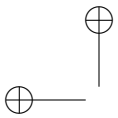


Chapter 4

Fundamentals of Probability, Random Processes and Statistics

We summarize in this chapter those aspects of probability, random processes and statistics that are employed in subsequent chapters. The discussion is necessarily brief and additional details can be found in the references cited both in the text and noted in Section 4.9.

4.1 Random Variables, Distributions and Densities

When constructing statistical models for physical and biological processes, we will consider parameters and measurement errors to be random variables whose statistical properties or distributions we wish to infer using measured data. The classical probability space provides the basis for defining and illustrating these concepts.

Definition 4.1 (Probability Space). A probability space $(\Omega, \mathcal{F}, P)$ is comprised of three components:

Ω: sample space is the set of all possible outcomes from an experiment;

$\mathcal{F}$: σ-field of subsets of Ω that contains all events of interest;

$P : \mathcal{F} \to [0, 1]$: probability or measure that satisfies the postulates

(i) $P(\emptyset) = 0$,

(ii) $P(\Omega) = 1$,

(iii) if $A_i \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

We note that the concept of probability depends on whether one is considering a frequentist (classical) or Bayesian perspective. In the frequentist view, probabilities are defined as the frequency with which an event occurs if the experiment is repeated a large number of times. The Bayesian perspective treats probabilities as a distribution of subjective values, rather than a single frequency, that are constructed or updated as data is observed.

Example 4.2. Consider an experiment in which we flip two individual coins (e.g., a quarter and nickel) multiple times and record the outcome which consists of an ordered pair. The sample space and σ-field of events are thus

$$\begin{aligned} \Omega &= \{(H,H),(T,H),(H,T),(T,T)\} \\ \mathcal{F} &= \{\emptyset,(H,H),(T,H),(H,T),(T,T),\Omega,\{(H,T),(T,H),\cdots\}\}. \end{aligned} \tag{4.1}$$

Note that $\mathcal{F}$ contains all countable intersections and unions of elements in Ω. If we flip the pair twice, two possible events are

$$A = \{(H,H),(T,H)\} \ , \ B = \{(H,H),(H,T)\}.$$

For fair coins, the frequentist perspective yields the probabilities

$$P(A) = \frac{1}{2} \quad , \ P(B) = \frac{1}{2} \quad , \ P(A \cap B) = \frac{1}{4} \quad , \ P(A \cup B) = \frac{3}{4}.$$

We note that because the events are independent, $P(A \cap B) = P(A)P(B)$. We will revisit the probabilities associated with flipping a coin from the Bayesian perspective in Example 4.66 of Section 4.8.2.

We now define univariate random variables, distributions and densities.

4.1.1 Univariate Concepts

Definition 4.3 (Random Variable). A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$; i.e., it is measurable. A random variable is said to be discrete if it takes values in a countable subset $\{x_1, x_2, \cdots\}$ of $\mathbb{R}$.

Definition 4.4 (Realization). The value

$$x = X(\omega)$$

of a random variable X for an event $\omega \in \Omega$ is termed a realization of X.

We note that in the statistics literature, many authors employ the same notation for the random variable and realization and let the context dictate the meaning. For those who are new to the field, this can obscure the meaning and, to the degree possible, we will use different notation for random variables and their realizations.

Definition 4.5 (Cumulative Distribution Function). Associated with every random variable X is a cumulative distribution function (cdf) $F_X : \mathbb{R} \to [0,1]$ given by

$$F_X(x) = P\{\omega \in \Omega | X(\omega) \leq x\}. \tag{4.2}$$

This is often expressed as $F_X(x) = P\{X \leq x\}$ which should be interpreted in the sense of (4.2). The following example illustrates the construction of a cdf for a discrete random variable.

Example 4.6. Consider the experiment of Example 4.2 in which our event ω consists of a single flip of a pair of coins. We define $X(\omega)$ to be the number of heads associated with the event so that

$$\begin{aligned} X(H,H) &= 2 \\ X(H,T) &= X(T,H) = 1 \\ X(T,T) &= 0. \end{aligned}$$

For $x < 0$, the probability of finding an event $\omega \in \Omega$ such that $X(\omega) \leq x$ is 0 so $F_X(x) = 0$ for $x < 0$. Similar analysis yields the cdf relation

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & , \quad 0 \leq x < 1 \\ 3/4 & , \quad 1 \leq x < 2 \\ 1 & , \quad x \geq 2 \end{cases}$$

which is plotted in Figure 4.1.

It is observed that, by construction, the cdf satisfies the properties

$$\begin{aligned} &\text{(i)} \quad \lim_{x \to -\infty} F_X(x) = 0 \\ &\text{(ii)} \quad x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \\ &\text{(iii)} \quad \lim_{x \to \infty} F_X(x) = 1. \end{aligned} \tag{4.3}$$

This is an example of a càdlàg (French "continue à droite, limite à gauche) function that is right-continuous and has left limits everywhere. These functions also arise in stochastic processes that admit jumps.

For continuous and discrete random variables the probability density function (pdf) and probability mass function are defined as follows.

Definition 4.7 (Probability Density Function). The random variable X is continuous if its cumulative distribution function is absolutely continuous and hence can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds \ , \ x \in \mathbb{R}$$

where the derivative $f_X = \frac{dF_x}{dx}$ mapping $\mathbb{R}$ to $[0, \infty)$ is called the probability density function (pdf) of X.

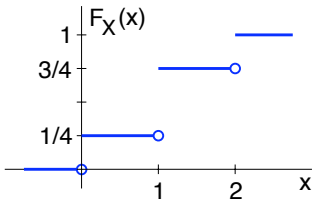


Figure 4.1. *Cumulative distribution function for Example 4.6.*

Definition 4.8 (Probability Mass Function). The probability mass function of a discrete random variable X is given by $f_X(x) = P(X = x)$.

The pdf properties

$$\text{(i)} \quad f_X(x) \geq 0$$

$$\text{(ii)} \quad \int_{\mathbb{R}} f_X(x)dx = 1$$

$$\text{(iii)} \quad P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx$$

follow immediately from the definition and (4.3). The attributes of density functions can be further specified by designating their location or centrality, their spread or variability, their symmetry, and the contribution of tail behavior. In general, this information is provided by moments

$$\mathbb{E}(X^n) = \int_{\mathbb{R}} x^n f_X(x)dx$$

or central moments. For example, the mean

$$\mu = \mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)dx,$$

also termed the first moment or expected value, provides a measure of the density's central location whereas the second central moment

$$\sigma^2 = \text{var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{\mathbb{R}} (x - \mu)^2 f_X(x)dx \tag{4.4}$$

provides a measure of the density's variability or width. This typically is termed the variance of X and σ is called the standard deviation. One often employs the relation

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2$$

which results directly from (4.4). We note that the third moment (skewness) quantifies the density's symmetry about μ whereas the fourth moment (kurtosis) quantifies the magnitude of tail contributions.

Important Distributions for Inference and Model Calibration

We summarize next properties of the univariate normal, uniform, chi-squared, Student's t, beta, gamma, inverse-gamma and inverse chi-squared distributions which are important for frequentist and Bayesian inference and model calibration.

Definition 4.9 (Normal Distribution). In uncertainty quantification, a commonly employed univariate density is the normal density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \ , \ -\infty < x < \infty.$$

The associated cumulative distribution function is

$$F_X(x) = \int_{-\infty}^{x} f(s)ds = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

where the error function is defined to be

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-s^2}ds.$$

The notation $X \sim N(\mu, \sigma^2)$ indicates that the random variable X is normally distributed with mean μ and variance σ^2. For the normal density, 68.29% of the area is within 1σ of the mean μ and 95.45% is within 2σ as illustrated in Figure 4.2(a).

Definition 4.10 (Continuous Uniform Distribution). A random variable X is uniformly distributed on the interval $[a, b]$, denoted $X \sim \mathcal{U}(a, b)$, if any value in the interval is achieved with equal probability. The pdf and cdf are thus

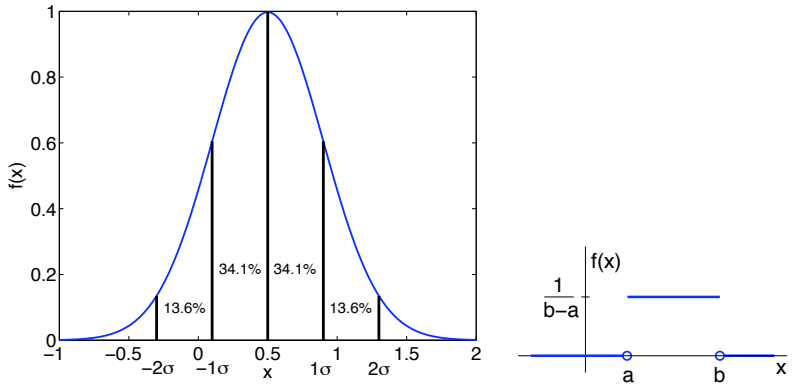
$$f_X(x) = \frac{1}{b-a}\chi_{[a,b]}(x) \tag{4.5}$$

and

$$F_X(x) = \begin{cases} 0 & , \quad x < a \\ \frac{x-a}{b-a} & , \quad a \le x < b \\ 1 & , \quad x \ge b \end{cases} \tag{4.6}$$

where the characteristic function $\chi_{[a,b]}(x)$ is defined to be unity on the interval $[a, b]$ and 0 elsewhere. The pdf is plotted in Figure 4.2(b). It is established in Exercise 4.1 that the mean and variance of X are

$$\mathbb{E}(X) = \frac{a+b}{2} \ , \ \text{var}(x) = \frac{(b-a)^2}{12} \tag{4.7}$$



(a) (b)

Figure 4.2. *(a) Normal density with $\mu = 0.5$ and $\sigma = 0.4$ and areas within 1σ and 2σ of μ. (b) Uniform density on the interval $[a, b]$.*

and the relationship between $X \sim \mathcal{U}(a,b)$ and $Z \sim \mathcal{U}(-1,1)$ is established in Exercise 4.6. When prior information is lacking, it is often assumed that model parameters have a uniform density.

