; \\ \qquad \text{and then } (\theta | \theta \in \Theta_2) \text{ has a proper prior density.} \end{cases} \tag{29}$$

Denote the proper density restricted over Θ_2 as $\pi(\theta | \theta \in \Theta_2)$. Formally, the mixture prior distribution for θ can be expressed as this mixture form

$$p\delta_{\theta_1} + (1 - p) Y \tag{30}$$

where δ_{θ_1} represents a random variable of $\{\theta = \theta_1\}$ with probability one and Y is a random variable defined on Θ_2, having a density of $\pi(\theta | \theta \in \Theta_2)$. Then, the two posterior probabilities, $P[\theta = \theta_1 | \mathbf{x}]$ and $P[\theta \in \Theta_2 | \mathbf{x}]$, can be computed using the techniques in the previous two sections separately up to a normalization constant; the posterior probability that $\{\theta = \theta_1\}$ is

$$P[\theta = \theta_1 | \mathbf{x}] \propto P[\theta = \theta_1] \times \pi(\mathbf{x} | \theta_1) = p\pi(\mathbf{x} | \theta_1), \tag{31}$$

whereas the posterior probability that $\{\theta \in \Theta_2\}$ is

$$\begin{aligned} P[\theta \in \Theta_2 | \mathbf{x}] &\propto \int P[\theta \in \Theta_2] \, \pi(\theta | \theta \in \Theta_2) \, \pi(\mathbf{x} | \theta) \, d\theta \\ &= (1 - p) \int \pi(\theta | \theta \in \Theta_2) \, \pi(\mathbf{x} | \theta) \, d\theta. \end{aligned} \tag{32}$$

The normalization constants are

$$p\pi\left(\mathbf{x}|\theta_1\right)+(1-p)\int\pi\left(\theta|\theta\in\Theta_2\right)\pi\left(\mathbf{x}|\theta\right)d\theta.$$

Example 7.4 *Refer to example 7.1, the housewife's husband believes 99% in his wife's claim. After observing 62 up-days for stocks in the current 100 days, will he still stand by her? (Assume a flat prior.) [Hint: Sterling formula:* $n!\approx\sqrt{2\pi}\times n^{n+0.5}e^{-n}$*.]*

Example 7.5 *Refer to example 7.3, (i) will you agree with the claim that* $\theta=78$ *if you have a fair chance in mind about the claim (i.e.,* $P\left[\theta=78\right]=1/2$*) and assume a* $N\left(79,3\right)$ *prior for* θ *when* $\theta\neq78$*?*

7.4 Hypothesis tests with nuisance parameters

Very often, parameters of interest are restricted to a lower-dimensional subset Θ of the full parameter set $\widetilde{\Theta}$, like in the Normal population with unknown mean and variance in which it is very often that only the population mean is of interest. One constructs a prior on the full parameter $\widetilde{\Theta}$ as usual. Then, proceeds as before after applying the double expectation formula to integrate out the nuisance/unwanted parameters.

Suppose $(x_1, \ldots, x_n | \theta, \lambda)$ has a joint density $\pi(\mathbf{x}|\theta, \lambda)$. One is interested in choosing between $\{\theta \in \Theta_1\}$ or $\{\theta \in \Theta_2\}$, without any knowledge of the value of λ. We assign these two choices to have prior probabilities p and $1-p$, respectively. That is, $P[\theta \in \Theta_1] = p$ and $P[\theta \in \Theta_2] = 1-p$. To continue, one needs to get $\pi(\mathbf{x}|\theta)$ by integrating out λ via a conditional density $\pi(\lambda|\theta)$,

$$\pi(\mathbf{x}|\theta) = \int \pi(\mathbf{x}|\theta, \lambda)\pi(\lambda|\theta)\, d\lambda. \tag{33}$$

Then, the test follows from the previous sections.

