1

Introduction

1.1 Statistical Inference

Probability is the way we quantify uncertainty. It is based on three axioms that develop the whole probability theory. We suggest the book Mood et al. (1974) to review the main concepts.

Statistics is the science of data. It is the science of collecting, exploring, presenting and making decisions from data. As a science, it is divided into two branches: descriptive statistics and inferential statistics. The former involves sampling and exploration, whereas the latter deals with decision making, which includes estimation, hypothesis testing and predictions.

To understand the previous definitions and some other concepts involved in statistics, let X denote a characteristic of interest that is measurable for individuals in a particular population. For instance, X could be the income, height or age of a person in a particular population (school, county, country, etc.). The *statistical population* is the collection of all possible values x_i for the individuals $i = 1, 2, \ldots$ of the population. In notation, $Pop = \{x_1, x_2, \ldots\}$, where the population size could be infinite. If we were able to obtain all possible values for the whole population, we could summarise them in a relative frequency table and plot it in a histogram, like the one depicted in Figure 1.1. If we make the histogram bins narrower and take the limit as the length of the bins goes to zero, by appropriately dividing by the bins' length we obtain a smooth curve such as the one plotted on top of the histogram in Figure 1.1. Let us denote this curve mathematically as $f(x \mid \theta_0)$, which is a function of x, the possible values of the variable of interest, and θ_0, which is the true population parameter. In other words, the population is fully characterised by the curve, that is, $Pop \iff f(x \mid \theta_0)$. The curve is actually a probability model or a density function for a random variable X indexed by the parameter θ_0.

In statistics, it is not common to have access to all population values, but usually we have access to a subset of them that we call a *sample*.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

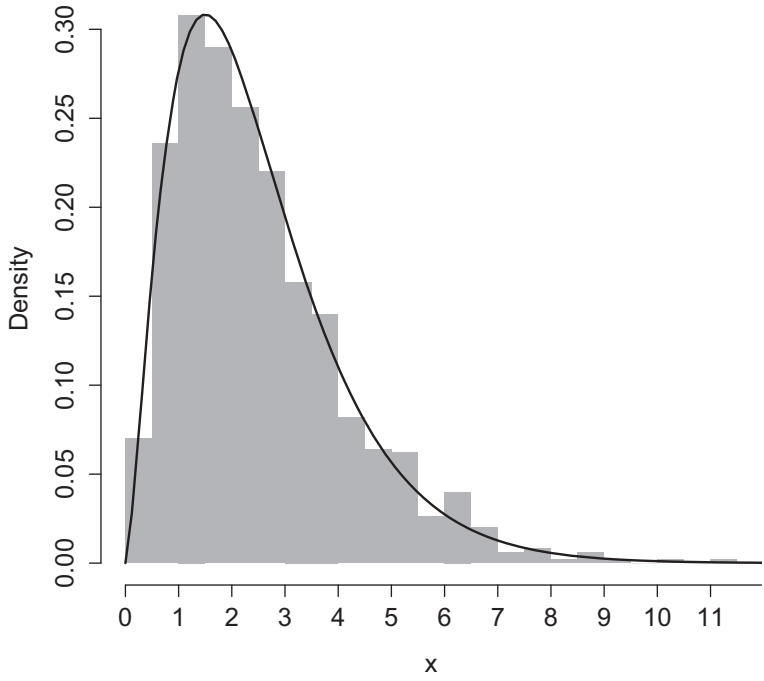

Figure 1.1 Histogram (shaded) and probability curve (solid line) for simulated data.

Formally, a sample is a finite collection of size $n < \infty$ of the characteristic of interest $\{X_1, X_2, \ldots, X_n\}$. Since these values are unknown to the researcher beforehand, they can be assumed to be (conditionally) independent random variables whose possible values are determined by the probability model $f(x \mid \theta)$, where the population parameter is usually unknown but belongs to a specific parameter space Θ, that is, $\theta_0, \theta \in \Theta$. For instance, for the data depicted in Figure 1.1, a possibility would be to assume a gamma model, namely $X \sim \text{Ga}(\alpha, \beta)$, where $\theta = (\alpha, \beta) \in \Theta = (\mathbb{R}^+)^2$.

There are two main approaches for statistical inference: classical or frequentist, and Bayesian. Within either inferential approach we could assume two possibilities for the population: a parametric assumption like the one we mentioned earlier where $X \sim f(x \mid \theta)$ and $\theta \in \Theta$; or a non-parametric assumption where the population is not characterised by a parametric model, namely $X \sim f(x)$ with $f \in \mathcal{F}$ and $\mathcal{F}$ the space of all probability models. This leads to four types of inferential procedures, which are summarised in Table 1.1.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

Assumption\Procedure	Frequentist	Bayesian
Parametric	(1)	(3)
Non-parametric	(2)	(4)

Table 1.1 *Types of inferential procedures.*

We now briefly describe the generalities of each inferential procedure with a little more emphasis on (3) since most of the ideas discussed in this book belong to that context.

(1) **Frequentist Parametric**: The assumptions here are that observed data $\mathbf{X} = \{X_1, \ldots, X_n\}$ are a sample from population $X_i \sim f(x \mid \theta)$ of independent random variables where $\theta \in \Theta$. Sample information about θ is summarised in the joint distribution function $f(\mathbf{x} \mid \theta)$, which in the case of independent data is given by $\prod_{i=1}^{n} f(x_i \mid \theta)$, where, if seen as a function of θ, it is called likelihood. The frequentist inferential procedure is based entirely on likelihood, usually through maximisation. See, for example, Mood et al. (1974).

(2) **Frequentist Nonparametric**: The assumptions here are that observed data $\mathbf{X} = \{X_1, \ldots, X_n\}$ are a sample from population $X_i \sim f(x)$ of independent random variables and $f \in \mathcal{F}$. Sample information about f, or F, the corresponding cumulative distribution function (CDF), is summarised as $f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i)$, which is the likelihood for f (and F). For instance, the maximum likelihood estimator (MLE) of F is the empirical distribution function $\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$, where $I_A(x)$ denotes the indicator function of set A that takes the value of one if $x \in A$ and zero otherwise. See, for example, Conover (1999).