Definition 4.11 (Chi-Squared Distribution). Let $X \sim N(0,1)$ be normally distributed. The random variable $Y = X^2$ then has a chi-squared distribution with 1 degree of freedom, denoted $Y \sim \chi^2(1)$. Furthermore, if $Y_i, i = 1, \cdots, k$, are independent $\chi^2(1)$ random variables, then their sum $Z = \sum_{i=1}^{k} Y_i$ is a χ^2 random variable with k degrees of freedom, denoted $Z \sim \chi^2(k)$ or $Z \sim \chi^2_k$. The probability density function

$$f_Z(z;k) = \begin{cases} \frac{z^{k/2-1}e^{-z/2}}{2^{k/2}\Gamma(k/2)} & , \quad z \geq 0 \\ 0 & , \quad z < 0 \end{cases} \tag{4.8}$$

can be compactly expressed in terms of the gamma function, where $\Gamma(k/2) = \sqrt{\pi}\frac{(k-2)!!}{2^{(k-1)/2}}$ for odd k, and exhibits the behavior shown in Figure 4.3(a). The mean and variance of Z are

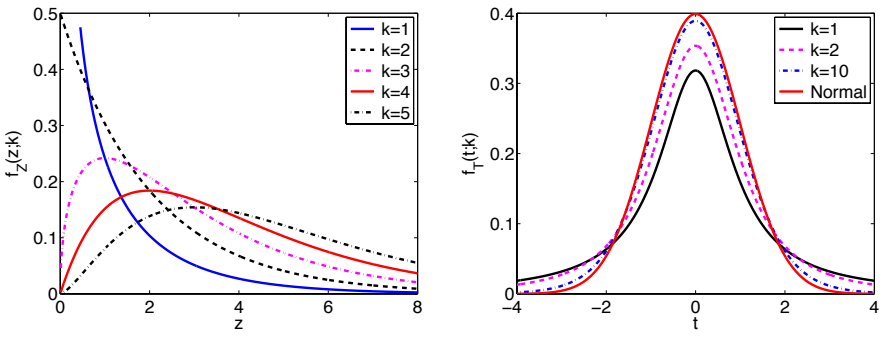
$$\mathbb{E}(Z) = k \ , \ \text{var}(Z) = 2k.$$

Chi-squared distributions naturally arise when evaluating the sum of squares error between measured data and model values when estimating model parameters.

Definition 4.12 (Student's *t*-Distribution). Let $X \sim N(0,1)$ and $Z \sim \chi^2(k)$ be independent random variables. The random variable

$$T = \frac{X}{\sqrt{Z/k}}$$

has a Student's t-distribution (or simply t-distribution) with k degrees of freedom.



(a) (b)

Figure 4.3. *(a) Chi-squared density for $k = 1, \cdots, 5$ and (b) Student's t-density with $k = 1, 2, 10$ compared with the normal density with $\mu = 0, \sigma = 1$.*

The probability density function can be expressed as

$$f_T(t;k) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$

where Γ again denotes the gamma function. Note that

$$f_T(t;1) = \frac{1}{\pi(1+t^2)}$$

is a special case of the Cauchy distribution. As illustrated in Figure 4.3(b), the density is symmetric and bell-shaped, like the normal density, but exhibits heavier tails.

It will be shown in Section 7.2 that the t-distribution naturally arises when estimating the mean of a population when the sample size is relatively small and the population variance is unknown.

On a historic note, aspects of this theory were developed by William Sealy Gosset, an employee of the Guinness brewery in Dublin, in an effort to select optimally yielding varieties of barley based on relatively small sample sizes. To improve perception following the recent disclosure of confidential information by another employee, Gosset was only allowed to publish under the pseudonym "Student." The importance of his work was advocated by both Karl Person and R.A. Fisher.

Definition 4.13 (Gamma Distribution). The gamma distribution is a two-parameter family with two common parameterizations: (i) shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ or (ii) shape parameter α and inverse scale or rate parameter $\beta = 1/\lambda$. We employ the second since the inverse-gamma distribution formulated in terms of α and β is a conjugate prior for likelihoods associated with normal distributions with known mean and unknown variance; see Example 4.69. For $X \sim \text{Gamma}(\alpha, \beta)$, the density is

$$f_X(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \ , \ x > 0,$$

and the expected value and variance are $\mathbb{E}(X) = \alpha/\beta$ and $\text{var}(X) = \alpha/\beta^2$.

In MATLAB, random values from a gamma distribution can be generated using the command `gamrnd.m` which uses the first parameterization based on the shape and scale parameters α and λ.

We point out that the one-parameter χ^2_k distribution with k degrees of freedom is a special case of the gamma distribution with $\alpha = \frac{k}{2}$ and $\beta = \frac{1}{2}$.

Definition 4.14 (Inverse-Gamma Distribution). If X has a gamma distribution, then $Y = X^{-1}$ has an inverse-gamma distribution with parameters that satisfy

$$X \sim \text{Gamma}(\alpha,\beta) \Leftrightarrow Y \sim \text{Inv-gamma}(\alpha,\beta). \tag{4.9}$$

Hence the density is

$$f_Y(y;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y} \ , \ y > 0,$$

and the mean and variance are $\mathbb{E}(Y) = \frac{\beta}{\alpha-1}$ for $\alpha > 1$ and $\text{var}(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$.

As noted in Definition 4.13 and illustrated in Example 4.69, the inverse-gamma distribution is the conjugate prior for normal likelihoods that are functions of the variance. The equivalence (4.9) can be used to generate random inverse-gamma values using the MATLAB Statistics Toolbox command `gamrnd.m`. Since $x = \texttt{gamrnd}(\alpha, \lambda)$ is parameterized in terms of the scale parameter, one would employ the command $y = \texttt{gamrnd}(\alpha, \beta)$, with $\beta = 1/\lambda$, to generate realizations of $Y \sim \text{Inv-gam}(\alpha, \beta)$. A technique to construct random realizations from the inverse-gamma distribution, if `gamrnd.m` is not available, is discussed at the end of this section.

Definition 4.15 (Inverse Chi-Squared Distribution). The inverse chi-squared distribution is a special case of Inv-gamma(α, β) with $\alpha = \frac{k}{2}, \beta = \frac{1}{2}$ so the density is

$$f_Y(y; k) = \frac{2^{-k/2}}{\Gamma(k/2)} y^{-(k/2+1)} e^{-1/2y}$$

for $y > 0$. This reparameterization can facilitate manipulation of conjugate families when constructing Bayesian posterior distributions.

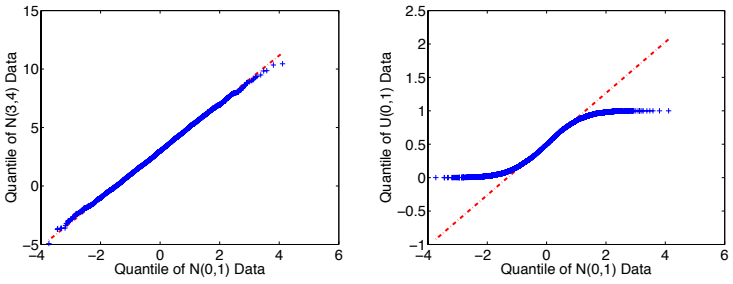
Definition 4.16 (Beta Distribution). The random variable $X \sim \text{Beta}(\alpha, \beta)$ has a beta distribution if it has the density

$$f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

for $x \in [0, 1]$. As illustrated in Example 4.68, it is the conjugate prior for the binomial likelihood. It is observed that if $\alpha = \beta = 1$, the beta distribution is simply the uniform distribution which is often used to provide noninformative priors. Realizations from the beta distribution can be generated using the MATLAB command `betarnd.m`.

Definition 4.17 (Quantile-Quantile (Q-Q) Plots). A Q-Q plot is a graphical method for comparing data from two distributions by plotting their quantiles against each other. We will typically use this to determine the degree to which data is Gaussian but the technique can be used to compare any distributions. If distributions are linearly related, Q-Q plots will be approximately linear. In MATLAB, Q-Q plots can be generated using the command `qqplot.m`.

To illustrate, we compare in Figure 4.4 realizations from $N(3, 4)$ and $\mathcal{U}(0, 1)$ distributions with data from a $N(0, 1)$ distribution. The linearity in the first case illustrates that the two are from the same family whereas the quantiles differ significantly in the comparison between the uniform and normal data.



(a) (b)

Figure 4.4. *Q-Q plot for (a) $N(3,4)$ and (b) $\mathcal{U}(0,1)$ data as compared with $N(0,1)$ data.*

Kernel Density Estimation

When estimating parameter densities in Chapter 8, we will determine the frequency with which values occur at the n points x_i. From this, we wish to compute density values $f_X(x)$ at arbitrary points x in the sample space. We consider non-parametric estimation procedures that do not pre-assume a parametric form for the density.

In principle, this can be achieved from a histogram of the computed values as illustrated in Figure 4.5(a). After dividing the sample space into a set of N bins, the density is approximated using the relation

$$\tilde{f}(x) = \frac{1}{N} \frac{\text{Number of } x_i \text{ in same bin as } x}{\text{Width of bin}}.$$

Whereas this approach is simple to implement in one dimension, it has the following disadvantages: the choice of bin locations and numbers can determine the structure of the density, and it is difficult to implement in multiple dimensions.

Instead, one often employs kernel density estimation (kde) techniques in which densities are formulated in terms of known kernel functions as shown in Figure 4.5(b). In 1-D, kernel density representations have the form

$$\tilde{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4.10}$$

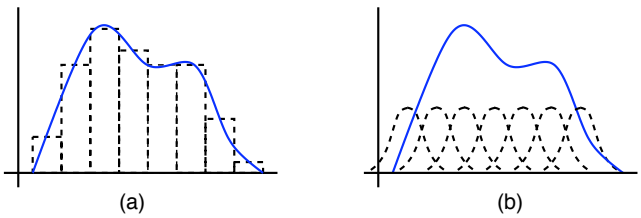


Figure 4.5. *(a) Histogram and approximating density. (b) Kernel basis function and kernel density estimate.*

where K is a specified, symmetric, pdf (e.g., normal) and h is a smoothing parameter termed the bandwidth [37,218]. Representations in higher dimensions are analogous.

If one has access to the MATLAB Statistics Toolbox, the function `ksdensity.m` can be employed to construct kernel density estimates. Alternatively, the functions `kde.m` and `kde2d.m`, which implement automatic bandwidth selection, are available from the MATLAB Central File Exchange.

Inverse Transform Sampling

In Definition 4.14, we discussed the use of the function `gamrnd.m`, from the MATLAB Statistics Toolbox, to construct random realizations from the inverse-gamma distribution. Here we summarize a technique to construct realizations of a general continuous random variable X with absolutely continuous distribution function $F_X(x)$.

For $U \sim \mathcal{U}(0,1)$, we assume that we have a random number generator capable of generating realizations of U. We define the random variable $Y = F_X^{-1}(U)$ which has the same distribution as X since

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X^{-1}(U) \leq y) \\ &= P(U \leq F_X(y)) \\ &= F_X(y). \end{aligned} \tag{4.11}$$

To generate a realization x of X, we generate a realization u of U and define

$$x = F_X^{-1}(u).$$

One typically computes $F_X^{-1}(u)$ using numerical algorithms. Even for an arbitrarily fine mesh, the cost of this procedure is typically low.

This technique can be used in lieu of calling `gamrnd.m` if the MATLAB Statistics Toolbox is unavailable.

4.1.2 Multiple Random Variables

For most applications, we have multiple parameters, responses and measurements with each being represented by a random variable. We discuss here multiple random variables with associated distributions.

Definition 4.18 (Random Vector). Let $X_1, \cdots, X_n$ be random variables. The vector $X : \Omega \to \mathbb{R}^n$ given by $X = [X_1, X_2, \cdots, X_n]$ is termed a random vector.

Definition 4.19 (Joint CDF). For a random vector X, the associated joint cdf $F_X : \mathbb{R}^n \to [0,1]$ is defined by

$$F_X(x_1, \cdots, x_n) = P\{\omega \in \Omega | X_j(\omega) \leq x_j\} \ , \ j = 1, \cdots, n$$

which is often written $F_X(x) = P\{X_1 \leq x_1, \cdots, X_n \leq x_n\}$.

