2 Writing Statistical Models

We had an initial exposure to various kinds of data in Chapter 1. Practical inference problems arise from someone wanting to answer a question about a population on the basis of some sample data. The sample data are measurements on some random variable on a sample of individuals. The population can be of anything, e.g., people, insects, cars, plants, rivers, stars, restaurants, or absolutely anything at all. Likewise, the random variable in question can be something simple, like height, or rather exotic, such as a randomly constructed cube or tetrahedron! We tend to equate the word sample to the sample data, and similarly equate the random variable, or more precisely, its distribution, to the word population, and we try to model the distribution of the random variable. Models are almost never unique, and also, all models are usually wrong. We just want to choose a useful model. Model choice requires sound knowledge of what choices you have. In this chapter, we talk about some standard choices in standard scenarios. The word model actually means more than a distribution for a random variable. Its implied meaning depends on the context. For example, if one random variable $Y$ is related to another random variable $X$, but we do not know the exact mathematical form of the relationship, we would want to model that. This chapter is about models for some problems that we often encounter; but, primarily we discuss models in the context of distributions for random variables. A few excellent references are Kendall and Stuart (2009), Johnson, Kotz, and Balakrishnan (1995), Rao (1973), Ross (2000), and Davison (2003); you may also consult DasGupta (2011).

2.1 Parameters, Statistics, Functionals

The most fundamental point that a student of statistics needs to understand is the distinction between the population and the sample, or equivalently, the distinction between the distribution of the concerned random variable and the particular sample data on it. The entire population is never observed by us; we have to infer about it on the basis of what was observed by us, namely the sample. Distributions involve parameters, and these parameters are unknown to us. We want to infer about the parameters by using the sample data, or functions of the sample data. These are statistics. Statistics can be computed, and we use them to infer about unknown parameters; the conceptual distinction between parameters and statistics is critical, and the two must never be confused.

Definition 2.1. Let $X$ be a random variable taking values in some suitable space $\mathcal{X}$ and let $P$ denote the distribution of $X$. Then, any functional of $P$, including $P$ itself, is called a parameter of $P$.

Some examples are needed.
**Example 2.1.** Suppose $X \sim N(\mu, 1)$. Then, our $P$ is just the normal distribution with mean $\mu$ and variance one. Note that $P$ depends on the exact value of $\mu$; $\mu$ is a parameter of $P$. We often write $P_\mu$ in order to explicitly point out the dependence of $P$ on $\mu$. In general, the parameter $\mu$ can be any real number. The set of all possible values of the parameter is called the parameter space. So, in this example, the parameter space is $\mathcal{R} = (-\infty, \infty)$. Functions of a parameter are also considered as parameters; for example, $\mu^2, |\mu|, e^{2\mu}, I_{\mu>0}$, these are all parameters also. Often, to avoid ambiguity and confusion, we pick a basic parameter and refer to the basic parameter as a parameter, and refer to functions of that basic parameter as parametric functions. Thus, in this example, we may decide to refer to $\mu$ as the basic parameter and refer to $\mu^2, |\mu|$ etc. as parametric functions.

**Example 2.2.** Suppose $X \sim \text{Poi}(\lambda)$, the Poisson distribution with mean $\lambda$. The basic parameter is $\lambda$ and in general $\lambda$ can take any positive value. So the parameter space is $\mathcal{R}_+ = (0, \infty)$. For all Poisson distributions, the mean and the variance are equal. So, the standard deviation of $X$ is $\sqrt{\text{Var}(X)} = \sqrt{\lambda}$; this would be a parametric function.

**Example 2.3.** Suppose $X \sim G(\alpha, \lambda)$, the Gamma distribution having the density function

$$f(x|\alpha, \lambda) = \frac{e^{-x/\lambda}x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)}, x > 0, \alpha, \lambda > 0$$

In this example, our $P$ is the Gamma distribution with parameters $\alpha$ and $\lambda$. Notice that in this example, in general there are two parameters. If one of them, say $\alpha$ was set equal to a specific value, such as $\alpha = 1$, then the distribution is left with only one parameter, namely $\lambda$. But, in general, in this example, the parameter space is the cartesian product $\mathcal{R} \otimes \mathcal{R} = (0, \infty) \otimes (0, \infty)$.