(3) **Bayesian Parametric**: The assumptions here are that observed data $\mathbf{X} = \{X_1, \ldots, X_n\}$ are a sample from population $X_i \mid \theta \sim f(x \mid \theta)$ of conditional independent random variables and $\theta \in \Theta$. The word *conditional* is included because the Bayesian inferential procedure depends on an axiomatic theory that establishes that all unknown quantities must be quantified using the researcher's prior (uncertain) knowledge through $f(\theta)$. This prior knowledge is updated with the observed data through Bayes's theorem which states that

$$f(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta) f(\theta)}{f(\mathbf{x})}, \tag{1.1}$$

where $f(\mathbf{x} \mid \theta)$ is the likelihood for θ and $f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x} \mid \theta) f(\theta)$ or $f(\mathbf{x}) = \sum_{\theta \in \Theta} f(\mathbf{x} \mid \theta) f(\theta)$ is a normalising constant. Bayes's theorem is

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

therefore a learning rule and $f(\theta \mid \mathbf{x})$ is called the *posterior distribution* for θ that contains all available information. To make decisions, the axiomatic theory establishes that preferences on consequences must be quantified by a utility (or loss) function which must be maximised (or minimised) after marginalising all uncertain quantities using the prior or posterior distribution, whichever is available. For instance, if we want to estimate θ with $\widehat{\theta}$ we could represent our preferences via a quadratic loss function $v(\widehat{\theta}, \theta) = a(\widehat{\theta} - \theta)^2$ for $a > 0$. If the posterior distribution is available, we obtain the expected loss as $\bar{v}(\widehat{\theta}) = \mathrm{E}\{v(\widehat{\theta}, \theta)\} = \int_{\Theta} a(\widehat{\theta} - \theta)^2 f(\theta \mid \mathbf{x}) \mathrm{d}\theta$, from which, after minimisation, we obtain $\widehat{\theta} = \mathrm{E}\{\theta \mid \mathbf{x}\}$; that is, our point estimate for θ is the posterior mean. See, for example, Bernardo and Smith (2000).

(4) **Bayesian Nonparametric**: The assumptions here are that observed data $\mathbf{X} = \{X_1, \ldots, X_n\}$ are a sample from population $X_i \mid \theta \sim f(x)$ of conditional independent random variables and $f \in \mathcal{F}$. The axiomatic theory establishes that the researcher must quantify prior knowledge on f or F via $\mathcal{P}(f)$ or $\mathcal{P}(F)$. This is usually done via stochastic processes whose paths are densities or distribution functions. The two most typical choices are the Dirichlet process with precision parameter c and centring measure F_0, denoted by $\mathcal{DP}(c, F_0)$, see Ferguson (1973); and the Pólya tree with precision parameter c, variance function ϱ and centring measure F_0, denoted as $\mathrm{PT}(c, \varrho, F_0)$; see for example, Nieto-Barajas and Núñez-Antonio (2021). This prior distribution is updated with the observed data through Bayes's theorem (1.1), but adapted to stochastic processes, to obtain the posterior law $F \mid \mathbf{x}$. If we further represent our preferences via a quadratic loss function, the posterior point estimate for F will be $\mathrm{E}(F \mid \mathbf{x})$, which is known as the posterior predictive function. See Hjort et al. (2010).

Therefore, statistical procedures can be summarised as shown in the diagram of Figure 1.2. The arrow pointing down corresponds to descriptive statistics, whereas the arrow pointing up corresponds to inferential statistics.

1.2 Common Probability Distributions

In the following chapters we will use several common probability distributions as well as their first two moments. We summarise them here.

Discrete Distributions

- *Bernoulli distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \theta^x (1 - \theta)^{1-x} I_{\{0,1\}}(x),$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

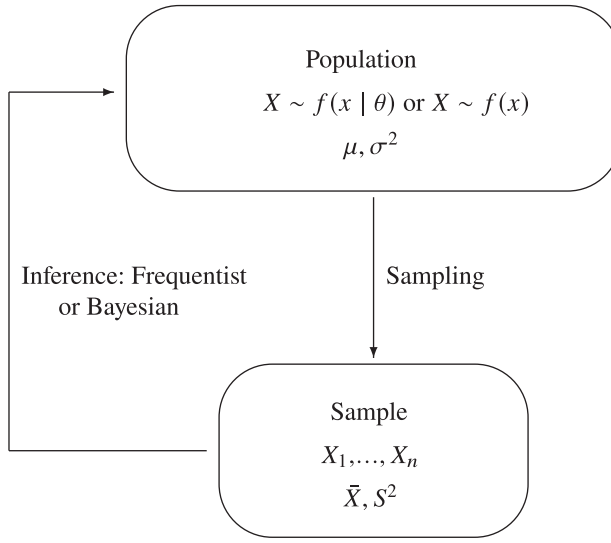

Figure 1.2 Diagram of statistics.

valid for $\theta \in (0, 1)$. This is denoted as $\text{Ber}(\theta)$. The first two moments are

$$\text{E}(X \mid \theta) = \theta \quad \text{and} \quad \text{Var}(X \mid \theta) = \theta(1 - \theta).$$

- *Binomial distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} I_{\{0,1,\ldots,n\}}(x),$$

with $\theta = (n, \pi)$ and valid for $\pi \in (0, 1)$ and $n \in \mathbb{N}$. This is denoted as $\text{Bin}(n, \pi)$. The first two moments are

$$\text{E}(X \mid \theta) = n\pi \quad \text{and} \quad \text{Var}(X \mid \theta) = n\pi(1 - \pi).$$

- *Geometric distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \theta(1 - \theta)^x I_{\{0,1,\ldots\}}(x),$$

valid for $\theta \in (0, 1)$. This is denoted as $\text{Geo}(\theta)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{(1 - \theta)}{\theta} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{(1 - \theta)}{\theta^2}.$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

- *Negative Binomial distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \binom{r+x-1}{x} \pi^r (1-\pi)^x I_{\{0,1,\ldots\}}(x),$$

with $\theta = (r, \pi)$ and valid for $\pi \in (0,1)$ and $r \in \mathbb{N}$. This is denoted as $\mathrm{NB}(r, \pi)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{r(1-\pi)}{\pi} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{r(1-\pi)}{\pi^2}.$$

- *Poisson distribution*: this is characterised by the following density:

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

valid for $\theta > 0$. This is denoted as $\mathrm{Po}(\theta)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \theta \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \theta.$$

- *Beta-Binomial distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} I_{\{0,\ldots,n\}}(x),$$

where $\Gamma(\cdot)$ is the gamma function that satisfies $\Gamma(a) = (a-1)\Gamma(a-1)$, with $\theta = (a, b, n)$ and valid for $a, b > 0$ and $n \in \mathbb{N}$. This is denoted as $\mathrm{BBin}(a, b, n)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{na}{a+b} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{nab(a+b+n)}{(a+b)^2(a+b+1)}.$$