Example 7.6 *Suppose $x_1, ..., x_n$ are iid $N(\mu, 1/\tau)$ where μ, τ are both unknown. The choices are $\{\mu = \mu_1\}$ and $\{\mu = \mu_2\}$. Assume independently $(\tau|\mu = \mu_1)$ is Gamma $(\alpha_1, 1/\beta_1)$ and $(\tau|\mu = \mu_2)$ is Gamma $(\alpha_2, 1/\beta_2)$. What is the posterior odds on μ_1 against μ_2?*

8 Bayesian computation

Most parametric models came across or dealt with in previous chapters possess a conjugate family for their parameters. This alleviates a lot of difficulties in determination of posterior distribution of the parameters since it can avoid any complex and undesirable integration (i.e., the evaluation of the normalization constant according to a given kernel. For instance, the normalization constant given a density kernel $k(\mathbf{x}, \theta)$ is $\int_\Theta k(\mathbf{x}, \theta)\, d\theta$). This chapter concerns **numerical approximation** techniques that can be used to approximate any expectation or probability when the answers are **NOT** readily available in some known forms or in statistical tables.

8.1 Noniterative Monte Carlo methods

8.1.1 Monte Carlo integration

Monte Carlo integration depends on random samples of the target random variables that are available from **direct** simulations. The most basic definition of Monte Carlo integration is described as follows: Suppose we want to calculate

$$\gamma_g = \int_\Theta g(\theta)\, p(\theta)\, d\theta = E[g(\theta)]$$

where the expectation is based on θ which has a density $p(\theta)$. Then, the integral can be approximated by the average

$$\widetilde{\gamma}_g = \frac{1}{M} \sum_{i=1}^{M} g(\theta_i) \tag{34}$$

if $\theta_1, \ldots, \theta_M \overset{iid}{\sim} p(\theta)$ since, by the **Strong Law of Large Numbers**, $\widetilde{\gamma}_g$ converges to $E[g(\theta)]$ with probability one as $M \to \infty$. Usually, M is called the **Monte Carlo size**.

As an application in Bayes inference, we have

$$\boxed{p(\theta) \iff \pi(\theta|\mathbf{x}) \qquad \text{and} \qquad \gamma_g \iff E[g(\theta)|\mathbf{x}].}$$

Hence numerical approximations of posterior expectations require only a sample of size M, denoted by $\theta_1, \ldots, \theta_M$, from the posterior distribution $\pi(\theta|\mathbf{x})$. For example,

(i) Suppose $g(\theta) = \theta$. Then, the posterior mean of θ

$$\gamma_g = E[\theta|\mathbf{x}] = \int_\Theta \theta\pi(\theta|\mathbf{x})\, d\theta$$

is approximated by

$$\widetilde{\gamma_g} = \frac{1}{M}\sum_{i=1}^{M}\theta_i.$$

(ii) Suppose $g(\theta) = \mathbf{I}_{\{\theta\in A\}}$. Then, the posterior probability that θ is in a subset A of the parameter space,

$$\gamma_g = E\left[\mathbf{I}_{\{\theta\in A\}}|\mathbf{x}\right] = \int_\Theta \mathbf{I}_{\{\theta\in A\}}\pi(\theta|\mathbf{x})\, d\theta$$

is approximated by

$$\widetilde{\gamma_g} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{I}_{\{\theta_i\in A\}}.$$

The quality of the approximation in (34) improves as we increase the Monte Carlo size M. [rather than the sample size n of the dataset which is typically beyond our control.] It can be measured by **the standard error of $\widetilde{\gamma_g}$**, which is given by the square root of

$$Var[\widetilde{\gamma_g}] = Var\left[\frac{1}{M}\sum_{i=1}^{M}g(\theta_i)\right] = \frac{1}{M^2}MVar[g(\theta)] = \frac{1}{M}Var[g(\theta)]. \tag{35}$$

Let

$$Var[g(\theta)] = \gamma_h = E[h(\theta)] \equiv E\{g(\theta) - E[g(\theta)]\}^2.$$