If $X \sim G(\alpha, \lambda)$, then $E(X) = \alpha \lambda$; this is a parametric function. Likewise, $\text{Var}(X) = \alpha \lambda^2$ is also a parametric function. So are $|\alpha \lambda - 1|$, or $\log \lambda$, as these are all functions of the basic parameters $\alpha, \lambda$.

**Example 2.4.** Suppose a bivariate random vector $(X, Y)$ is jointly distributed as a bivariate normal with means $\mu_1, \mu_2$, variances $\sigma_1^2, \sigma_2^2$, and correlation coefficient $\rho$. Our $P$ in this example is this bivariate normal distribution, which in general has five parameters, $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. If we set $\sigma_1 = 1, \sigma_2 = 2$, just as an example, then $P$ will be left with three parameters. What are some examples of parametric functions? $\mu_1 - \mu_2, \max\{\sigma_1, \sigma_2\}, \rho \sigma_1 \sigma_2, |\rho|$, these are all examples of parametric functions.

**Example 2.5. (A Nonregular Example)** Suppose $X \sim U[0, \theta], \theta > 0$. Now, our $P$ is the uniform distribution on the interval $[0, \theta]$; the basic parameter is $\theta$. The variance of $X$ in this case is $\frac{\theta^2}{12}$; this is an example of a parametric function. This example differs in
a fundamental manner from the first three examples. For instance, in Example 2.1, the set of possible values of $X$ is always the entire real line, no matter what the value of the parameter $\theta$ is. Similarly, in Example 2.4, the set of possible values of the bivariate vector $(X, Y)$ is always $\mathbb{R}^2$, no matter what the values of the five parameters are. But, in this example, the set of possible values of $X$ is $0 \leq X \leq \theta$. Thus, the set of possible values of $X$ depends on the value of the parameter $\theta$. In such a case, we call the corresponding distribution $P$ a nonregular distribution. Two other examples of nonregular distributions would be $U[\theta_1, \theta_2]$, or an exponential distribution with a left endpoint at a parameter $\theta$:

$$f(x|\mu, \lambda) = \frac{e^{-(x-\mu)/\lambda}}{\lambda}, x \geq \mu.$$ 

In general, the parameter space in this example, namely the set of all possible values of $(\alpha, \mu)$ is $(0, \infty) \otimes (-\infty, \infty)$.

**Example 2.6. (Infinite Dimensional Parameter)** You will now see an example where the underlying distribution $P$ does not have one, two, three, or indeed any finite number of parameters, but an infinite number of parameters. The meaning of there being an infinite number of parameters has to be carefully understood. Here is the model. Suppose we have a continuous real valued random variable $X$ with a density function $f(x)$. But, we do not assume that $f$ is of any particular form with a name, such as normal, or Cauchy, or $t$, or anything else of a specified functional form. *We only assume* that $f$ is symmetric, i.e., $f(-x) = f(x)$ for all $x$, and that $f$ is unimodal. So, in particular, $f$ can be normal, or Cauchy, or $t$, or of many other functional forms. In this case, the parameter is the density $f$ itself, and the parameter space is a set of functions:

$$\mathcal{F} = \{f : f \text{ is a density, } f(-x) = f(x) \text{ for all } x, f \text{ is decreasing on } (0, \infty)\}$$

This is an infinite dimensional set. In this sense, the parameter space in this example is infinite dimensional, and it is common usage to call such a model nonparametric. Nonparametric does not mean that there are no parameters! To the contrary, it is supposed to mean that the parameter is infinite dimensional.

As we remarked before, unknown parameters of distributions of random variables are to be estimated by using the sample data, or more precisely, suitable functions of the sample data. These are called statistics.

**Definition 2.2.** Let for some $n \geq 1, X^{(n)} = (X_1, X_2, \ldots, X_n)$ be the vector of sample observations. Then, any function $T(X^{(n)}) = T(X_1, X_2, \ldots, X_n)$ is called a statistic. If we consider the collection of statistics $T(X_1, X_2, \ldots, X_n)$ as $n$ runs through the integers $1, 2, \cdots$, we call $T(X_1, X_2, \cdots, X_n), n \geq 1$ a sequence of statistics.

To decide if something is a statistic, ask yourself: *can I compute its numerical value by*
knowing only the sample data? If the answer is yes, then it is a statistic. Once again, let us see a few examples.