- *Beta-Negative Binomial distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \binom{r+x-1}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+r)\Gamma(b+x)}{\Gamma(a+b+r+x)} I_{\{0,1,\ldots\}}(x),$$

with $\theta = (a, b, r)$ and valid for $a, b > 0$ and $r \in \mathbb{N}$. This is denoted as $\mathrm{BNB}(a, b, r)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{rb}{a-1}$$

if $a > 1$, and

$$\mathrm{Var}(X \mid \theta) = \frac{rb(a+r-1)(a+b-1)}{(a-1)^2(a-2)}$$

if $a > 2$.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

- *Gamma-Poisson distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+x)c^x}{x!(b+c)^{a+x}} I_{\{0,1,\ldots\}}(x),$$

with $\theta = (a, b, c)$ and valid for $a, b, c > 0$. This is denoted as $\text{Gpo}(a, b, c)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{ca}{b} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{ca(b+c)}{b^2}.$$

- *Multinomial distribution*: this is a multivariate distribution characterised by the following density:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = n! \prod_{j=1}^{k} \frac{\pi_j^{x_j}}{x_j!} I\left(\sum_{j=1}^{k} x_j = n\right),$$

with $\mathbf{x} = (x_1, \ldots, x_k)$, $x_j \in \mathbb{N}$, $\theta = (n, \boldsymbol{\pi})$ and valid for $\pi_j \in (0, 1)$, $\sum_{j=1}^{k} \pi_j = 1$ and $n \in \mathbb{N}$. This is denoted as $\text{Mult}(n, \boldsymbol{\pi})$. The first two moments are

$$\text{E}(X_j \mid \theta) = n\pi_j, \quad \text{Var}(X_j \mid \theta) = n\pi_j(1 - \pi_j) \quad \text{and}$$
$$\text{Cov}(X_i, X_j) = -n\pi_i\pi_j$$

for $i \neq j$.

- *Dirichlet-Multinomial distribution*: this is a multivariate distribution characterised by the following density:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\Gamma(n+1)\Gamma(a_0)}{\Gamma(a_0+n)} \prod_{j=1}^{k} \frac{\Gamma(a_j+x_j)}{\Gamma(a_j)\Gamma(x_j)} I\left(\sum_{j=1}^{k} x_j = n\right),$$

with $\mathbf{x} = (x_1, \ldots, x_k)$, $x_j \in \mathbb{N}$, $\boldsymbol{\theta} = (n, \mathbf{a})$ where $\mathbf{a} = (a_1, \ldots, a_k)$ and valid for $a_j > 0$ and $n \in \mathbb{N}$, with $a_0 = \sum_{j=1}^{k} a_j$. This is denoted as $\text{DMult}(\mathbf{a}, n)$. The first two moments are

$$\text{E}(X_j \mid \theta) = n\frac{a_j}{a_0}, \quad \text{Var}(X_j \mid \theta) = \frac{n(n+a_0)a_j(a_0-a_j)}{a_0^2(a_0+1)}$$

$$\text{and} \quad \text{Cov}(X_i, X_j) = -\frac{n(n+a_0)a_ia_j}{a_0^2(a_0+1)}$$

for $i \neq j$.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

Continuous Distributions

- *Uniform distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{1}{b-a} I_{(a,b)}(x),$$

with $\theta = (a, b)$ and valid for $a < b \in \mathbb{R}$. We denote it as $\mathrm{Un}(a, b)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{(b-a)^2}{12}.$$

- *Beta distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x),$$

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\mathrm{Be}(a, b)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

- *Inverse beta distribution*: this is also known as beta prime or beta of the second kind and is characterised by the following density:

$$f(x \mid \theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1+x)^{-a-b} I_{(0,\infty)}(x),$$

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\mathrm{Ibe}(a, b)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{a}{b-1} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{a(a+b-1)}{(b-2)(b-1)^2}$$

if $b > 1$ and $b > 2$, respectively.

- *Exponential distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \theta e^{-\theta x} I_{(0,\infty)}(x),$$

valid for $\theta > 0$. We denote it as $\mathrm{Exp}(\theta)$. The first two moments are

$$\mathrm{E}(X \mid \theta) = \frac{1}{\theta} \quad \text{and} \quad \mathrm{Var}(X \mid \theta) = \frac{1}{\theta^2}.$$

- *Gamma distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} I_{(0,\infty)}(x),$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\text{Ga}(a, b)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{a}{b} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{a}{b^2}.$$

- *Inverse gamma distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x} I_{(0,\infty)}(x),$$

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\text{Iga}(a, b)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{b}{a-1} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{b^2}{(a-1)^2(a-2)}$$

if $a > 1$ and $a > 2$, respectively.

- *Gamma-gamma distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{b^a \Gamma(a+c) x^{c-1}}{\Gamma(a)\Gamma(c)(b+x)^{a+c}} I_{(0,\infty)}(x),$$

with $\theta = (a, b, c)$ and valid for $a, b, c > 0$. We denote it as $\text{Gga}(a, b, c)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{cb}{a-1} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{b^2 c(a+c-1)}{(a-1)^2(a-2)}$$

if $a > 1$ and $a > 2$, respectively.

- *Normal distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \left(\frac{2\pi}{\tau}\right)^{-1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} I_{\mathbb{R}}(x),$$

with $\theta = (\mu, \tau)$ and valid for $\mu \in \mathbb{R}$ and $\tau > 0$. We denote it as $\text{N}(\mu, \tau)$. The first two moments are

$$\text{E}(X \mid \theta) = \mu \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{1}{\tau}.$$

- *Pareto distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{ab^a}{x^{a+1}} I_{[b,\infty)}(x),$$

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\text{Pa}(a, b)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{ba}{a-1} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{b^2 a}{(a-1)^2(a-2)}$$

if $a > 1$ and $a > 2$, respectively.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

- *Inverse Pareto distribution*: this is characterised by the following density:

$$f(x \mid \theta) = ab^a x^{a-1} I_{(0,1/b)}(x),$$

with $\theta = (a, b)$ and valid for $a, b > 0$. We denote it as $\text{Ipa}(a, b)$. The first two moments are

$$\text{E}(X \mid \theta) = \frac{a}{b(a+1)} \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{a}{b^2(a+1)^2(a+2)}.$$

- *Student-t*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(\frac{\tau}{\nu\pi}\right)^{1/2} \left(1 + \frac{\tau(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2} I_{\mathbb{R}}(x),$$

with $\theta = (\mu, \tau, \nu)$ and valid for $\mu \in \mathbb{R}$ and $\tau, \nu > 0$. We denote it as $\text{St}(\mu, \tau, \nu)$. The first two moments are

$$\text{E}(X \mid \theta) = \mu \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{\nu}{\tau(\nu-2)}$$

if $\nu > 1$ and $\nu > 2$, respectively.