Based on the same idea in approximating $E[g(\theta)]$ by $\widetilde{\gamma_g}$, one can approximate $Var[g(\theta)]$ based on the same sample of size M from $p(\theta)$ by

$$\widetilde{\gamma_h} = \frac{1}{M-1}\sum_{i=1}^{M}[g(\theta_i) - \widetilde{\gamma_g}]^2.$$

The denominator $M-1$ is due to the fact that one degree of freedom is lost in approximating the unknown $E[g(\theta)]$ by $\widetilde{\gamma_g}$. Substituting $\widetilde{\gamma_h}$ into (35) gives the estimated standard error of the estimator $\widetilde{\gamma_g}$ to be

$$\sqrt{\frac{1}{M(M-1)}\sum_{i=1}^{M}[g(\theta_i) - \widetilde{\gamma_g}]^2}. \tag{36}$$

Remark 8.1 *In this chapter, the computational procedures are emphasized; we prefer working out the details by hands to computers for clarity. We have to rely on small Monte Carlo samples. Hence,* ***most examples to be given in this chapter do not satisfy the condition that the Monte Carlo size*** M ***is large, in turn, do not have accurate answers.***

Example 8.1 *Suppose* $X \sim N\left(\mu, \sigma^2\right)$. *Given five independent samples from* X *as* $1.3, -0.5, -1.75, 0.33, -1.17$. *Give estimates for (i)* μ, *(ii)* σ, *(iii)* $P[X > 0.3]$, *(iv) the mean of* $Y \equiv X^2/(X-1)$, *(v)* $P\left[X^2 > 0.3\right]$, *and (vi) variance of the estimate for* $P[X > 0.3]$.

Example 8.2 *Suppose we obtain five independent samples* $1.3, -0.5, -1.75, 0.33, -1.17$ *from the posterior of the mean* μ *of a population* X. *Approximate (i) your Bayes estimate for the population mean* μ, *(ii) the posterior variance of* μ, *and (iii) the posterior probability that* $\mu > 0.3$.

Example 8.3 *Consider the Normal model with both mean μ and precision τ unknown. Describe how to approximate the Bayes estimate for coefficient of variation $CV = \mu\sqrt{\tau}$.*

For simplicity, assume the posterior distribution of (μ, τ) is $Gamma - Normal\,(\alpha, 1/\beta; m, 1/t)$. Then, we can draw independent pairs $(\mu_1, \tau_1), \ldots, (\mu_M, \tau_M)$ by (i) drawing $\tau \sim Gamma\,(\alpha, 1/\beta)$, and then (ii) drawing $\mu|\tau \sim N\,(m, 1/\,(\tau t))$.

Hence, the posterior mean of $\mu\sqrt{\tau}$, which is the Bayes estimate for CV,

$$E\left[\mu\sqrt{\tau}|\mathbf{x}\right] \approx \frac{1}{M}\sum_{i=1}^{M}\mu_i\sqrt{\tau_i},$$

since $(\mu_1, \tau_1), \ldots, (\mu_M, \tau_M)$ are independent samples from the posterior distribution of (μ, τ), which is $Gamma - Normal\,(\alpha, 1/\beta; m, 1/t)$.

It seems that almost any quantity of interest in a statistical model can be approximated once independent samples from the posterior distribution of the parameters are available. But **what if we can't directly sample from the posterior distribution?**

It is very often that some intractable integrations remain in many practical problems such that the posterior distributions cannot be easily recognized to be within certain parametric models.

Example 8.4 *As in example 5.1, which concerns a Normal model with known variance r where a Student's t prior distribution with 40 degrees of freedom is assumed for the unknown mean μ, one can neither recognize the posterior distribution of μ from*

$$\pi\,(\mu|\mathbf{x}) \propto \left[1 + \frac{\mu^2}{40}\right]^{-20.5} \exp\left[-\frac{nr}{2}\,(\mu - \overline{x})^2\right],$$

to be within which known parametric model nor evaluate the normalization constant via integrating the above kernel over the real line.