Consider now the random variables $X_1, \cdots, X_n$ each having an expectation $\mathbb{E}(X_i)$. It follows immediately that

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i) \tag{4.12}$$

where $a_1, \cdots, a_n$ are real constants. Furthermore, if the n random variables are independent, then

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\cdots\mathbb{E}(X_n). \tag{4.13}$$

Definition 4.20 (Covariance and Correlation). The covariance of random variables X and Y is the number

$$\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{4.14}$$

and the correlation or correlation coefficient is

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{4.15}$$

We note that if X and Y are independent, then $\text{cov}(X,Y) = \rho_{XY} = 0$ and the random variables are uncorrelated. The converse is not true in general since the relation (4.15) quantifies only linear dependencies among random variables.

Returning to the case of n random variables, it is shown in [95] that

$$\text{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2\sum_{i<j} a_i a_j \text{cov}(X_i, X_j) \tag{4.16}$$

which simplifies to

$$\text{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{var}(X_i) \tag{4.17}$$

if the random variables are pairwise uncorrelated.

Theorem 4.21. Let $X_1, \cdots, X_n$ be mutually independent, normally distributed, random variables with $X_i \sim N(\mu_i, \sigma_i^2)$ and let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be fixed constants. As proven in Corollary 4.6.2 of [60], it then follows that

$$Z = \sum_{i=1}^{n}(a_i X_i + b_i) \sim N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right). \tag{4.18}$$

Like the univariate normal, the multivariate normal distribution plays a central role in uncertainty quantification and model validation.

Definition 4.22 (Multivariate Normal Distribution). The random n-vector X is said to be normally distributed with mean $\mu = [\mu_1, \cdots, \mu_n]$ and covariance matrix

$$V = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}(X_n) \end{bmatrix}, \tag{4.19}$$

designated $X \sim N(\mu, V)$, if the associated density is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left[-\frac{1}{2}(x-\mu)V^{-1}(x-\mu)^T \right].$$

Here $x = [x_1, x_2, \cdots, x_n]$ and $|V|$ is the determinant of V.

We use the next theorem when constructing proposal functions for the MCMC algorithms detailed in Chapter 8.

Theorem 4.23. Let $Y = [Y_1, \cdots, Y_n]^T$ be a normally distributed random vector, $Y \sim N(\mu, V)$, where V is positive definite. Let $Z \sim N(0, I_n)$ where I_n is the $n \times n$ identity. Then $Y = (RZ + \mu)$ where $V = RR^T$ and R is a lower triangular matrix.

A proof of this theorem can be found in [95]. We note that the decomposition $V = RR^T$ can be efficiently computed using a Cholesky decomposition.

Finally, the concepts of marginal and conditional distributions and densities will play an important role in statistical inference. We summarize the definitions for continuous random variables and refer the reader to [110, 168] for analogous definitions for discrete random variables.

Definition 4.24 (Marginal PDF). Let X_1 and X_2 be jointly continuous random variables with joint pdf $f_X(x_1, x_2)$. The marginal density functions of X_1 and X_2 are respectively given by

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2) dx_2 \quad , \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2) dx_1.$$

A representative marginal density is plotted in Figure 4.6(a). Similarly for jointly continuous random variables $X_1, \cdots, X_n$ with joint density function $f_X(x_1, \cdots, x_n)$, the marginal pdf of X_1 is

$$f_{X_1}(x_1) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_X(x_1, x_2, \cdots, x_n) dx_2 \cdots dx_n.$$

Definition 4.25 (Conditional PDF). Let X_1 and X_2 be jointly continuous random variables with joint pdf $f_X(x_1, x_2)$ and marginal pdf $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. The conditional density of X_1 given $X_2 = x_2$ is

$$f_{X_1|X_2}(x_1|x_2) = \begin{cases} \frac{f_X(x_1,x_2)}{f_{X_2}(x_2)} & , \quad f_{X_2}(x_2) > 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

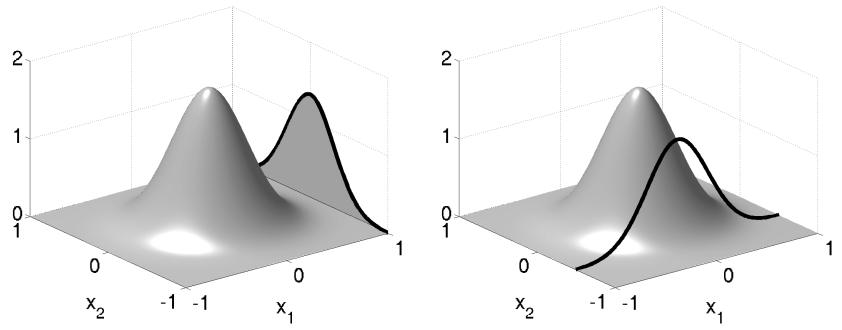


Figure 4.6. *(a) Marginal density $f_{X_2}(x_2)$ and (b) conditional density $f_{X_1|X_2}(x_1|x_2)$ at $x_2 = -\frac{1}{2}$ for a normal joint density $f_X(x_1, x_2)$ with covariance matrix $V = 0.09I$.*

as plotted in Figure 4.6(b). We note that $f_{X_1|X_2}(x_1|x_2)$ is a function of x_1. The definition for $f_{X_2|X_1}(x_2|x_1)$ is analogous. Similarly, for n jointly continuous random variables $X_1, \cdots, X_n$ with joint density function $f_X(x_1, \cdots, x_n)$ and marginal density $f_{X_1}(x_1)$, the conditional pdf of $X_2, \cdots, X_n$ given $X_1 = x_1$ is

$$f_{X_2,\cdots,X_n|X_1}(x_2, \cdots, x_n|x_1) = \frac{f_X(x_1, x_2, \cdots, x_n)}{f_{X_1}(x_1)}.$$

Definition 4.26 (iid Random Variables). Random variables $X_1, \cdots, X_n$ are said to be independent and identically distributed (iid) with pdf $f(x)$ if they are mutually independent and the marginal pdf f_{X_i} for each X_i is the same function $f(x) = f_{X_1}(x) = \cdots f_{X_n}(x)$. The joint pdf for iid random variables is

$$f_X(x_1, \cdots x_n) = \prod_{i=1}^{n} f(x_i). \tag{4.20}$$

4.2 Estimators, Estimates and Sampling Distributions

In this section, we summarize concepts pertaining to the estimation of unknown parameters through samples, observations, or measurements. In Section 4.3, we will detail specific techniques to estimate parameters in the context of model calibration. More general theory pertaining to frequentist and Bayesian inference is provided in Section 4.8.

Definition 4.27 (Point and Interval Estimates). Consider a fixed but unknown parameter $q \in \mathbb{Q} \subset \mathbb{R}^p$. A point estimate is a vector in $\mathbb{R}^p$ that represents q. An interval estimate provides an interval that quantifies the plausible location of components of q. The mean, median, or mode of a sampling distribution are examples of point estimates whereas confidence intervals are interval estimates.

Definition 4.28 (Estimator and Sampling Distribution). An estimator is a rule or procedure that specifies how to construct estimates for q based on random samples $X_1, \cdots, X_n$. Hence the *estimator is a random variable* with an associated distribution, termed the *sampling distribution*, which quantifies attributes of the estimation process. The *estimate is a realization of the estimator* so it is a function of the realized values $x_1, \cdots, x_n$. An estimator is said to be *unbiased* if its mean is equal to the value of the parameter being estimated. Otherwise it is said to be biased. Two estimators that we will employ for model calibration are *ordinary least squares* and *maximum likelihood* estimators. We will also employ mean, variance and interval estimators at various points in the discussion.

Definition 4.29 (Statistic). A statistic is a measurable function of one or more random variables that does not depend on unknown parameters.

Example 4.30. Let $X_1, \cdots, X_n$ be random variables associated with a sample of size n. Suppose we wish to estimate the population mean μ and variance σ^2 which are assumed unknown. This can be accomplished using the estimators, or statistics,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4.21}$$

which are the sample mean and variance. We note that we employ $n-1$ rather than n in the expression for S^2 to ensure that it is unbiased. If we additionally assume that $X_i \sim N(\mu, \sigma^2)$, it is illustrated in [168] that the sampling distributions for $\bar{X}$ and S^2 are

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad , \quad S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1). \tag{4.22}$$

Definition 4.31 (Interval Estimator and Confidence Interval). The goal when constructing an interval estimate is to determine functions $q_L(x)$ and $q_R(x)$ that bound the location $q_L(x) < q < q_R(x)$ of q based on realizations $x = [x_1, \cdots, x_n]$ of a random sample $X = [X_1, \cdots, X_n]$. The random interval $[q_L(X), q_R(X)]$ is termed an *interval estimator*. An interval estimator in combination with a confidence coefficient is commonly called a *confidence interval*. The confidence coefficient can be interpreted as the frequency of times, in repeated sampling, that the interval will contain the target parameter q. The $(1-\alpha)\times 100\%$ confidence interval is the pair of statistics $(q_L(X), q_R(X))$ such that for all $q \in \mathbb{Q}$,

$$P[q_L(X) \leq q \leq q_R(X)] = 1 - \alpha. \tag{4.23}$$

Example 4.32. Consider a sequence of n random variables $X_1, \cdots, X_n$ from a normal distribution with known variance σ^2 and unknown mean μ; that is, $X_i \sim N(\mu, \sigma^2)$. To determine information about the unknown mean, we consider the sample mean $\bar{X}$ given by (4.21) which has the sampling distribution given in (4.22).

It thus follows that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ so that

$$P\left(-2 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2\right) \approx 0.9545$$

since 95.45% of the area of a normal distribution lies within 2 standard deviations of the mean. This implies that

$$P\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) \approx 0.9545.$$

Here $[\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n}]$ is an *interval estimator* for μ where both endpoints are statistics since σ^2 is considered known. A $(1-\alpha)\times 100\%$ confidence interval is $[\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}]$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the realized sample mean based on n measurements, or realizations, x_i of the random variables X_i.

Example 4.33. We now turn to the problem of determining the confidence interval for the mean μ of a normal distribution when the variance σ^2 is also unknown. To estimate σ^2, we employ the statistic S^2 given by (4.21) which has the χ^2 distribution (4.22). We thus have

$$X = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1) \quad , \quad Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

so that the quotient

$$T = \frac{X}{\sqrt{Z/(n-1)}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S}$$

has a t-distribution with $n-1$ degrees of freedom; see Definition 4.12. To determine a $(1-\alpha)\times 100\%$ confidence interval for a given value of n, we seek values a and b such that

$$P\left(a < \frac{\sqrt{n}(\bar{X}-\mu)}{S} < b\right) = 1-\alpha.$$

In Figure 4.3(b), it is shown that the t-distribution is symmetric so that $b = -a$ which we denote by $t_{n-1,1-\alpha/2}$ to reflect the $n-1$ degrees of freedom and interval $1-\alpha/2$. It then follows that

$$P\left(\bar{X} - \frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}} < \mu < \bar{X} + \frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}}\right) = 1-\alpha.$$

On can employ standard tables of t-distributions to determine $t_{n-1,1-\alpha/2}$ given α and n and thus specify the $(1-\alpha)\times 100\%$ confidence interval $[\bar{X} - t_{n-1,1-\alpha/2}S/\sqrt{n}\,,$ $\bar{X} + t_{n-1,1-\alpha/2}S/\sqrt{n}\,]$. We remind the reader that for $\alpha = 0.05$, this is a random interval that has a 95% chance of containing the unknown but fixed (deterministic) parameter μ. The interval is constructed by obtaining measurements $x_1, \cdots, x_n$ and employing the realizations $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ to obtain

$$\left[\bar{x} - \frac{t_{n-1,1-\alpha/2}\, s}{\sqrt{n}}\,,\ \bar{x} + \frac{t_{n-1,1-\alpha/2}\, s}{\sqrt{n}}\right].$$

We will use t-distributions in this manner in Chapter 7 to construct confidence intervals for model parameters determined using least squares estimators when σ is unknown and the degrees of freedom is relatively small.