- *Generalised Scaled Student-t*: this is characterised by the following density:

$$f(x \mid \theta) = k(\mu, \tau) \frac{\exp\{\tan^{-1}(x)\tau\mu\}}{(1+x^2)^{1+\tau/2}} I_{\mathbb{R}}(x),$$

with $k(\mu, \tau)$ a normalising constant, $\theta = (\mu, \tau)$ and valid for $\mu \in \mathbb{R}$ and $\tau > 0$. We denote it as $\text{GSSt}(\mu, \tau)$. The first two moments are

$$\text{E}(X \mid \theta) = \mu \quad \text{and} \quad \text{Var}(X \mid \theta) = \frac{1+\mu^2}{\tau-1}$$

if $\tau > 1$.

- *Generalised Hyperbolic Secant distribution*: this is characterised by the following density:

$$f(x \mid \theta) = \frac{2^{\tau-2}}{\Gamma(\tau)} \prod_{k=0}^{\infty} \left\{1 + \frac{x^2}{(\tau+2k)^2}\right\}^{-1} \frac{\exp\{\tan^{-1}(\mu)x\}}{(1+\mu^2)^{\tau/2}} I_{\mathbb{R}}(x),$$

with $\theta = (\mu, \tau)$ and valid for $\mu \in \mathbb{R}$ and $\tau > 0$. We denote it as $\text{GHS}(\mu, \tau)$. The first two moments are

$$\text{E}(X \mid \theta) = \mu \quad \text{and} \quad \text{Var}(X \mid \theta) = \tau + \mu^2/\tau.$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

- *Dirichlet distribution*: This is a multivariate distribution characterised by the following density:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\Gamma(\sum_{j=1}^{k} \theta_j)}{\prod_{j=1}^{k} \Gamma(\theta_j)} \prod_{j=1}^{k} x_j^{\theta_j - 1} I\left(\sum_{j=1}^{k} x_j = 1\right),$$

where $\mathbf{x} = (x_1, \ldots, x_k)$, $x_j \in (0,1)$ for $j = 1, \ldots, k$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and valid for $\theta_j > 0$ for $j = 1, \ldots, k$. We denote it as $\text{Dir}(\boldsymbol{\theta})$. The first two moments are

$$\text{E}(X_j \mid \boldsymbol{\theta}) = \frac{\theta_j}{\sum_{j=1}^{k} \theta_j}, \quad \text{Var}(X_j \mid \boldsymbol{\theta}) = \frac{\theta_j(\sum_{i \neq j} \theta_i)}{(\sum_{i=1}^{k} \theta_i)^2(\sum_{i=1}^{k} \theta_i + 1)} \quad \text{and}$$

$$\text{Cov}(X_i, X_j) = \frac{-\theta_i \theta_j}{(\sum_{l=1}^{k} \theta_l)^2(\sum_{l=1}^{k} \theta_l + 1)}$$

for $i \neq j$.

- *Multivariate normal distribution*: This is a multivariate distribution characterised by the following density:

$$f(\mathbf{x} \mid \theta) = (2\pi)^{-p/2} |\mathbf{C}|^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \mathbf{C} (\mathbf{x} - \boldsymbol{\mu})\right\} I_{\mathbb{R}^p}(\mathbf{x}),$$

with $\mathbf{x} = (x_1 \ldots, x_p)$, $\theta = (\boldsymbol{\mu}, \mathbf{C})$ and valid for $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\mathbf{C}$ a precision matrix of dimension $p \times p$. We denote it as $\text{N}_p(\boldsymbol{\mu}, \mathbf{C})$. The first two moments are

$$\text{E}(\mathbf{X} \mid \theta) = \boldsymbol{\mu} \quad \text{and} \quad \text{Var}(\mathbf{X} \mid \theta) = \mathbf{C}^{-1}.$$

In general, we will use a tilde '$\sim$' to denote 'distributed as', for example $X \sim \text{Ber}(\theta)$ means that the random variable X has a Bernoulli distribution with parameter θ. We will put an argument in front to denote density, for example $\text{Ber}(x \mid \theta)$ denotes the Bernoulli density. In some cases, to make our statements clear, we will explicitly denote the random variables involved as well as the arguments for densities, for example $f_X(x)$ denotes the density for random variable X evaluated at value x and $f_{X|Y}(x \mid y)$ denotes the conditional density of random variable X given random variable $Y = y$ evaluated at value x. Whenever we can, we will avoid the sub-indexes.

1.3 Moments

In Section 1.1 we used the notation $f(x \mid \theta)$ to denote a parametric density. In this section we remove the explicit dependence on the parameter θ to avoid burdening the notation and simply denote a density as $f(x)$.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

Let us recall the definition of moments, marginal and joint. Let X be a real random variable with probability distribution $f(x)$ and let $g(\cdot)$ be a real function. Then the expectation operator E of function g of X is defined as

$$\mathrm{E}\{g(X)\} = \int_{\mathbb{R}} g(x)f(x)\,\mathrm{d}x \overset{\text{or}}{=} \sum_{x\in\mathbb{R}} g(x)f(x) \tag{1.2}$$

according to whether X is continuous or discrete.

There are two particular cases for g. These are:

- If $g(x) = x$, then $\mathrm{E}\{g(X)\} = \mathrm{E}(X) = \mu$ and it is called the *mean*. This is also known as the first non-central moment.
- If $g(x) = (x-\mu)^2$, then $\mathrm{E}\{g(X)\} = \mathrm{E}\{(X-\mu)^2\} = \sigma^2$ and it is called the *variance*. This is also known as the second central moment.

The mean is a measure of central tendency of the values in a random variable, whereas the variance is a measure of dispersion. It is common to consider the squared root of the variance σ, called standard deviation, to measure the dispersion in the same units as the random variable.

Let (X,Y) be a random vector with joint probability distribution $f(x,y)$ and let g be a real function such that $g\colon \mathbb{R}^2 \to \mathbb{R}$. Then the expectation of g of (X,Y) is defined as

$$\mathrm{E}\{g(X,Y)\} = \int_{\mathbb{R}^2} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \overset{\text{or}}{=} \sum_{(x,y)\in\mathbb{R}^2} g(x,y)f(x,y) \tag{1.3}$$

according to whether the vector (X,Y) is continuous or discrete. Mixture nature of the random variables is also possible with the appropriate changes to expression (1.3). There is one particular case for g that we are interested in. This is

- If $g(x,y) = (x-\mu_x)(y-\mu_Y)$, with μ_X and μ_Y the mean of X and Y, respectively, then $\mathrm{E}\{g(X,Y)\} = \mathrm{E}\{(X-\mu_X)(Y-\mu_Y)\} = \mathrm{E}(XY) - \mu_X\mu_Y = \mathrm{Cov}(X,Y)$ and is called the *covariance*. This is also known as the second cross central moment.