Remark 8.2 *Even though in some cases, the denominator in the posterior densities can be evaluated, any inference (e.g., computations of moments and quantiles) leads to more integrations.*

The rest of this chapter discusses some numerical approximation methods that are indispensible especially when closed-form expressions of posterior densities are unavailable. The methods rely on random samples that are closely drawn/simulated from the desired posterior densities.

8.1.2 Importance sampling

Importance sampler is one technique which provides an **indirect** sampling procedure from the target density. Suppose we wish to approximate a posterior expectation, say,

$$E\left[g\left(\theta\right)|\mathbf{x}\right] = \int_{\Theta} g\left(\theta\right)\pi\left(\theta|\mathbf{x}\right)d\theta,$$

but we are unable to sample from the posterior density $\pi\left(\theta|\mathbf{x}\right)$, which is our target density. However, if we can sample easily from a certain density $h\left(\theta\right)$ such that that

$$\omega\left(\theta\right)h\left(\theta\right) = \pi\left(\theta\right)\pi\left(\mathbf{x}|\theta\right).$$

Define the **importance weight function**

$$\omega\left(\theta\right) = \frac{\pi\left(\theta\right)\pi\left(\mathbf{x}|\theta\right)}{h\left(\theta\right)}. \tag{37}$$

Hence,

$$\begin{aligned} E\left[g\left(\theta\right)|\mathbf{x}\right] &= \int_{\Theta} g\left(\theta\right)\pi\left(\theta|\mathbf{x}\right)d\theta \\ &= \frac{\int_{\Theta} g\left(\theta\right)\pi\left(\theta\right)\pi\left(\mathbf{x}|\theta\right)d\theta}{\int_{\Theta}\pi\left(\theta\right)\pi\left(\mathbf{x}|\theta\right)d\theta} = \frac{\int_{\Theta} g\left(\theta\right)\omega\left(\theta\right)h\left(\theta\right)d\theta}{\int_{\Theta}\omega\left(\theta\right)h\left(\theta\right)d\theta} \\ &\approx \frac{\sum_{i=1}^{M} g\left(\theta_i\right)\omega\left(\theta_i\right)}{\sum_{i=1}^{M}\omega\left(\theta_i\right)}, \end{aligned} \tag{38}$$

where $\theta_1, \ldots, \theta_M \overset{iid}{\sim} h\left(\theta\right)$. The approximation is due to individual applications of the Strong Law of Large Numbers on the numerator and the denominator. Here, $h\left(\theta\right)$ is called the **importance density** or the **trial density**; how closely it resembles the posterior density controls how good the approximation in (38) is.

Remark 8.3 *One natural candidate of an importance density in Bayes inference is the* ***prior density*** *of the unknown parameter, that is,* $h\left(\theta\right) = \pi\left(\theta\right)$. *In that case,*

$$\begin{aligned} E\left[g\left(\theta\right)|\mathbf{x}\right] &= \int g\left(\theta\right)\pi\left(\theta|\mathbf{x}\right)d\theta \\ &= \frac{\int g\left(\theta\right)\pi\left(\mathbf{x}|\theta\right)\pi\left(\theta\right)d\theta}{\int \pi\left(\mathbf{x}|\theta\right)\pi\left(\theta\right)d\theta} \\ &\approx \frac{\sum_{i=1}^{M} g\left(\theta_i\right)\pi\left(\mathbf{x}|\theta_i\right)}{\sum_{i=1}^{M}\pi\left(\mathbf{x}|\theta_i\right)} \end{aligned} \tag{39}$$

where $\theta_1, \ldots, \theta_M \overset{iid}{\sim} \pi(\theta)$ *and* $\pi(\mathbf{x}|\theta)$ *is the importance weight function. As we mentioned before, the accuracy of the approximation depends on how well the importance density* $h(\theta)$ *can approximate the target density. Since the prior density* $\pi(\theta)$ *might not carry much information about* θ*, especially when* $\pi(\theta)$ *is a flat prior, such that it dislikes the posterior density which is proportional to the prior times likelihood but not only the prior, the above approximation (39) sometimes could be terrible.*