4.3 Ordinary Least Squares and Maximum Likelihood Estimators

The process of model calibration entails estimating model parameters, and possibly initial and boundary conditions, based on measured data. More generally, the estimation of model parameters, based on observations, comprises a significant component of statistical inference which is further discussed in Section 4.8.

To motivate, consider the statistical model

$$\Upsilon_i = f(t_i, q_0) + \varepsilon_i \quad , \quad i = 1, \cdots, n \tag{4.24}$$

where Υ_i are random variables whose realizations υ_i are a set of n measurements from an experiment and $f(t_i, q)$ is the parameter-dependent model response or quantity of interest at corresponding times. The random variables ε_i account for errors between the model and measurements. Finally, q_0 denotes the true, but unknown, parameter value that we cannot measure directly but instead must infer from realizations of the random variables Υ_i. We emphasize that in this context, q_0 is *not* a random variable.

4.3.1 Ordinary Least Squares (OLS) Estimator

Consider (4.24) with the assumption that errors ε_i are independent and identically distributed (iid), unbiased so $\mathbb{E}(\varepsilon_i) = 0$, and have true but unknown variance $\text{var}(\varepsilon_i) = \sigma_0^2$. We assume that the true parameter q_0 is in an admissible parameter space $\mathbb{Q}$ and we let $\mathcal{Q}$ denote the corresponding sample space. As illustrated in the examples of Chapter 7, these spaces typically coincide.

The ordinary least squares estimator and estimate[1]

$$\begin{aligned} \hat{q}_{OLS} &= \underset{q \in \mathcal{Q}}{\text{argmin}} \sum_{i=1}^{n} [\Upsilon_i - f(t_i, q)]^2 \\ q_{OLS} &= \underset{q \in \mathbb{Q}}{\text{argmin}} \sum_{i=1}^{n} [\upsilon_i - f(t_i, q)]^2 \end{aligned} \tag{4.25}$$

are the random variable and realization in $\mathbb{R}^p$ that minimize the respective sum of squares errors as illustrated in Figure 4.7(a). Details regarding the distribution of $\hat{q}_{OLS}$ based on various assumptions regarding the distribution of the errors ε_i are provided in Chapter 7.

[1]The use of the notation $\hat{q}_{OLS}$ to indicate the estimator is not universal and many texts denote the least squares estimate by the hat-notation. Hence care must be taken to establish the convention employed in the specific text.

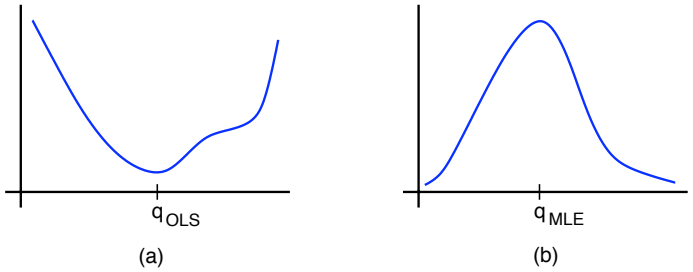


Figure 4.7. *(a) Ordinary least squares solution q_{OLS} to (4.25) and (b) maximum likelihood estimate q_{MLE} given by (4.27).*

4.3.2 Maximum Likelihood Estimator (MLE)

Maximum likelihood estimators can also be used to achieve the objective of estimating a parameter q based on random samples $\Upsilon_1, \cdots, \Upsilon_n$.

Definition 4.34 (Likelihood Function). Let $f_\Upsilon(\upsilon; q)$ be a parameter-dependent joint pdf associated with a random vector $\Upsilon = [\Upsilon_1, \cdots, \Upsilon_n]$, where $q \in \mathbb{Q}$ is an unknown parameter vector, and let $\upsilon = [\upsilon_1, \cdots, \upsilon_n]$ be a realization of Υ. The likelihood function $L : \mathbb{Q} \to [0, \infty)$ is defined by

$$L_\upsilon(q) = L(q|\upsilon) = f_\Upsilon(\upsilon; q) \tag{4.26}$$

where the observed sample υ is fixed and q varies over all admissible parameter values. The notation $L_\upsilon(q)$ is somewhat nonstandard but it highlights the fact that the independent variable is q. Some authors use the notation

$$L(q) = L(q|d) = f_\Upsilon(d; q),$$

where $d = [d_1, \cdots d_n]$ denotes the outcome from a random experiment, to reinforce this concept.

We note that because L is function of q, it is *not* a probability density function and the notation $L(q|\upsilon)$, while standard, should not be interpreted as a conditional pdf. If Y is discrete, then $L_\upsilon(q)$ is the probability of obtaining the data υ for a given parameter value q. For continuous Υ, the fact that L is only defined to within a constant of proportionality can be combined with Riemann sum approximations of the integral to obtain a similar interpretation.

For n iid random variables, it follows from (4.20) that the likelihood function is

$$L(q|\upsilon) = \prod_{i=1}^{n} f_\Upsilon(\upsilon_i; q).$$

Finally, we denote the log-likelihood function by

$$\ell_\upsilon(q) = \ell(q|\upsilon) = \ln L(q|\upsilon).$$

Example 4.35. Consider the binomial distribution with probability of success p. The probability mass function

$$f_\Upsilon(\upsilon; p, n) = P(\Upsilon = \upsilon | n, p) = \binom{n}{\upsilon} p^\upsilon (1-p)^{n-p}$$

quantifies the probability of obtaining exactly $\upsilon = 0, 1, \cdots, n$ successes in a sequence of n experiments. In this function, p and n are known and υ is unknown. Although the likelihood

$$L(q|\upsilon, n) = \binom{n}{\upsilon} p^\upsilon (1-p)^{n-p}$$

has the same functional form, the independent variable now is p, and υ and n are known. Hence the likelihood function is continuous whereas the probability mass function is discrete.

Estimates for q_0 are commonly constructed by computing the value of q that maximizes the likelihood which is termed a *maximum likelihood estimate (MLE).* For iid samples, the maximum likelihood estimate is

$$q_{MLE} = \underset{q \in \mathbb{Q}}{\operatorname{argmax}} \prod_{i=1}^{n} f_\Upsilon(\upsilon_i | q).$$

To illustrate, we consider (4.24) with the assumption that errors are iid, unbiased, and normally distributed with true but unknown variance σ_0^2 so that $\varepsilon_i \sim N(0, \sigma_0^2)$ and hence $\Upsilon_i \sim N(f(t_i, q_0), \sigma_0^2)$. In this case q and σ^2 are both parameters so the likelihood function is

$$\begin{aligned} L(q, \sigma | \upsilon) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-[\upsilon_i - f(t_i, q)]^2 / 2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n} [\upsilon_i - f(t_i, q)]^2 / 2\sigma^2} \end{aligned} \tag{4.27}$$

and the maximum likelihood estimate is

$$q_{MLE} = \underset{\substack{q \in \mathbb{Q} \\ \sigma^2 \in (0, \infty)}}{\operatorname{argmax}} L(q, \sigma | \upsilon) \tag{4.28}$$

as depicted in Figure 4.7(b).

Due to the monotonicity of the logarithm function, maximizing $L(q, \sigma|\upsilon)$ is equivalent to maximizing the log likelihood

$$\ell(q, \sigma | \upsilon) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [\upsilon_i - f(t_i, q)]^2 .$$

From a computational perspective, however, the log likelihood is advantageous so it is commonly employed in algorithms. For fixed σ^2, the condition $\frac{d}{dq}\ell(\upsilon|q, \sigma) = 0$ yields

$$\sum_{i=1}^{n} [\upsilon_i - f(t_i, q)] \nabla f(t_i, q) = 0 \tag{4.29}$$

where ∇f denotes the gradient of f with respect to q. It is observed that with the assumption of iid, unbiased, normally distributed errors, the maximum likelihood solution q_{MLE} to (4.29) is the same as the least squares estimate q_{OLS} specified by (4.25). The equivalence between minimizing the sum of squares error and maximizing the likelihood will be utilized when we construct proposal functions for the MCMC techniques in Chapter 8.

In frequentist inference, the maximum likelihood estimate q_{MLE} is the *parameter value that makes the observed output most likely.* It should *not* be interpreted as the *most likely* parameter value resulting from the data since this would require it to be a random variable which contradicts the tenets of frequentist analysis.

4.4 Modes of Convergence and Limit Theorems

There are several modes of convergence for sequences of random variables and distributions that are important for our discussion. We summarize the definitions and refer the reader to [60, 80, 81] for additional details, examples and proofs of related theorems.

Definition 4.36 (Convergence in Probability). A sequence $X_1, X_2, \cdots$ of random variables converges in probability to a random variable X, written as $X_n \xrightarrow{P} X$, if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0 \text{ or, equivalently, } \lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1.$$

Note that $X_1, X_2, \cdots$ are typically not iid in this and following definitions. This mode of convergence is weaker than *almost sure convergence.*

Definition 4.37 (Almost Sure Convergence). A sequence $X_1, X_2, \cdots$ of random variables converges almost surely to a random variable X, written $X_n \xrightarrow{a.s.} X$, if for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1.$$

Examples of sequences that converge in probability but not almost surely are provided in [60]. This is sometimes referred to as convergence with probability 1.

Definition 4.38 (Convergence in Distribution). Let $X_1, X_2, \cdots$ be a sequence of random variables with corresponding distributions $F_{X_1}(x), F_{X_2}(x), \cdots$. If $F_X(x)$ is a distribution function and

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous, then X_n is said to have a limiting random variable X with distribution function $F_X(x)$. In this case, X_n is said to converge in distribution to X, which is often written as $X_n \xrightarrow{D} X$. Care must be taken when using this notation since the convergence of random variables is defined in terms

of the convergence of the distributions. Hence this mode of convergence is quite different from the previous two.

We note that almost sure convergence implies convergence in probability which in turn implies convergence in distribution. Hence convergence in distribution is the weakest of the three concepts.

Definition 4.39 (Consistent Estimator). A sequence $\hat{q}_n$ of estimators is said to be consistent, or weakly consistent, if it converges in probability to the value q_0 of the parameter being estimated. In practice, we often construct estimators that are a function of the sample size n. In this case, the estimator is consistent if the sequence converges in probability to q_0 as the number of samples tends to infinity.