The covariance is a measure of the linear dependence between the two random variables X and Y and it can take any real value. To better interpret the linear dependence, it is customary to compute the covariance of the standardised variables, which produces the correlation, denoted by ρ, and defined as

$$\rho = \mathrm{Corr}(X,Y) = \mathrm{E}\left\{\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

The correlation satisfies $\rho \in [-1, 1]$, which makes it easier to determine when the correlation is strong, for values close to -1 or 1, or weak, for values close to zero.

The conditional distributions of X given Y, denoted as $f(x \mid y)$, and the conditional distribution of Y given X, denoted as $f(y \mid x)$ are defined as

$$f(x \mid y) = \frac{f(x,y)}{f(y)} \quad \text{and} \quad f(y \mid x) = \frac{f(x,y)}{f(x)}$$

if the denominators $f(y)$ and $f(x)$ are positive, respectively.

Then, the theorem of total probability states that

$$f(x) = \int_{\mathbb{R}} f(x \mid y) f(y) \, \mathrm{d}y \overset{\text{or}}{=} \sum_{y \in \mathbb{R}} f(x \mid y) f(y) = \mathrm{E}_Y\{f(x \mid Y)\}, \qquad (1.4)$$

that is, we can recover the marginal distribution of X by taking the expected value with respect to Y of the conditional distribution of X given Y. Similarly, $f(y) = \mathrm{E}_X\{f(y \mid X)\}$. With the conditional distributions we can define conditional moments. Let g be a real function; then the expected value of the function g of x with respect to the conditional distribution of X given Y is given by

$$\mathrm{E}\{g(X) \mid y\} = \int_{\mathbb{R}} g(x) f(x \mid y) \, \mathrm{d}x \overset{\text{or}}{=} \sum_{x \in \mathbb{R}} g(x) f(x \mid y).$$

We note that this conditional expected value $\mathrm{E}\{g(X) \mid y\}$ is a function of the conditioning variable Y. There are two particular cases of interest:

- If $g(x) = x$, then $\mathrm{E}\{g(X) \mid y\} = \mathrm{E}\{X \mid y\} = \mu_{X|y}$ is the conditional mean of X given Y.
- If $g(x) = (x - \mu_{X|y})^2$, then $\mathrm{E}\{g(X) \mid y\} = \mathrm{E}\{(X - \mu_{X|y})^2 \mid y\} = \mathrm{Var}(X \mid y)$ is the conditional variance of X given Y.

Let (X, Y, Z) be a vector of dimension three with joint distribution function $f(x, y, z)$. The conditional distribution of (X, Y) given Z is defined as

$$f(x, y \mid z) = \frac{f(x, y, z)}{f(z)}.$$

Let $g(x, y) = (x - \mu_{X|z})(y - \mu_{Y|z})$; then the conditional covariance of (X, Y) given Z is defined as

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

$$\begin{aligned}\mathrm{Cov}(X,Y \mid z) &= \mathrm{E}\{g(X,Y) \mid z\} = \mathrm{E}\{(X-\mu_{X|z})(Y-\mu_{Y|z}) \mid z\} \\ &= \int_{\mathbb{R}^2} (x-\mu_{X|z})(y-\mu_{Y|z}) f(x,y \mid z)\,\mathrm{d}x\,\mathrm{d}y \\ &\overset{\text{or}}{=} \sum_{(x,y)\in\mathbb{R}^2} (x-\mu_{X|z})(y-\mu_{Y|z}) f(x,y \mid z) \\ &= \mathrm{E}\{XY \mid z\} - \mathrm{E}(X \mid z)\mathrm{E}(Y \mid z).\end{aligned}$$

From the conditional mean, variance and covariance, we can recover the marginal mean, variance and covariance, via the iterative result, which is given as follows.

Proposition 1.1 Mood et al. (1974) *Let* (X,Y,Z) *be a random vector of dimension three. If the conditional expectations, variances and covariances exist, then*

i). $E(X) = E_Y\{E(X \mid Y)\}$
ii). $Var(X) = E_Y\{Var(X \mid Y)\} + Var_Y\{E(X \mid Y)\}$
iii). $Cov(X,Y) = E_Z\{Cov(X,Y \mid Z)\} + Cov_Z\{E(X \mid Z), E(Y \mid Z)\}$

Proposition 1.1 is the most important result of this section that we will exploit throughout the remaining chapters of the book. Let us present some examples.

Example 1.2 Let (X,N) be a bivariate random vector, whose probability distribution is given by $X \mid N = n \sim \mathrm{Bin}(n,p)$ and $N \sim \mathrm{Bin}(m,q)$. Explicitly, we have

$$f(x \mid n) = \binom{n}{x} p^x (1-p)^{n-x} I_{\{0,1,\ldots,n\}}(x)$$

and

$$f(n) = \binom{m}{n} q^n (1-q)^{m-n} I_{\{0,1,\ldots,m\}}(n).$$

The objective is to find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ in two ways: (a) obtaining the marginal distribution of X using the theorem of total probability (1.4); and (b) using the iterative mean and variance formulae given in Proposition (1.1). For (a) we use the theorem of total probability:

$$\begin{aligned}f(x) &= \mathrm{E}\{f(x \mid N)\} = \sum_n f(x \mid n) f(n) \\ &= \sum_{n=x}^{m} \frac{n!}{(n-x)!x!} \frac{m!}{(m-n)!n!} p^x (1-p)^{n-x} (1-q)^m \left(\frac{q}{1-q}\right)^n I_{\{0,1,\ldots,m\}}(x).\end{aligned}$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

After cancelling some factorials, doing the change of variable $u = n - x$ and completing the combinations, we get

$$f(x) = (pq)^x(1-q)^{m-x}\binom{m}{x} I_{\{0,1,\ldots,m\}}(x) \sum_{u=0}^{m-x} \binom{m-x}{u} \left(\frac{q(1-p)}{1-q}\right)^u.$$