Example 8.5 *Refer to example 4.1(i), based on five independent samples* $1.3, -0.5, -1.75, 0.33, -1.17$ *from the prior of the mean* θ*, which is a standard Normal density, what are your Bayes estimates for (i) the population mean* θ*, and (ii) the probability that the population has a positive mean.*

Example 8.6 *Repeat the above problem using a* $N(16.334, 1/10)$ *density, which is the density in* θ *from the* ***normalized likelihood function****, as the importance density based on five independent observations* $16.55, 16.6, 16.02, 16.81, 15.9.$

8.2 Markov chain Monte Carlo approximations

Suppose we are interested in approximating the following expectation,

$$E\left[g\left(X,Y\right)\right]=\int\int g\left(x,y\right)\pi\left(x,y\right)dxdy \tag{40}$$

where $\pi\left(x,y\right)$ is a bivariate density of the pair $\left(X,Y\right)$, which is our target density to draw samples from. Sometimes, the noniterative methods mentioned in the previous section could be irrelevant. In particular, on one hand, we CANNOT use the Monte Carlo integration if we are NOT able to draw exact samples from the target density.[1] On the other hand, we CANNOT find an importance density that are both easy to sample and close to the target density such that the importance sampler is not desired. Under this situation, we could recourse to the Markov chain Monte Carlo (MCMC) method. It is primarily based on sampling a sequence of random samplesthat are **correlated with each other**. The sequence constitutes a **stationary Markov chain** with a unique **stationary distribution** that coincides with the target distribution.

We will illustrate here the MCMC idea in the two-variable case as in the above scenario, of which we are interested in drawing samples from a bivariate distribution $\pi\left(x,y\right)$. Suppose $\pi\left(x,y\right)$ is known up to a proportional constant, that is,

$$\pi\left(x,y\right)=\frac{h\left(x,y\right)}{C}, \tag{41}$$

where $h\left(x,y\right)$ is the kernel of the joint density and C is the normalization constant.

According to the Markov chain Monte Carlo method, the sampled Markov chain is represented by $\left(x_0,y_0\right),\left(x_1,y_1\right),\ldots,\left(x_M,y_M\right),\ldots$. In order to make use of them, the stationary distribution of the Markov chain has to be **exactly identical to** our desired bivariate distribution $\pi\left(x,y\right)$. It suffices to answer the **QUESTION**,

How to construct such a chain possessing the desired stationary distribution?

[1] The reason can be two-fold; it could be too difficult to draw exact samples from the target density, or the target density is only known up to a normalization constant in terms of a complex integral.

or in other words,

$$\boxed{\text{How to define moves from one state } (x_i, y_i) \text{ to the next state } (x_{i+1}, y_{i+1})?}$$

8.2.1 The Gibbs sampler

The Gibbs sampler is constructed based on rather natural predictions between the multiple variables which defines the domain of the multiple integrals. It is one of many methods producing one such chain that can address the concerned **QUESTION**. The **transition function of a Gibbs chain**, which governs the movement from one state to the next, is defined through **predictions between** X **and** Y, and is particularly appealing.