Law of Large Numbers and Central Limit Theorem

The *Law of Large Numbers* and *Central Limit Theorem* are two of the pillars of probability theory. To motivate them, we consider the problem of estimating the unknown mean μ and variance σ^2 of a population based on samples $x_1, x_2, \cdots$ and associated random variables $X_1, X_2, \cdots$. An estimator for the mean is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.30}$$

so a natural question is the following: Does $\lim_{n\to\infty} \bar{X}_n = \mu$? This is addressed by the strong and weak laws of large numbers.

Theorem 4.40 (Strong Law of Large Numbers). Let $X_1, X_2, \cdots$ be iid random variables with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$ and define $\bar{X}_n$ by (4.30). Then for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \varepsilon\right) = 1 \text{ or } \bar{X}_n \xrightarrow{a.s.} \mu.$$

The formulation of the weak Law of Large Numbers is similar except $\bar{X}_n \xrightarrow{P} \mu$. These laws are of fundamental importance since they establish that the random sample adequately represents the population in the sense that $\bar{X}_n$ converges to the mean μ.

Given the central role of the sample mean, it is natural to question the degree to which its sampling distribution can be established. In Example 4.30, we noted that if $X_i \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N(\mu, \sigma^2/n)$. The requirement of normally distributed random variables is quite restrictive, however, so we relax this assumption and pose the same question in the context of iid random variables from an arbitrary distribution. The remarkable answer is provided by the Central Limit Theorem.

Theorem 4.41 (Central Limit Theorem). Let $X_1, \cdots, X_n$ be iid random variables with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Furthermore, let $\bar{X}_n$ be given by (4.30) and let $G_n(x)$ denote the cdf of the random variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then

$$\lim_{n\to\infty} G_n = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

so that the limiting distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is a normal distribution $N(0,1)$. The theorem is often expressed as

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z$$

where $Z \sim N(0,1)$.

Because

$$\bar{X}_n \xrightarrow{D} \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \tag{4.31}$$

$\bar{X}_n$ is approximately normal for sufficiently large n. This result is similar to that noted in Example 4.30 for $X_i \sim N(\mu, \sigma^2)$ but with the major difference that (4.31) holds in an asymptotic sense for X_i from an *arbitrary distribution* as long as n is sufficiently large.

From a broad perspective, the combination of the Law of Large Numbers and Central Limit Theorem establishes that for sufficiently large n, samples are representative of the population (in the sense of the means) and the means of these samples behave asymptotically as normal distributions. The question as to how large n must be to ensure this asymptotic behavior is problem dependent and the assumption of approximate normality can be questionable when sample sizes are small.

We will invoke the asymptotic normality provided by the Central Limit Theorem in Chapter 7 when constructing sampling distributions for model parameters.

4.5 Random Processes

In Section 4.1, we summarized the framework associated with random variables and random vectors. However, uncertainty quantification in the context of differential equation models can yield variables that exhibit time or space dependence in addition to randomness. This necessitates the discussion of stochastic or random processes and fields. We will also see that the Markov chain Monte Carlo (MCMC) techniques of Chapter 8 rely on the theory of stochastic processes.

To motivate our discussion of random processes, consider first the ODE

$$\begin{aligned} &\frac{du}{dt} = -\alpha(\omega)u \ , \ t > 0 \\ &u(0,\omega) = \beta(\omega) \end{aligned} \tag{4.32}$$

where α and β are random variables and $\omega \in \Omega$ is an event in an underlying probability space. It was noted in Example 3.1 that for every time instance t, the random solution $u(t,\omega)$ is an example of a stochastic or random process.

Now consider the partial differential equation

$$\begin{aligned} &\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(x,\omega)\frac{\partial T}{\partial x}\right) = f(t,x) \quad , \ -1 < x < 1 \ , \ t > 0 \\ &T(t,-1) = T_\ell \ , \ T(t,1) = T_r \quad , \ t \geq 0 \\ &T(0,x) = T_0(x) \quad , -1 \leq x \leq 1 \end{aligned} \tag{4.33}$$

which, as detailed in Example 3.5, models the flow of heat u in a structure having uncertain diffusivity α. Here α is an example of a random field and the solution $T(t, x, \omega)$ is random for all pairs (t, x) of independent variables.

Definition 4.42 (Stochastic Process). A stochastic or random process is an indexed collection

$$X = \{X_t, t \in \mathbb{T}\} = \{X(t), t \in \mathbb{T}\}$$

of random variables, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. The index set is typically assumed to be totally ordered and often is taken to be time. Taking $\mathbb{T}$ to be a subset of consecutive integers yields a discrete random process whereas taking $\mathbb{T}$ to be an interval of real numbers yields a continuous process.

The random solution $u(t, \omega)$ to (4.32) is an example of a continuous random process. In the next section, we will devote significant discussion to Markov chains, which are discrete random processes, since they are central to the MCMC methods used in the Bayesian analysis of Chapter 8 to quantify parameter densities.

Other ordered index sets can be considered including spatial points or intervals. However, the ordering in dimensions greater than one is complicated so we employ the terminology stochastic or random fields for spatially varying quantities.

A stochastic process can be interpreted three ways.

(i) X is a function on $\mathbb{T} \times \Omega$ with the realization $X_t(\omega)$ for $t \in \mathbb{T}$ and $\omega \in \Omega$;

(ii) For fixed $t \in \mathbb{T}$, X_t is a random variable;

(iii) For an outcome $\omega \in \Omega$, the realization $X_t(\omega)$ is a function of t that is often called the sample path or trajectory associated with ω.

We note that continuous stochastic processes are infinite-dimensional and extreme care must be taken when extending finite-dimensional convergence results to these cases. The following class of random processes is important since the concepts of mean, covariance and correlation functions are well-defined for these processes.

Definition 4.43 (Second-Order Stochastic Process). A second-order stochastic process is one for which $\mathbb{E}(X_t^2) < \infty$ for all $t \in \mathbb{T}$.

For second-order random processes, the random variable concepts of mean and covariance can be directly extended using the interpretation (ii). Specifically, the expectation and covariance functions of X are defined as

$$\begin{aligned} &\mu(t) = \mathbb{E}(X_t) \ , \ t \in \mathbb{T} \\ &C(t, s) = \operatorname{cov}(X_t, X_s) = \mathbb{E}\left[(X_t - \mu(t))(X_s - \mu(s))\right] \ , \ t, s \in \mathbb{T}. \end{aligned} \tag{4.34}$$

Hence $\mu(t)$ quantifies the centrality of sample paths whereas $C(t, s)$ quantifies their variability about $\mu(t)$.

Definition 4.44 (Gaussian Process). A Gaussian process (GP) is a continuous-time stochastic process X such that all finite-dimensional vectors $X_t = [X_{t_1}, \cdots, X_{t_n}]$ have a multivariate normal distribution; that is

$$X_t \sim N(\mu(t), C(t))$$

where $t = [t_1, \cdots, t_n], \mu(t) = [\mathbb{E}(X_{t_1}), \cdots, \mathbb{E}(X_{t_n})]$ and $[C(t)]_{ij} = \text{cov}(X_{t_i}, X_{t_j})$ for all $1 \leq i, j \leq n$. A Gaussian process is thus a probability distribution for a function.

The concept of stationarity is important in the theory of Markov chains since it provides criteria specifying when MCMC methods can be expected to converge to posterior distributions for parameters. We consider this in the context of a discrete index set $\mathbb{T}$ but note that a similar definition holds for continuous index sets.

Definition 4.45 (Stationary Random Process). The random process X is said to be stationary if, for any $t_1, t_2, \cdots, t_n \in \mathbb{T}$ and s such that $t_1+s, \cdots, t_n+s \in \mathbb{T}$, the random vectors $[X_{t_1}, \cdots, X_{t_n}]$ and $[X_{t_1+s}, \cdots, X_{t_n+s}]$ have the same distribution. For a stationary process, $\mu(t)$ is constant for all $t = [t_1, \cdots, t_n]$ and $C(t, s) = C(t-s)$ is a function only of the time difference $|t - s|$.

Definition 4.46 (Autoregressive (AR) Models). An AR(1) process, or time series, X satisfies

$$X_t = \rho_1 X_{t-1} + \varepsilon_t \ , \ \varepsilon_t \sim N(0, \sigma^2) \tag{4.35}$$

where ρ_1 is a parameter. If $|\rho_1| < 1$, the process is said to be wide-sense stationary. In this case, $\mathbb{E}(X_t) = \mathbb{E}(X_{t-1})$ so that $\mathbb{E}(X_t) = 0$ and $\text{var}(X_t) = \mathbb{E}(X_t^2) = \rho_1^2 \mathbb{E}(X_{t-1}^2) + \sigma^2$ so that $\text{var}(X_t^2) = \frac{\sigma^2}{1-\rho_1^2}$. We note that an AR(1) process smooths the output in the sense of a low-pass filter.

An AR(p) process satisfies

$$X_t = \sum_{k=1}^{p} \rho_k X_{t-k} + \varepsilon_t \ , \ \varepsilon_t \sim N(0, \sigma^2). \tag{4.36}$$

We note that AR(p) processes are a type of Gaussian process.

Definition 4.47 (Random Field). The concept of a random field generalizes that of a random process by allowing indices that are vector-valued or points on a manifold. Specifically, a random field is a collection

$$X = \{X_x, x \in \mathcal{X}\}$$

of random variables indexed by elements x in a topological space $\mathcal{X}$. For our applications, we will employ random fields to quantify uncertain spatially-varying parameters such as $\alpha(x, \omega)$ in (4.33).

For the definitions of random processes and random fields, we have considered indexed families of random variables which, for fixed values of the index, map Ω

to $\mathbb{R}$. When describing Markov processes, however, it is advantageous to generalize this concept to include random variables that map into a state space S. This is established in the following definitions.

Definition 4.48 (S-Valued Random Variable). Let S be a finite or countable set termed the *state space*. An S-valued random variable is a function $X : \Omega \to S$ such that $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in S$. Note that this is exactly Definition 2.3 if $S = \mathbb{R}$.

Definition 4.49. A random process X is said to have a state space S if X_t is an S-valued random variable for each $t \in \mathbb{T}$.

4.6 Markov Chains

In Chapter 8, we will employ Markov chain Monte Carlo (MCMC) methods to construct posterior densities for model parameters. We summarize here the fundamental properties of Markov chains necessary for that development.

Broadly stated, a stochastic process is said to satisfy the Markov property if the probability of future states is dependent only on the present state rather than the sequence of past events that precede it. This is completely analogous to the state space concept of modeling in which a system is defined in terms of state variables that uniquely define the behavior at time t. When combined with dynamics encompassed in the model, the future state behavior can be completely defined. Both Markov processes and state space models are memoryless in the sense that the past history is not required to make future predictions. Whereas Markov processes can be defined for both continuous and discrete index sets $\mathbb{T}$, we focus solely on the latter since it provides the setting necessary for MCMC analysis. Discrete-time Markov processes are usually called *Markov chains* although some authors also use this designation for continuous time processes.