After computing the last sum with Newton's theorem, we get

$$f(x) = \binom{m}{x} (pq)^x(1-pq)^{m-x} I_{\{0,1,\ldots,m\}}(x).$$

Therefore, the marginal distribution of X is another binomial of the form $X \sim \text{Bin}(m, pq)$. In this case, $\text{E}(X) = mpq$ and $\text{Var}(X) = mpq(1-pq)$. Now, for (b) we use the iterative results and obtain that the mean becomes

$$\text{E}(X) = \text{E}\{\text{E}(X \mid N)\} = \text{E}(Np) = p\text{E}(N) = mpq$$

and the variance is

$$\begin{aligned}\text{Var}(X) &= \text{E}\{\text{Var}(X \mid N\} + \text{Var}\{\text{E}(X \mid N)\} = \text{E}(Np(1-p)) + \text{Var}(Np) \\ &= p(1-p)\text{E}(N) + p^2\text{Var}(N) = p(1-p)mq + p^2mq(1-q) \\ &= mpq - mp^2q + mp^2q - mp^2q^2 = mpq(1-pq),\end{aligned}$$

which correspond to the previous computed values. As a further illustration, we can compute the conditional distribution $f(n \mid x)$ by using Bayes's theorem (1.1):

$$f(n \mid x) = \frac{\binom{n}{x}p^x(1-p)^{n-x} I_{\{0,1,\ldots,n\}}(x) \binom{m}{n} q^n(1-q)^{m-n} I_{\{0,1,\ldots,m\}}(n)}{\binom{m}{x}(pq)^x(1-pq)^{m-x} I_{\{0,1,\ldots,m\}}(x)}.$$

After re-writing the product of indicator variables in the numerator as $I_{\{x,x+1,\ldots,m\}}(n) I_{\{0,1,\ldots,m\}}(x)$ and cancelling some common terms, we get

$$f(n \mid x) = \binom{m-x}{m-n} \left\{\frac{(1-p)q}{1-pq}\right\}^{n-x} \left(\frac{1-q}{1-pq}\right)^{m-n} I_{\{x,x+1,\ldots,m\}}(n),$$

which can be identified as a shifted binomial, that is, $N - x \mid X = x \sim \text{Bin}\left(m-x, \frac{(1-p)q}{1-pq}\right)$.

Example 1.3 Let (X, N) be a bivariate random vector, whose probability distribution is given by $X \mid N = n \sim \text{Bin}(n, p)$ and $N \sim \text{Po}(\lambda)$. Explicitly, we have

$$f(x \mid n) = \binom{n}{x} p^x(1-p)^{n-x} I_{\{0,1,\ldots,n\}}(x)$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

and

$$f(n) = e^{-\lambda}\frac{\lambda^n}{n!}I_{\{0,1,\ldots\}}(n).$$

As in the previous example, the objective is to find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ via (a) the marginal distribution of X using the theorem of total probability and (b) using the iterative mean and variance formulae. For (a) we use the theorem of total probability (1.4):

$$\begin{aligned} f(x) &= \mathrm{E}\{f(x \mid N)\} = \sum_n f(x \mid n) f(n) \\ &= \sum_{n=x}^{\infty} \frac{n!}{(n-x)!x!}\frac{1}{n!}p^x(1-p)^{n-x}e^{-\lambda}\lambda^n I_{\{0,1,\ldots\}}(x). \end{aligned}$$

After cancelling some factorials, doing the change of variable $u = n - x$, we get

$$f(x) = \frac{e^{-\lambda}}{x!}\left(\frac{p}{1-p}\right)^x I_{\{0,1,\ldots\}}(x)\sum_{u=0}^{\infty}\frac{1}{u!}\left(\lambda(1-p)\right)^{u+x}.$$

After computing the last sum with Taylor expansion of the exponential function and cancelling some elements, we get

$$f(x) = e^{-\lambda p}\frac{(\lambda p)^x}{x!}I_{\{0,1,\ldots\}}(x).$$

Therefore, the marginal distribution of X is a Poisson of the form $X \sim \mathrm{Po}(\lambda p)$. In this case, $\mathrm{E}(X) = \lambda p$ and $\mathrm{Var}(X) = \lambda p$. Now, for (b) we use Proposition (1.1) and obtain that the mean becomes

$$\mathrm{E}(X) = \mathrm{E}\{\mathrm{E}(X \mid N)\} = \mathrm{E}(Np) = p\mathrm{E}(N) = p\lambda$$

and the variance is

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\{\mathrm{Var}(X \mid N\} + \mathrm{Var}\{\mathrm{E}(X \mid N)\} = \mathrm{E}(Np(1-p)) + \mathrm{Var}(Np) \\ &= p(1-p)\mathrm{E}(N) + p^2\mathrm{Var}(N) = p(1-p)\lambda + p^2\lambda \\ &= p\lambda(1-p+p) = p\lambda. \end{aligned}$$

We can compute the conditional distribution $f(n \mid x)$ by using Bayes's theorem (1.1):

$$\begin{aligned} f(n \mid x) &= \frac{f(x \mid n)f(n)}{f(x)} \\ &= \frac{\binom{n}{x}p^x(1-p)^{n-x}I_{\{0,1,\ldots,n\}}(x)e^{-\lambda}\lambda^n\frac{1}{n!}I_{\{0,1,\ldots\}}(n)}{e^{-\lambda p}(\lambda p)^n\frac{1}{n!}I_{\{0,1,\ldots\}}(x)}. \end{aligned}$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

After re-writing the product of indicator variables in the numerator as $I_{\{x,x+1,\ldots\}}(n)I_{\{0,1,\ldots\}}(x)$ and cancelling some common terms, we get

$$f(n \mid x) = e^{-\lambda(1-p)} \frac{\{\lambda(1-p)\}^{n-x}}{(n-x)!} I_{\{x,x+1,\ldots\}}(n),$$

which can be identified as a shifted Poisson, that is, $N - x \mid X = x \sim \text{Po}(\lambda(1-p))$.

1.4 Stochastic Processes

Definition 1.4 A stochastic process, denoted by $\{X(t) : t \in \mathbb{T}\}$, is a family or collection of random variables, where t is a parameter that takes values in $\mathbb{T}$. For each t, $X(t)$ is a random variable.

In general, the most common parameter t that indexes a stochastic process is time. In this case $X(t)$ would be the state of the process at time t. However, t could be space in any dimension, for instance in $\mathbb{R}^2$, if $\mathbf{t} = (t_1, t_2)$ then $X(\mathbf{t})$ would be the state of the process at location (t_1, t_2). Sometimes we interchange $X(t)$ with X_t to simplify the notation, avoiding the use of parentheses.

$\mathbb{T}$ is the index set of the process. If $\mathbb{T}$ is enumerable, then $X(t)$ is a process in discrete time, for example $\{X(t) : t \in \mathbb{N}\}$. If $\mathbb{T}$ is a non-enumerable subset of $\mathbb{R}$, then $X(t)$ is a process in continuous time, for example $\{X(t) : t > 0\}$.

The state of the process is the set of all possible values $X(t) \in \mathbb{X}$ for all $t \in \mathbb{T}$. The state space $\mathbb{X}$ can be discrete or continuous.