To describe it, given the present state (x_i, y_i). Select X and Y consecutively to form the next state (x_{i+1}, y_{i+1}) according to the following two "predictive" densities:

$$\begin{cases} (x_{i+1}|x_i, y_i) & \text{has density} \quad \pi(x_{i+1}|y_i) \propto h(x_{i+1}, y_i); \\ (y_{i+1}|x_i, y_i, x_{i+1}) & \text{has density} \quad \pi(y_{i+1}|x_{i+1}) \propto h(x_{i+1}, y_{i+1}). \end{cases} \tag{42}$$

Remark 8.4 *It should be noted that the knowledge of the normalization constant C is not required in identifying the two "predictive" densities. By the product rule, the conditional density of $(x_{i+1}, y_{i+1}|x_i, y_i)$,*

$$k(x_{i+1}, y_{i+1}|x_i, y_i) = \pi(x_{i+1}|y_i)\,\pi(y_{i+1}|x_{i+1}), \tag{43}$$

gives the ***transition function of the Gibbs chain****. The above transition function guarantees that the Gibbs chain is a stationary Markov chain with a unique stationary distribution of $\pi(x, y)$. That is, if the present state (x_i, y_i) is sampled from a distribution $\pi(x, y)$, then the next state (x_{i+1}, y_{i+1}) will follow the same distribution $\pi(x, y)$.*

In practice, a **Gibbs sampler can be described as follows**:

(i) Start with an arbitrary initial value of state at time 0, (x_0, y_0), which should be a possible pair of variates from the bivariate density.[2]

(ii) Move to the next state (x_{i+1}, y_{i+1}) at time $i + 1$ from the present state (x_i, y_i) at time i via the transition function $k(x_{i+1}, y_{i+1}|x_i, y_i)$, $i = 0, 1, 2, \ldots$.

[2] Note that (x_0, y_0) could be from any distribution other than the stationary distribution $\pi(x, y)$.

(iii) Repeat step (ii) for a large number of times to get a Markov chain $(x_0, y_0), (x_1, y_1), \ldots,$ $(x_w, y_w), \ldots, (x_{w+M}, y_{w+M})$.

Finally, the first w members in the chain will be discarded, while the next M members, $(x_{w+1}, y_{w+1}), \ldots, (x_{w+M}, y_{w+M})$, will be kept to calculate estimates for the expectation (40).

Remark 8.5 *Usually, we run the chain for a long period of time, say, $w + M$, where w is called a "burn-in" period and M is the required Monte Carlo size. If w is "large", we are safe to assume that (x_w, y_w) at time w is sampled from the stationary distribution $\pi(x, y)$, in addition, the following states at time $t > w$ all come from the same distribution $\pi(x, y)$ due to remark* 8.4.

The following law of large numbers for a stationary Markov chain prevails: Suppose $\int\int g(x, y)\, \pi(x, y)\, dxdy$ is finite. If a Markov chain $(x_0, y_0), (x_1, y_1), \ldots, (x_M, y_M)$ with a unique stationary distribution $\pi(x, y)$ is obtained following the above 3 steps, the average

$$\boxed{\frac{1}{M}\sum_{i=1}^{M} g(x_i, y_i) \text{ converges to } E[g(X, Y)].} \tag{44}$$

The convergence in (44) is the backbone of the Markov chain Monte Carlo method. The above assertion enables us to treat the correlated variates, $(x_1, y_1), \ldots, (x_M, y_M)$, as if they were iid from $\pi(x, y)$.

Example 8.7 *Consider a random vector $(X_1, X_2, X_3) \sim Multinomial(n; p_1, p_2, p_3)$ where n is a known integer greater than 1 and p_1, p_2, p_3 are known probabilities such that $\sum_{i=1}^{3} p_i = 1$. Suppose we are interested in knowing the probability that $X_1 > X_2$. Via a Gibbs sampler, it suffices to draw samples from the conditional densities $\pi(x_1|x_2)$ and $\pi(x_2|x_1)$. According to example 2.8,*

$$(X_2|X_1 = x_1) \sim Bin\left(n - x_1, \frac{p_2}{1 - p_1}\right).$$

Then, by symmetry, we have

$$(X_1|X_2 = x_2) \sim Bin\left(n - x_2, \frac{p_1}{1 - p_2}\right).$$

These two conditional densities define a Gibbs sampler as follows: Choose an initial state $(x_{1,0}, x_{2,0})$ as, say, $(1, 2)$.[3] *Conditional on the present state $(x_{1,i}, x_{2,i})$ at time i, to select the next state $(x_{1,i+1}, x_{2,i+1})$ at time $i + 1$, we go through the two following prediction steps in a Gibbs cycle,*

[3] The two numbers have to be positive and their sum do not exceed n.