Definition 4.50 (Markov Chain). A Markov chain is a sequence of S-valued random variables

$$X = \{X_i, i \in \mathbb{Z}\}$$

that satisfy the Markov property that X_{n+1} depends only on X_n; that is

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n) \qquad (4.37)$$

where x_i is the state of the chain at time i.

A Markov chain is characterized by three components: a state space S, an initial distribution p^0, and a transition or Markov kernel. As indicated in Definition 4.48, the state space is the range of all random variables so it is the set of all possible realizations. We assume a finite number k of discrete states so $S = \{x_1, \cdots, x_k\}$. The initial distribution quantifies the starting configuration for the chain whereas the transition kernel quantifies the probability of transitioning from state x_i to x_j so it establishes how the chain evolves. For our discussion,

we assume that the transition probabilities are the same for all time which yields a *homogeneous* Markov chain.

We let p_{ij} denote the probability of moving from x_i to x_j in one step so that

$$p_{ij} = P(X_{n+1} = x_j | X_n = x_i).$$

The resulting transition matrix is

$$P = [p_{ij}] \quad , \ 1 \leq i, j \leq k.$$

We will also be interested in the probability of transitioning between states in m-steps which we denoted by

$$p_{ij}^{(m)} = P(X_{n+m} = x_j | X_n = x_i)$$

with the corresponding m-step transition matrix

$$P_m = \left[p_{ij}^{(m)} \right] = P^m.$$

The initial density, which is often termed mass when it is discrete, is given by

$$p^0 = [p_1^0, \cdots, p_k^0]$$

where $p_i^0 = P(X_0 = x_i)$. Because p^0 and P contain probabilities, their entries are nonnegative and the elements of p^0 and rows of P must sum to unity. Matrices satisfying the property are termed *row-stochastic matrices*.

Given an initial distribution and transition kernel, the distribution after 1 step is $p^1 = p^0 P$ and

$$p^n = p^{n-1} P = p^0 P^n$$

after n steps. We illustrate these concepts in the next example.

Example 4.51. Various studies have indicated that factors such as weather, injuries, and unquantifiable concepts such as hitting streaks lend a random nature to baseball [7]. We assume that a team that won its previous game has a 70% chance of winning their next game and 30% chance of losing whereas a losing team wins 40% and loses 60% of their next games. Hence the probability of winning or losing the next game is conditioned on a team's last performance.

This yields the two-state markov chain illustrated in Figure 4.8 where

$$S = \{\text{win}, \text{lose}\}.$$

The resulting transition matrix is

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

There are 30 teams in major league baseball so

$$p^0 = [p_w^0 \ , \ p_\ell^0] \ , \ p_w^0 + p_\ell^0 = 1$$

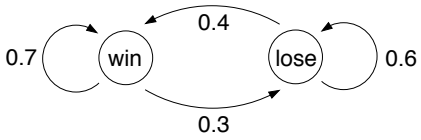


Figure 4.8. *Markov chain quantifying the probability of winning or losing based on the last performance.*

is the percentage of teams who won and lost their last games. To illustrate, we take $p^0 = [0.8, 0.2]$. We assume a schedule in which teams play at different times so p_w^0 and p_ℓ^0 do not both have to be 0.5.

The percentage of teams who win/lose their next game is given by

$$\begin{aligned} p^1 &= [0.8\ ,\ 0.2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \\ &= [0.64\ ,\ 0.32] \end{aligned}$$

so the distribution after n games is

$$p^n = [0.8\ ,\ 0.2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}^n.$$

The distributions for $n = 0, \cdots, 10$ are compiled in Table 4.1. These numerical results indicate that the distribution is limiting to a stationary value.

For this example, we can explicitly compute a limiting distribution π by solving the constrained relation

$$\begin{aligned} &\pi = \pi P \quad , \quad \sum \pi_i = 1 \\ &\Rightarrow [\pi_{win}\ ,\ \pi_{lose}] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [\pi_{win}\ ,\ \pi_{lose}] \quad , \quad \pi_{win} + \pi_{lose} = 1 \end{aligned}$$

to obtain

$$\pi = [0.5714\ ,\ 0.4286].$$

In general, however, we cannot solve explicitly for a stationary value and instead must establish the manner in which p^n limits to π. We next discuss the nature of this convergence and summarize criteria that guarantee the existence of a unique limiting value.

n	p^n	n	p^n	n	p^n
0	[0.8000 , 0.2000]	4	[0.5733 , 0.4267]	8	[0.5714 , 0.4286]
1	[0.6400 , 0.3600]	5	[0.5720 , 0.4280]	9	[0.5714 , 0.4286]
2	[0.5920 , 0.4080]	6	[0.5716 , 0.4284]	10	[0.5714 , 0.4286]
3	[0.5776 , 0.4224]	7	[0.5715 , 0.4285]		

Table 4.1. *Iteration and distributions for Example 4.51.*

As detailed in Section 4.4, it does not make sense to directly consider limits $\lim_{n\to\infty} X_n$ of random variables. Instead, we consider the limit

$$\lim_{n\to\infty} p^n = \pi$$

which is convergence in distribution. We note that if this limit exists, it must satisfy

$$\pi = \lim_{n\to\infty} p^0 P^n = \lim_{n\to\infty} p^0 P^{n+1} = \left(\lim_{n\to\infty} p^0 P^n\right) P = \pi P.$$

Definition 4.52 (Stationary Distribution). For a Markov chain with transition kernel P, distributions π that satisfy

$$\pi = \pi P \tag{4.38}$$

are termed equilibrium or stationary distributions of the chain. In a measure theoretic framework, π is an invariant measure.

For every finite Markov chain, there exists at least one stationary distribution. However, it may not be unique and it may not be equal to $\lim_{n\to\infty} p^n$. Criteria necessary to establish a unique limiting distribution $\pi = \lim_{n\to\infty} p^n$ are motivated by the following definitions and examples.

Definition 4.53 (Irreducible Markov Chain). A Markov chain is irreducible if any state x_j can be reached from any other state x_i in a finite number of steps; that is $p_{ij}^{(m)} > 0$ for all states in finite m. Otherwise it is reducible.

Example 4.54. Consider the Markov chain depicted in Figure 4.9(a) with the transition matrix

$$P = \begin{bmatrix} 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The chain is clearly reducible since $p_{3j} = 0$ for $j = 1, 2, 4$. Furthermore, it is easy to verify that $\pi = [0, 0, 1, 0]$ and $\pi = [0, 0, 0, 1]$ are both stationary distributions. The property of irreducibility is required to guarantee that π is unique.

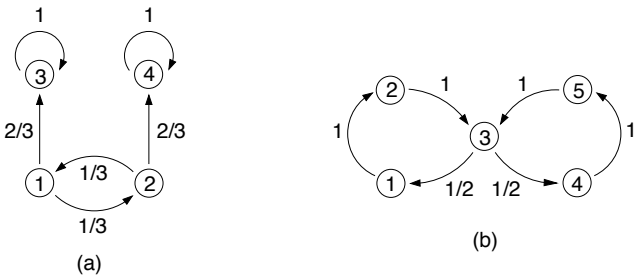


Figure 4.9. *(a) Reducible chain for Example 4.54, and (b) periodic chain for Example 4.56.*

Definition 4.55 (Periodic Markov Chain). A Markov chain is periodic if parts of the state space are visited at regular intervals. The period k is defined as

$$\begin{aligned} k &= \gcd\left\{m | \pi_{ii}^{(m)} > 0\right\} \\ &= \gcd\left\{m | P(X_{n+m} = x_i | X_n = x_i) > 0\right\}. \end{aligned}$$

The chain is aperiodic if $k = 1$.

Example 4.56. The Markov chain depicted in Figure 4.9(b) with the transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

has the unique stationary distribution $\pi = [1/6\,,\, 1/6\,,\, 1/3\,,\, 1/6\,,\, 1/6]$. It is established in Exercise 4.8 that if $p^0 = [1, 0, 0, 0, 0, 0]$, then $p^3 = p^6 = p^9 = \cdots = p^0$ so the period is $k = 3$. Because mass cycles through the chain at a regular interval, it does not converge so $\lim_{n\to\infty} p^n$ does not exist. Furthermore, it is demonstrated in Exercise 4.9 that if the limit of a periodic chain exists for one initial distribution, other distributions can yield different limits. Hence aperiodicity is required to guarantee that the limit exists.

For infinite chains, one must additionally include conditions regarding the persistence or recurrence of states. However, we will focus on finite Markov chains for which it can be shown that if the chain is irreducible, all states are positive persistent [119].

Before providing a theorem that establishes the convergence $\lim_{n\to\infty} p^n = \pi$, we summarize relevant results from matrix theory.

Definition 4.57. A $k \times k$ matrix A is

(i) nonnegative, denoted $A \geq 0$, if $a_{ij} \geq 0$ for all i, j

(ii) strictly positive, denoted $A > 0$, if $a_{ij} > 0$ for all i, j.

Theorem 4.58 (Perron–Frobenius). Let A be an $k \times k$ nonnegative matrix such that $A^m > 0$ for some $m \geq M$. Then

(i) A has a positive eigenvalue λ_0 with corresponding left eigenvector x_0 where the entries of x_0 are positive,

(ii) If $\lambda \neq \lambda_0$ is any other eigenvalue of A, then $|\lambda| < \lambda_0$,

(iii) λ_0 has geometric and algebraic multiplicity 1.

There are several statements of the Perron-Frobenius theorem, and details and proofs can be found in [119, 128, 217].

Theorem 4.59. For all finite stochastic matrices P, the largest eigenvalue is $\lambda_0 = 1$.

See [119] for a proof of this theorem.

Theorem 4.60. Let P be a finite transition matrix for an irreducible aperiodic Markov chain. Then there exists $M \geq 1$ such that $P^m > 0$ for all $m \geq M$.

Further details are provided in [119] and the theorem is illustrated in Exercise 4.10. The following theorem establishes the convergence of the Markov chain.

Theorem 4.61. Every finite, homogeneous Markov chain that is irreducible and aperiodic, with transition matrix P, has a unique stationary distribution π. Moreover, chains converge in the sense of distributions, $\lim_{n\to\infty} p^n = \pi$, for every initial distribution p^0.

Proof. It follows from Theorems 4.58–4.60 that the largest eigenvalue of P is $\lambda_0 = 1$ which has multiplicity 1. There is thus a unique left eigenvector π that satisfies $\pi P = \pi$ and $\sum \pi_i = 1$. To establish the convergence, we first consider the eigendecomposition

$$UPV = \Lambda = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_k \end{bmatrix}$$

where $1 > |\lambda_2| \geq \cdots \geq |\lambda_k|$ and $V = U^{-1}$. It follows that

$$\lim_{n\to\infty} P^n = \lim_{n\to\infty} V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^n & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_k^n \end{bmatrix} U = V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & 0 \end{bmatrix} U.$$

Furthermore, we observe that $UP = \Lambda U$ implies that

$$\begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \begin{bmatrix} & P & \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix}$$

and $V = U^{-1}$ implies that

$$UV = \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \begin{bmatrix} 1 & \cdots & v_{1k} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

since $\sum \pi_i = 1$. This establishes that the first column of V is all ones. Finally

$$\begin{aligned}
\lim_{n\to\infty} p^n &= \lim_{n\to\infty} p^0 P^n \\
&= \lim_{n\to\infty} \left[p_1^0, \cdots, p_k^0\right] \begin{bmatrix} 1 & \cdots & v_{k1} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \\
&= \begin{bmatrix} p_1^0 & \cdots & p_k^0 \end{bmatrix} \begin{bmatrix} 1 & \cdots & v_{k1} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \\
&= [\pi_1, \cdots, \pi_k] \\
&= \pi
\end{aligned}$$

thus establishing the required convergence.