One particular type of process of interest is the Markov process. We define it here.

Definition 1.5 A stochastic process $\{X(t) : t \in \mathbb{T}\}$ is a Markov process if it satisfies the Markovian property that states given the present, $X(t)$, the values of the future, $X(s)$ for $s > t$, do not depend on the past, $X(u)$ for $u < t$. In notation,

$$\begin{aligned} P\{X(s) \in A \mid X(u_0) = x_{u_0}, X(u_1) = x_{u_1}, \ldots, X(u_n) = x_{u_n}, X(t) = x_t\} \\ = P\{X(s) \in A \mid X(t) = x_t\} \end{aligned}$$

for arbitrary $A \subset \mathbb{X}$ and $u_0 < u_1 < \cdots < u_n \leq t < s$.

Another property of interest of stochastic processes is *stationarity*. This is a condition that can be achieved strictly or weakly.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

Definition 1.6 A stochastic process $\{X(t)\colon t \in \mathbb{T}\}$ is strictly stationary if it satisfies that for all $n, s, t_1, \ldots, t_n$, the vectors $(X(t_1), \ldots, X(t_n))$ and $(X(t_1 + s), \ldots, X(t_n + s))$ have the same joint distribution.

In other words, Definition 1.6 says that there must be a kind of invariant distribution if we shift the process a specific amount of time. In particular, it must be satisfied that $X(t)$ and $X(t + s)$ must have the same distribution.

A weaker version of stationarity is defined as follows.

Definition 1.7 A stochastic process $\{X(t)\colon t \in \mathbb{T}\}$ is weakly stationary, or second-order stationary, if for $s, t \in \mathbb{T}$, $X(t)$ satisfies the following two conditions:

i). $E\{X(t)\} = \mu$, *and*
ii). $Cov\{X(t), X(t + s)\} = \sigma(s)$,

That is, the first two moments do not depend on t.

Second-order stationarity, given in Definition 1.7, only requires that the first two moments of the process, mean and variance, remain constant for all times. And after shifting the process, the covariance does not depend on the specific time t, it only depends on the time difference s.

Let us consider a first example.

Example 1.8 Autoregressive process of order 1, $AR(1)$. Let $Z_1, Z_2, \ldots$ be random variables such that $\mathrm{E}(Z_t) = 0$ for all t, $\mathrm{Var}(Z_t) = \sigma^2$ for $t = 0$, $\mathrm{Var}(Z_t) = (1 - \theta^2)\sigma^2$ for $t \geq 1$, and $\mathrm{Cov}(Z_t, Z_s) = 0$ for all $t \neq s$. Let

$$X_0 = Z_0 \quad \text{and} \quad X_t = \theta X_{t-1} + Z_t, \text{ for } t \geq 1.$$

Therefore $\{X_t\colon t \in \mathbb{N}\}$ is an autoregressive process of order 1. If we iterate, we can re-write the process as

$$\begin{aligned} X_t &= \theta(\theta X_{t-2} + Z_{t-1}) + Z_t \\ &= \theta^2 X_{t-2} + \theta Z_{t-1} + Z_t \\ &\vdots \\ &= \sum_{i=0}^{t} \theta^{t-i} Z_i = \sum_{i=0}^{t} \theta^i Z_{t-i}. \end{aligned}$$

With this expression we can easily compute the first two moments of the process X_t and the covariance. The mean is

$$\mathrm{E}(X_t) = \sum_{i=0}^{t} \theta^{t-i} \mathrm{E}(Z_i) = 0.$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

The variance is

$$\begin{aligned}\text{Var}(X_t) &= \sum_{i=0}^{t} \theta^{2(t-i)}\text{Var}(Z_i) = \theta^{2t}\left\{\sigma^2 + \sum_{i=1}^{t}\theta^{-2i}(1-\theta^2)\sigma^2\right\} \\ &= \sigma^2\theta^{2t}\left[1 + (1-\theta^2)\left\{\frac{1-(\theta^{-2})^{t+1}}{1-\theta^{-2}} - 1\right\}\right] \\ &= \sigma^2.\end{aligned}$$

The covariance is

$$\begin{aligned}\text{Cov}(X_t, X_{t+s}) &= \text{Cov}\left(\sum_{i=0}^{t}\theta^{n-i}Z_i, \sum_{j=0}^{t+s}\theta^{t+s-j}Z_j\right) \\ &= \sum_{i=0}^{t}\sum_{j=0}^{t+s}\theta^{2t+s-i-j}\text{Cov}(Z_i, Z_j) \\ &= \sum_{i=0}^{t}\theta^{2t+s-i-j}\text{Var}(Z_i) = \theta^{2t+s}\left\{\sigma^2 + \sum_{i=1}^{t}\theta^{-2i}(1-\theta^2)\sigma^2\right\} \\ &= \sigma^2\theta^s\end{aligned}$$

for $s \geq 0$. Additionally,

$$\text{Corr}(X_t, X_{t+s}) = \theta^s.$$

Since the mean and variance of X_t are constant and the covariance between (X_t, X_{t+s}) only depends on the shift s, $\{X_t\}$ is a second-order stationary process. A further question would be, is $\{X_t\}$ a Markov process? The answer is yes if we add the independence assumption in the $\{Z_t\}$. In such a case

$$f(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_0) = f(x_t \mid x_{t-1}).$$

Note that the version of the autoregressive process of order one presented in Example 1.8 is a finite version of the process, in the sense that it is defined for $t = 0, 1, 2, \ldots, n$ with n finite or infinite. In such a case, to achieve second-order stationarity in $\{X_t\}$, the innovation terms Z_t have different variance for $t = 0$ than for $t \geq 1$. Common specifications of an $AR(1)$ process, for example Chatfield (2003), define the process for non-bounded times, that is, for $t \in \mathbb{Z}$. In such a case we do not need different variances to achieve second order stationarity.