**Example 2.7.** Consider general sample data (observations) \( X_1, X_2, \ldots, X_n \) in a generic problem. Then, each of the following is a function of \( X_1, X_2, \ldots, X_n \), and so, is a statistic:

- **Sample mean** \( \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \);
- **Sample variance** \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \);
- **Sample maximum** \( X_{(n)} = \max_{1 \leq i \leq n} X_i \);
- **Fraction of nonnegative observations** \( T_n = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \geq 0} \);
- **Sample observation closest to zero** \( T_n = \text{Sample observation closest to zero} \).

Sometimes we deal with quantities that are simultaneously functions of the sample values as well as unknown parameters. Since they depend on unknown parameters, they cannot be computed by knowing only the sample data; therefore, such functions do not count as statistics. They are called functionals.

**Definition 2.3.** Let the vector of sample observations \( X^{(n)} = (X_1, X_2, \ldots, X_n) \) have a joint distribution \( P = P_n \). Let \( T_n = T(X^{(n)}, P) \) be a function of \( X^{(n)} \) and \( P \). Then, \( T_n \) is called a functional.

Here are a few examples of functionals.

**Example 2.8.** Suppose \( X_1, X_2, \ldots, X_n \) \( \text{iid} \sim N(\mu, \sigma^2) \), where \( \mu, \sigma \) are both unknown parameters. Then, each of the following is a functional:

\[ \sum_{i=1}^{n} (X_i - \mu)^2; \sqrt[n]{n} (\bar{X} - \mu); \sqrt[n]{n} (\bar{X} - \mu) \sigma; \text{Sample observation closest to } \mu; I_{|\bar{X} - \mu| > 1}. \]

On the other hand, each of the following is a statistic:

\[ \sum_{i=1}^{n} (X_i - \bar{X})^3; \sqrt[n]{n} \bar{X}; \frac{X_{(n)} - \bar{X}}{s}. \]

And, each of the following is a parametric function:

\[ \frac{\mu}{\sigma}; E(|\bar{X} - \mu|); \text{Var}(\sum_{i=1}^{n} (X_i - \mu)^2); P(|\bar{X} - \mu| > 1). \]
2.1.1 Parametric Models

When we talk of a general distribution \( P \) with some finite number of parameters, we denote the parameter vector by the general notation \( \theta \). This notation is universal. Referring back to Example 2.3, the parameter vector is \( \theta = (\alpha, \lambda) \). On the other hand, in Example 2.1, there is only one parameter \( \mu \), and we associate \( \theta \) with \( \mu \), so that the parameter is a scalar parameter.

**Definition 2.4.** Let a random variable \( X \sim P = P_\theta \), where the parameter \( \theta \) belongs to a set (the parameter space) \( \Theta \). If \( \Theta \) is a subset of some finite dimensional Euclidean space \( \mathbb{R}^k, 1 \leq k < \infty \), we call \( \{P_\theta, \theta \in \Theta\} \) a parametric family and the model a parametric model.

**Example 2.9.** Each of these is a parametric model:

\[
X \sim \text{Poi}(\lambda), \lambda > 0; \\
X \sim N(\mu, 1), -\infty < \infty < \mu; \\
X \sim N(\mu, 1), 0 < \mu < 1; \\
X \sim N(\mu, \sigma^2), -\infty < \infty < \mu, \sigma > 0; \\
X \sim \text{Beta}(\alpha, \beta), \alpha, \beta > 0; \\
X \sim \text{U}[0, \theta], \theta > 0; \\
X \sim N_p(\mu, \Sigma), \mu \in \mathbb{R}^p, \Sigma \text{ positive definite}.
\]

2.1.2 Glimpse of Exponential Family

The basic ideas of statistical inference arose within the domain of parametric families, and a very large number of them originated in the work of R.A. Fisher. As you may have already suspected by reading the examples presented so far, parametric models such as the Poisson, normal, exponential, Gamma, etc. are frequently used to model the distribution of a random variable. Interestingly, many standard distributions such as these, discrete and continuous both included, can be put together in a general parametric family with a certain general functional form for the density or the mass function. This is the exponential family, and exponential families can handle one, two, or any finite number of parameters. Exponential families are so important that we will study them separately as a single chapter in Chapter 5; but we will have a quick first look at it right now.