■

Theorem 4.61 establishes that finite Markov chains which are irreducible and aperiodic will converge to a stationary distribution π. However, it is often difficult or impossible to solve for π using the relations $\pi P = \pi$ subject to $\sum \pi_i = 1$. The detailed balance condition provides an alternative that is straight-forward to implement in MCMC methods where the goal is to construct Markov chains whose stationary distribution π is the posterior distribution for parameters.

Definition 4.62 (Detailed Balance). A chain with transition matrix $P = [p_{ij}]$ and distribution $\pi = [\pi_1, \cdots, \pi_k]$ is *reversible* if the detailed balance condition

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{4.39}$$

is satisfied for all i, j. Since

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_j p_{ji} = \pi_j,$$

it follows immediately that $\pi P = \pi$ so that reversibility implies stationarity. Hence if the chains are irreducible and aperiodic, they will uniquely limit to this specified stationary distribution. In Chapter 8, we use the Metropolis algorithm to construct chains that satisfy (4.39) and converge to the posterior density.

4.7 Random Versus Stochastic Differential Equations

We briefly illustrate here the difference between random differential equations, which we consider throughout this text, and stochastic differential equations. This is done in part to allay a growing trend in the UQ community to treat these terms as synonymous when in fact they are distinctly different and they require completely different techniques for analysis and approximation.

Definition 4.63 (Random Differential Equation). Random differential equations are those in which random effects are manifested in parameters, initial or boundary conditions, or forcing conditions that are regular (e.g., continuous) with respect to time and space. An example is the ODE

$$\frac{dz}{dt} = a(\omega)z + b(t,\omega)$$
$$z(0) = z_0(\omega)$$

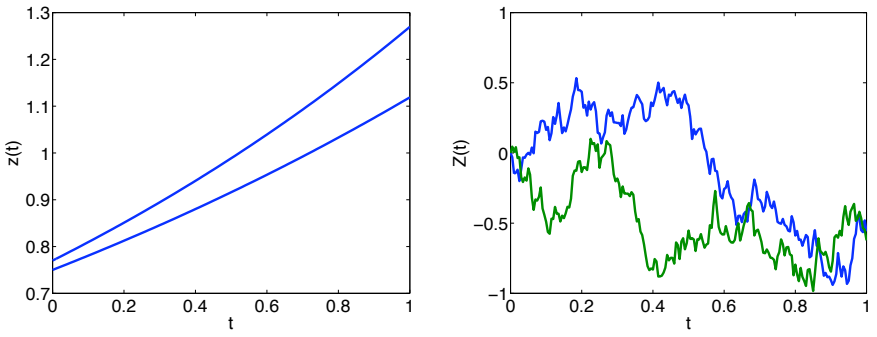
which has the solution

$$z(t;q) = e^{a(\omega)t}\left[z_0(\omega) + \int_0^t e^{-a(\omega)s} b(s,\omega))ds\right].$$

We emphasize that $b(t,\omega)$ is a random process, as defined in Definition 4.42, with the additional requirement that for an outcome $\omega \in \Omega$, the sample path $b(t,\omega)$ is taken to be smooth; e.g., in $C[0,t_f]$. This guarantees that sample paths of the solution $u(t,\omega)$ are at least differentiable functions as illustrated in Figure 4.10.

In summary, for each realization of ω, random differential equations are analyzed and solved sample path by sample path using the theory of standard differential equations [90, 136, 228]. The goal pursued in Chapters 9 and 10 is to determine distributions or uncertainty bounds for $u(t,\omega)$ based on those of inputs such as parameters or initial and boundary conditions.

Definition 4.64 (Stochastic Differential Equation). The role of uncertainty is fundamentally different in stochastic differential equations (SDE). In this case, the differential equations are forced by an irregular process such as a Wiener process or Brownian motion. Stochastic differential equations are typically written symbolically in terms of stochastic differentials but they are interpreted as Itō or Stratonovich stochastic integrals. For example, fluctuations in $Z(t)$ due to a Wiener



(a) (b)

Figure 4.10. *Realizations of (a) a random differential equation and (b) sample paths of a stochastic differential equation.*

process W could be formulated as

$$dZ(t) = -aZ(t)dt + bdW(t)$$

which is interpreted as

$$Z(t) = Z_0 - \int_0^t aZ(s)ds + \int_0^t bdW(s)$$

where the second integral is an Itō stochastic integral.

As illustrated in Figure 4.10, the solutions of SDE exhibit nondifferentiable sample paths due to the irregularity of the driving Wiener process. We do not further consider SDE in this text but rather include this definition to delineate them from random differential equations. The reader is referred to [90, 136] for further details about SDE.

4.8 Statistical Inference

The goal in statistical inference is to deduce the structure of, or make conclusions about, a phenomenon based on observed data. This often involves the determination of an unknown distribution based on observed data in which case the problem of statistical inference can be stated as follows. Given a set

$$S = \{x_1, \cdots, x_n\} \ , \ x_j \in \mathbb{R}^N$$

of observed realizations of a random variable X, we want to infer the underlying probability distribution that produces the data S.

Statistical inference can be roughly categorized as being *parametric* or *nonparametric* in nature. In parametric approaches, one assumes that the underlying distributions can be adequately described in terms of a parametric relation having a relatively small number of parameters; e.g., mean and variance. The inference problem is to estimate those parameters or the distribution of those parameters. This approach has the advantage of a typically small number of parameters but the disadvantage of limited accuracy if the assumed functional relation is incorrect. In nonparametric approaches, one does not presuppose a functional form but instead describes or constructs the distribution based solely on properties of the observations. This avoids errors associated with incorrect parametric relations but requires that some structure be imposed on algorithms to ensure that reasonable distributions are determined.

4.8.1 Frequentist Versus Bayesian Inference

Frequentist and Bayesian inference differ in the underlying assumptions made regarding the nature of probabilities, models, parameters, and confidence intervals. As detailed in [30], each approach, or a hybrid combination of the two, is advantageous for certain problems or applications. Hence it is necessary that scientists understand both.

From a frequentist perspective, probabilities are defined as the frequencies with which an event occurs if the experiment is repeated a large number of times. Hence they are objective and are not updated as data is acquired. Parameters are considered to be unknown but fixed; hence they are deterministic. To statistically establish confidence in the estimation process, one constructs estimators, such as ordinary least squares (OLS) or maximum likelihood estimators (MLE), to estimate the parameters in the manner detailed in Section 4.3. Based on either the assumption of normality for the errors or asymptotic theory resulting from the Central Limit Theorem, one can then construct sampling distributions and confidence intervals for the parameter estimators.

The interpretation of confidence intervals in the framework of frequentist inference is often a source of confusion. As detailed in Definition 4.31, a 90% confidence interval has the following interpretation: in repeated procedures, 90% of realized intervals would include the true parameter q_0. In model calibration, this means that if the estimation procedure is repeated 100 times using data having the same error statistics, and a 90% interval estimate is computed each time, then 90% of the intervals would include q_0 as illustrated in Figure 4.11(a). The sampling distribution and confidence intervals thus quantify the accuracy and variability of the estimation procedure rather than providing a density for the parameter. Hence they do not provide a direct measure of parameter uncertainty.

Because parameters are fixed, but unknown, values in this framework, it cannot be directly applied to obtain parameter densities that can be propagated through models to quantify model uncertainty. In some problems, the sampling distributions may be similar to parameter distributions but this needs to be verified either experimentally or using Bayesian analysis. This is discussed in more detail in Chapter 7.

Probabilities are treated as possibly subjective in the Bayesian framework and they can be updated to reflect new information. Moreover, they are considered to be a distribution rather than a single frequency value. Similarly, parameters are considered to be random variables with associated densities and the solution of the parameter estimation problem is the posterior probability density. The Bayesian perspective is thus natural for model uncertainty quantification since it provides densities that can be propagated through models. The interpretation of interval estimates, termed credible intervals, is also natural in the Bayesian framework.

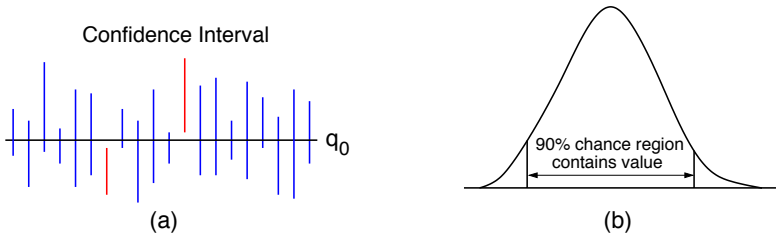


Figure 4.11. *Interpretation of a (a) frequentist 90% confidence interval and (b) Bayesian 90% credible interval.*

Definition 4.65 (Credible Interval). The $(1-\alpha)\times 100\%$ credible interval is that which has a $(1-\alpha)\times 100\%$ chance of containing the expected parameter. A 90% credible interval is illustrated in Figure 4.11(b).

We next provide details regarding Bayesian inference to provide the background necessary for Chapter 8.

4.8.2 Bayesian Inference

Bayesian inference is based on the supposition that probabilities, and more generally our state of knowledge regarding an observed phenomenon, can be updated as additional information is obtained. In the context of parametric models, parameters are treated as random variables having associated densities.

Because probabilities and parameter densities are conditioned on observations, Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

for probabilities provides a natural genesis for Bayesian inference. In the context of parameters $Q = [Q_1, \cdots, Q_p]$ that are quantified based on observations $\upsilon = [\upsilon_1, \cdots, \upsilon_n]$, one employs the relation

$$\pi(q|\upsilon) = \frac{\pi(\upsilon|q)\pi_0(q)}{\pi_\Upsilon(\upsilon)} \tag{4.40}$$

where $\pi_0(q)$ and $\pi(q|\upsilon)$ respectively denote the prior and posterior densities, $\pi(\upsilon|q)$ is a likelihood, and the marginal density $\pi_\Upsilon(\upsilon)$ is a normalization factor. Here $q = Q(\omega)$ denotes realizations of Q. We note that the subscripts which indicate specific random variables are typically dropped from the prior and posterior in Bayesian analysis.

The prior density $\pi_0(q)$ quantifies any prior knowledge that may be known about the parameter before data is taken into account. For example, one might have prior information based on similar previous models, data that is similar to previous data, or initial parameter densities that have been determined through other means such as related experiments.