Let us now consider a second example.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

Example 1.9 Moving average process of order q, $MA(q)$. Let $Z_1, Z_2, \ldots$ be random variables such that $\mathrm{E}(Z_t) = 0$, $\mathrm{Var}(Z_t) = \sigma^2$ and $\mathrm{Cov}(Z_t, Z_s) = 0$ for all $t \neq s$. Let

$$X_t = \theta_0 Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad \text{for } t \in \mathbb{Z}.$$

Therefore $\{X_t : t \in \mathbb{Z}\}$ is a moving average process of order q. We can re-write the process in two different ways

$$X_t = \sum_{i=0}^{q} \theta_i Z_{t-i} = \sum_{j=t-q}^{t} \theta_{t-j} Z_j.$$

With these expressions we can compute the first two moments of the process X_t as well as the covariance. The mean is

$$\mathrm{E}(X_t) = \mathrm{E}\left(\sum_{i=0}^{q} \theta_i Z_{t-i}\right) = \sum_{i=0}^{q} \theta_i \mathrm{E}(Z_{t-i}) = 0.$$

The variance is

$$\mathrm{Var}(X_t) = \sum_{i=0}^{q} \theta_i^2 \mathrm{Var}(Z_{t-i}) = \sigma^2 \sum_{i=0}^{q} \theta_i^2.$$

The covariance is, for $s \leq q$,

$$\begin{aligned}
\mathrm{Cov}(X_t, X_{t+s}) &= \mathrm{Cov}\left(\sum_{i=t-q}^{t} \theta_{t-i} Z_i, \sum_{j=t+s-q}^{t+s} \theta_{t+s-j} Z_j\right) \\
&\sum_{i=t-q}^{t} \sum_{j=t+s-q}^{t+s} \theta_{t-i}\theta_{t+s-j} \mathrm{Cov}(Z_i, Z_j) \\
&= \sum_{i=t+s-q}^{t} \theta_{t-i}\theta_{t+s-i} \mathrm{Var}(Z_i) \\
&= \sigma^2 \sum_{i=t+s-q}^{t} \theta_{t-i}\theta_{t+s-i} \quad \text{for } s \leq q.
\end{aligned}$$

By doing the change of variable $j = i - t - s + q$ in the previous sum, we have

$$\mathrm{Cov}(X_t, X_{t+s}) = \sigma^2 \sum_{j=0}^{q-s} \theta_{q-s-j}\theta_{q-j} \quad \text{for } s \leq q$$

and $\mathrm{Cov}(X_t, X_{t+s}) = 0$ if $s > q$. Additionally, the correlation becomes

$$\mathrm{Corr}(X_t, X_{t+s}) = \frac{\sum_{j=0}^{q-s} \theta_{q-s-j}\theta_{q-j}}{\sum_{i=0}^{q} \theta_i^2} \quad \text{for } s \leq q$$

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

and $\text{Corr}(X_t, X_{t+s}) = 0$ if $s > q$. Since the mean and variance of X_t are constants and the covariance does not depend on t, $\{X_t\}$ is a second-order stationary process. But, is $\{X_t\}$ a Markov process? The answer is no because there is no way of writing X_t in terms of X_{t-1} exclusively.

Let us consider a third example of a stochastic process in space instead of time.

Example 1.10 Conditionally autoregressive (CAR) process (Besag, 1974). Let $\{X_i : i = 1, \ldots, n\}$ be a stochastic process such that each random variable X_i is associated to an area i in a region. Let each X_i be defined conditionally on the other areas $j \neq i$ through a normal distribution of the form

$$X_i \mid X_j = x_j, j \neq i \sim \text{N}\left(\sum_j b_{ij} x_j \,,\, \tau_i\right).$$

We know that any joint distribution $f(x_1, \ldots, x_n)$ induces well-defined conditional densities $f(x_i \mid x_j, j \neq i)$; However, the converse is not always possible. Brook's lemma (Brook, 1964) states the conditions for obtaining a joint distribution based on its conditional distributions. In this case, it can be proved that

$$f(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{D}(\mathbf{I} - \mathbf{B})\mathbf{x}\right\},$$

where $\mathbf{B} = (b_{ij})$ and $\mathbf{D} = \text{diag}(\tau_1, \ldots, \tau_n)$. For this to be a well-defined joint density, we need the matrix $\mathbf{D}(\mathbf{I} - \mathbf{B})$ to be symmetric. This is satisfied if $b_{ij}\tau_i = b_{ji}\tau_j$ for all i and j. In particular, if $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i = \tau w_{i+}$, where $w_{ij} = I(i \smile j)$ with "$\smile$" denoting neighbour, and $w_{i+} = \sum_j w_{ij}$ is the number of neighbours of area i. In this case the conditional distributions become

$$f(x_i \mid x_j, j \neq i) = \text{N}\left(\sum_j \frac{w_{ij}}{w_{i+}} x_j \,,\, \tau\, w_{i+}\right) \tag{1.5}$$

and the joint distribution is

$$f(x_1, \ldots, x_n) \propto \exp\left\{-\frac{\tau}{2}\mathbf{x}'(\mathbf{D}_w - \mathbf{W})\mathbf{x}\right\}, \tag{1.6}$$

where $\mathbf{W} = (w_{ij})$ and $\mathbf{D}_w = \text{diag}(w_{1+}, \ldots, w_{n+})$. We note that $(\mathbf{D}_w - \mathbf{W})$ $\mathbf{1} = \mathbf{0}$, that is, the precision matrix $(\mathbf{D}_w - \mathbf{W})$ is singular, so the joint distribution (1.6) is improper. Expressions (1.5) and (1.6) define a stochastic process that is known as an *intrinsic CAR* process.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press

According to Banerjee et al. (2010), the impropriety condition can be corrected by adding an association parameter ρ such that the precision matrix $\mathbf{C} = \mathbf{D}_w - \rho\mathbf{W}$ becomes non-singular. This is achieved if $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)}$ and $\lambda_{(n)}$ are the minimum and maximum eigenvalues of $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$. In this case the conditional distributions become

$$f(x_i \mid x_j, j \neq i) = \mathrm{N}\left(\rho \sum_j \frac{w_{ij}}{w_{i+}} x_j \,, \tau\, w_{i+}\right)$$

and the joint distribution is $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \tau(\mathbf{D}_w - \rho\mathbf{W}))$. Alternatively, Cressie (1993) suggested correcting the impropriety condition by considering a parameter $\alpha \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)}$ and $\lambda_{(n)}$ are the minimum and maximum eigenvalues of the adjacency matrix $\mathbf{W}$ and defining a joint distribution $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \tau(\mathbf{I} - \alpha\mathbf{W}))$. Either of these two latter processes is called a *proper CAR* process.

On the other hand, none of the first two CAR processes are stationary. The first one because it is improper and the second one because $\mathrm{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1/(\tau w_{i+})$ and the marginal distribution is not invariant, that is, $X_i \sim \mathrm{N}(0, \tau w_{i+})$. However, the third specification does define a stationary process with invariant distribution $X_i \sim \mathrm{N}(0, \tau)$. Moreover, the three CAR processes satisfy a Markov property in space, because their law only depends on neighbours of the first kind. Therefore intrinsic and proper CAR models are known as *Markov random fields*.

https://doi.org/10.1017/9781009584128.002 Published online by Cambridge University Press