Step 1 $$\left(x_{2,i+1}\left|x_{1,i}\right.\right) \sim Bin\left(n - x_{1,i}, \frac{p_2}{1-p_1}\right);$$

Step 2 $$\left(x_{1,i+1}\left|x_{2,i+1}\right.\right) \sim Bin\left(n - x_{2,i+1}, \frac{p_1}{1-p_2}\right).$$

Repeating the two steps until getting a Markov chain $\left(x_{1,0}, x_{2,0}\right), \left(x_{1,1}, x_{2,1}\right), \ldots, \left(x_{1,w}, x_{2,w}\right), \ldots,$ $\left(x_{1,w+M}, x_{2,w+M}\right)$*. Then, the probability that* $X_1 > X_2$ *can be approximated by*

$$\frac{1}{M}\sum_{i=w+1}^{w+M} \mathbf{I}_{\{x_{2,i} > x_{1,i}\}} \qquad \textit{if } M \textit{ is large.}$$

Example 8.8 *Let us consider the problem in example 8.3 again. We want to approximate the posterior mean of coefficient of variation by a Gibbs sampler. First of all, note that we are approximating*

$$E\left[\mu\sqrt{\tau}\left|\mathbf{x}\right.\right] = \int\int \mu\sqrt{\tau}\pi\left(\mu, \tau|\mathbf{x}\right) d\tau d\mu$$

where $\pi\left(\mu, \tau|\mathbf{x}\right)$ *is jointly* $Gamma - Normal\left(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n}\right)$*. Here* $g\left(X, Y\right)$ *in (44) is* $\mu\sqrt{\tau}$*. We need to find the conditional densities, while these are already given in example 2.6 as*

$$\left(\mu\left|\tau, \mathbf{x}\right.\right) \sim N\left(m_n, \frac{1}{\tau t_n}\right)$$

and

$$\left(\tau\left|\mu, \mathbf{x}\right.\right) \sim Gamma\left(\alpha_n + \frac{1}{2}, \frac{1}{\beta_n + \frac{t_n}{2}\left(\mu - m_n\right)^2}\right).$$

These two conditional densities define a Gibbs sampler as follows: Choose an initial state $\left(\mu_0, \tau_0\right)$ *as, say,* $(0, 2)$.[4] *Conditional on the present state* $\left(\mu_i, \tau_i\right)$ *at time* i*, to select the next state* $\left(\mu_{i+1}, \tau_{i+1}\right)$ *at time* $i+1$*, we go through the following two prediction steps in a Gibbs cycle,*

Step 1 $$\left(\tau_{i+1}\left|\mu_i, \mathbf{x}\right.\right) \sim Gamma\left(\alpha_n + \tfrac{1}{2}, \frac{1}{\beta_n + \frac{t_n}{2}\left(\mu_i - m_n\right)^2}\right);$$

Step 2 $$\left(\mu_{i+1}\left|\tau_{i+1}, \mathbf{x}\right.\right) \sim N\left(m_n, \frac{1}{\tau_{i+1}t_n}\right).$$

Repeating the two steps until getting a Markov chain $\left(\mu_0, \tau_0\right), \left(\mu_1, \tau_1\right), \ldots, \left(\mu_w, \tau_w\right), \ldots,$ $\left(\mu_{w+M}, \tau_{w+M}\right)$*. Then, the posterior mean of coefficient of variation can be approximated by*

$$\frac{1}{M}\sum_{i=w+1}^{w+M} \mu_i\sqrt{\tau_i} \qquad \textit{if } M \textit{ is large.}$$

[4] The initial value for τ CANNOT be a negative number.