For most model calibration, however, one does not have such prior information so one uses instead what is termed a *noninformative prior*. A common choice of noninformative prior is the uniform density, or unnormalized uniform, posed on the parameter support. For example, one might employ

$$\pi_0(q) = \chi_{[0,\infty)}(q),$$

for a positive parameter. This choice is improper in the sense that the integral of $\pi_0(q)$ is unbounded. It is recommended that a noninformative prior be used unless good previous information is known since it is shown in Example 4.66 that incorrect prior information can degrade (4.40) far more than a noninformative prior.

In "empirical Bayes" inference, one also encounters data-dependent priors in which priors estimated using frequentist techniques such as maximum likelihood are

employed in the Bayesian model. It is argued in [35] that this double use of data is problematic with small sample sizes and is at odds with the tenets of Bayesian analysis.

The term $\pi(\upsilon|q)$, which is a function of q with υ fixed, quantifies the likelihood $L(q|\upsilon)$ of observing υ given parameter realizations q as detailed in Section 4.3.2. We will illustrate various choices for the likelihood function in the examples at the end of this section and at the beginning of Chapter 8. The joint density is given by

$$\pi(q, \upsilon) = \pi(\upsilon|q)\pi_0(q)$$

and is normalized to unity by the marginal density function $\pi_{\Upsilon}(\upsilon)$ of all possible observations.

Finally, the posterior density $\pi(q|\upsilon)$ quantifies the probability of obtaining parameters q given observations υ. It is the posterior density that we will be estimating using the Bayesian parameter estimation techniques of Chapter 8 and we point out that the data directly informs the posterior only through the likelihood. Finally, representation of $\pi_{\Upsilon}(\upsilon)$ as the integral over all possible joint densities yields the Bayes relation

$$\pi(q|\upsilon) = \frac{\pi(\upsilon|q)\pi_0(q)}{\int_{\mathbb{R}^p} \pi(\upsilon|q)\pi_0(q)dq} \tag{4.41}$$

commonly employed for model calibration and data assimilation.

A significant issue, which will be discussed in detail in Chapter 8, concerns the evaluation of the normalizing integral. It can be analytically evaluated only in special cases and classical tensored quadrature techniques are effective only in low dimensions; e.g., $p \leq 4$. This has spawned significant research on high dimensional quadrature techniques including adaptive sparse grids for moderate dimensionality and Monte Carlo techniques for high dimensions; see Chapter 11.

Example 4.66. To illustrate (4.41) in a setting where the posterior density can be computed explicitly, we consider the results from tossing a possibly biased coin. The random variable

$$\Upsilon_i(\omega) = \begin{cases} 0 & , \quad \omega = T \\ 1 & , \quad \omega = H \end{cases}$$

represents the result from the i^{th} toss and the parameter q is the probability of getting heads. We now consider the probability of obtaining N_1 heads and N_0 tails in a series of $N = N_0 + N_1$ flips of the coin.

Because coin flips are independent events with only two possible outcomes, the likelihood of observing a sequence $\upsilon = [\upsilon_1, \cdots, \upsilon_N]$, given the probability q, is

$$\begin{aligned} \pi(\upsilon|q) &= \prod_{i=1}^{N} q^{\upsilon_i}(1-q)^{1-\upsilon_i} \\ &= q^{\sum \upsilon_i}(1-q)^{N-\sum \upsilon_i} \\ &= q^{N_1}(1-q)^{N_0} \end{aligned}$$

which is simply a scaled binomial density. We consider first a noninformative prior

$$\pi_0(q) = \begin{cases} 1 & , \quad 0 \leq q \leq 1 \\ 0 & , \quad \text{else} \end{cases}$$

which yields the posterior density

$$\pi(q|\upsilon) = \frac{q^{N_1}(1-q)^{N_0}}{\int_0^1 q^{N_1}(1-q)^{N_0} dq} = \frac{(N+1)!}{N_0!N_1!} q^{N_1}(1-q)^{N_0}.$$

We note that in this special case, the denominator is the integral of a beta function which admits an analytic solution. In general, however, quadrature techniques must be employed to approximate the integral.

For a fair coin with $q_0 = \frac{1}{2}$, the posterior densities associated with various realizations N_1 and N_0 are plotted in Figure 4.12. It is first observed that Bayesian inference yields a posterior density with just one experiment whereas frequentist analysis would specify a probability of either 0 or 1. It is also observed that the variability of $\pi(q|\upsilon)$ decreases as N increases. Finally, the manner in which the data informs the density is illustrated by comparing the results with 5 Heads, 9 Tails, which has a mode of 0.36, to those of 49 Heads, 51 tails which has a mode of 0.495. This illustrates that the method is achieving the goal of having the data inform when there is no prior information.

We next illustrate the effect of a poor choice for the prior density. For the same fair coin ($q_0 = \frac{1}{2}$), we consider the choice

$$\pi_0(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(q-\mu)^2/2\sigma^2}$$

with $\mu = 0.3$ and $\sigma = 0.1$. We cannot analytically evaluate the denominator in this case so we instead employ Gaussian quadrature. As illustrated in Figure 4.13, even for a realization of 50 Heads and 50 Tails, the mean of the posterior is still smaller than $q_0 = \frac{1}{2}$ but is significantly better than the result for 5 Heads and 5 Tails. This illustrates the manner in which a poor informative prior can have negative impact for a large number of observations. Hence if the validity of an informative prior is in doubt, it is recommended that a noninformative prior be used instead.

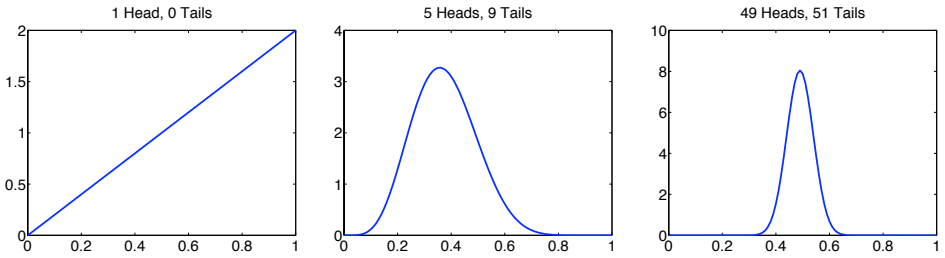


Figure 4.12. *Posterior densities associated with a noninformative prior for three realizations of the coin toss experiment.*

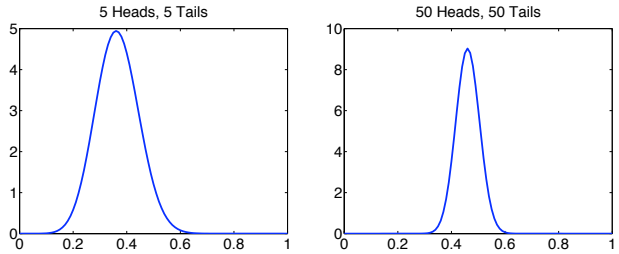


Figure 4.13. *Posterior densities associated with a poor informative prior for two realizations of the coin toss experiment.*

Conjugate Priors

Definition 4.67 (Conjugacy). The property that the prior and posterior distributions have the same parametric form is termed conjugacy. When this occurs, the prior $\pi_0(q)$ is termed a conjugate prior for the likelihood $\pi(\upsilon|q)$. Parameters in the prior relation are often termed *prior hyperparameters* to distinguish them from the model parameters q. The corresponding parameters in the posterior relation are called *posterior hyperparameters.*

The use of conjugate priors, when possible, is advantageous since closed form expressions for the posterior are then available. This will be used when estimating densities for measurement errors in Chapter 8.

Example 4.68. Consider the binomial model

$$\pi(\upsilon|q) = q^{N_1}(1-q)^{N-N_1} \ , \ N_1 = \sum_{i=1}^{N} \upsilon_i$$

used for the likelihood in the coin toss Example 4.66. We observe that if the prior is parameterized similarly, the product of the prior and likelihood will be in the same family. Specifically, we take $\pi_0(q)$ to be a beta density with hyperparameters α and β so that $\pi_0(q) \propto q^{\alpha-1}(1-q)^{\beta-1}$ as shown in Definition 4.16. It then follows that the posterior density satisfies

$$\begin{aligned}\pi(q|\upsilon) &\propto q^{N_1}(1-q)^{N-N_1} q^{\alpha-1}(1-q)^{\beta-1} \\ &= q^{N_1+\alpha-1}(1-q)^{N-N_1+\beta-1}\end{aligned}$$

so it is a beta density with shape parameters $N_1+\alpha$ and $N-N_1+\beta$. The beta prior distribution is thus a conjugate family for the binomial likelihood.

Example 4.69. Here we consider normally distributed random variables with known mean μ and unknown variance σ^2. This will illustrate techniques employed in Chapter 8 to estimate the unknown variance σ_0^2 of measurement errors. As detailed in Section 4.3.2, the likelihood of observing $\upsilon = [\upsilon_1, \cdots, \upsilon_N]$ iid measurements under

these assumptions is

$$\pi(\upsilon|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS/2\sigma^2}$$

where the sum of squares error is

$$SS = \sum_{j=1}^{n} (\upsilon_j - \mu)^2.$$

This likelihood is in the inverse-gamma family, defined in Definition 4.14, so the conjugate prior is $\pi_0(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{\beta/\sigma^2}$. The posterior density can then be expressed as

$$\begin{aligned}\pi(\sigma^2|\upsilon) &\propto \pi_0(\sigma^2)\pi(\upsilon|\sigma^2) \\ &\propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2} (\sigma^2)^{-n/2} e^{-SS/2\sigma^2} \\ &= (\sigma^2)^{-(\alpha+1+n/2)} e^{-(\beta+SS/2)/\sigma^2}\end{aligned}$$

so that

$$\sigma^2|\upsilon \sim \text{Inv-gamma}(\alpha + n/2, \beta + SS/2).$$

As shown in Definitions 4.13 and 4.14, if $X \sim \text{Gamma}(\alpha, \beta)$, then $Y = X^{-1} \sim \text{Inv-gamma}(\alpha, \beta)$. This equivalence can be exploited so that the MATLAB command `gamrnd.m` can be used to generate random numbers from a gamma distribution which can then be used to construct random values from an inverse-gamma distribution.

4.9 Notes and References

This chapter provides an overview of statistical topics that play a role in uncertainty quantification and we necessarily leave details to the following references. The text [60] provides a very accessible introduction to probability, point and interval estimation, hypothesis testing, analysis of variance and linear regression with clearly stated definitions. The texts [110, 168] are also excellent sources for obtaining an overview of probability and statistics at an upper undergraduate level. The book [95] delineates the difference between estimators and estimates by using different notation and is an excellent source for details regarding linear regression. Finally, [80, 81] are classics in the field of probability.

There are a number of excellent supplemental texts on random processes and Morkov chains including [98, 119, 124, 128, 156, 181, 251]. Additional theory, examples, and numerical algorithms for random and stochastic differential equations can be found in [90, 136, 183, 228]. We note that due to the mathematical nature of the underlying framework required for stochastic differential equations, these latter texts also provide a measure-theoretic framework for random variables and other concepts discussed in this chapter. Additional details regarding a measure theoretic basis for aspects of this material can be found in [36].

The reader is referred to [56, 221] for introductory concepts and examples regarding Bayesian analysis and computing and [34, 91] for a more in-depth treatment