**Definition 2.5.** Let \( X = (X_1, \cdots, X_d) \) be a \( d \)-dimensional random vector with a distribution \( P_\theta, \theta \in \Theta \subseteq \mathcal{R} \).
Suppose $X_1, \ldots, X_d$ are jointly continuous. The family of distributions $\{P_\theta, \theta \in \Theta\}$ is said to belong to the one parameter Exponential family if the density of $X = (X_1, \ldots, X_d)$ may be represented in the form

$$f(x | \theta) = e^{\eta(\theta)T(x) - \psi(\theta)h(x)},$$

for some real valued functions $T(x), \psi(\theta)$ and $h(x) \geq 0$.

If $X_1, \ldots, X_d$ are jointly discrete, then $\{P_\theta, \theta \in \Theta\}$ is said to belong to the one parameter Exponential family if the joint pmf $p(x | \theta) = P_\theta(X_1 = x_1, \ldots, X_d = x_d)$ may be written in the form

$$p(x | \theta) = e^{\eta(\theta)T(x) - \psi(\theta)h(x)},$$

for some real valued functions $T(x), \psi(\theta)$ and $h(x) \geq 0$.

**Example 2.10. (Normal Distribution with a Known Mean).** Suppose $X \sim N(0, \sigma^2)$. Then the density of $X$ is

$$f(x | \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} I_{x \in \mathbb{R}}.$$  

This density is parametrized by a single parameter $\sigma$. Writing

$$\eta(\sigma) = -\frac{1}{2\sigma^2}, T(x) = x^2, \psi(\sigma) = \log \sigma, h(x) = \frac{1}{\sqrt{2\pi}} I_{x \in \mathbb{R}},$$

we can represent the density in the form

$$f(x | \sigma) = e^{\eta(\sigma)T(x) - \psi(\sigma)h(x)},$$

for any $\sigma \in \mathbb{R}_+$.  

**Example 2.11. (Binomial Distribution).** Let $X \sim Bin(n, p)$, with $n \geq 1$ considered as known, and $0 < p < 1$ a parameter. We represent the pmf of $X$ in the one parameter Exponential family form.

$$f(x | p) = \binom{n}{x} p^x (1-p)^{n-x} I_{\{x \in \{0, 1, \ldots, n\}\}} = \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n I_{\{x \in \{0, 1, \ldots, n\}\}}$$

$$= \binom{n}{x} e^{x \log \frac{p}{1-p} + n \log(1-p)} I_{\{x \in \{0, 1, \ldots, n\}\}}.$$  

Writing $\eta(p) = \log \frac{p}{1-p}, T(x) = x, \psi(p) = -n \log(1-p)$, and $h(x) = \binom{n}{x} I_{\{x \in \{0, 1, \ldots, n\}\}}$, we have represented the pmf $f(x | p)$ in the one parameter Exponential family form, as long as $p \in (0, 1)$. For $p = 0$ or 1, the distribution becomes a one point distribution. Consequently, the family of distributions $\{f(x | p), 0 < p < 1\}$ forms a one parameter Exponential family, but if either of the boundary values $p = 0, 1$ is included, the family is not in the Exponential family.
Example 2.12. (Gamma Distribution). Suppose $X$ has the Gamma density $e^{-\frac{x^\alpha}{\lambda \Gamma(\alpha)}}I_{x>0}$. As such, it has two parameters $\lambda, \alpha$. If we assume that $\alpha$ is known, then we may write the density in the one parameter Exponential family form:

$$f(x | \lambda) = e^{-\frac{x^\alpha}{\lambda}} \frac{\alpha}{\Gamma(\alpha)} I_{x>0},$$

and recognize it as a density in the Exponential family with $\eta(\lambda) = -\frac{1}{\lambda}, T(x) = x, \psi(\lambda) = \alpha \log \lambda, h(x) = x^{\alpha-1} \Gamma(\alpha) I_{x>0}.$

If we assume that $\lambda$ is known, once again, by writing the density as

$$f(x | \alpha) = e^{\alpha \log x - \alpha \log \lambda - \log \Gamma(\alpha)} e^{-\frac{x}{\lambda}} I_{x>0},$$

we recognize it as a density in the Exponential family with $\eta(\alpha) = \alpha, T(x) = \log x, \psi(\alpha) = \alpha \log \lambda + \log \Gamma(\alpha), h(x) = e^{-\frac{x}{\lambda}} I_{x>0}.$

The beauty of the exponential family is that because many standard distributions are special members of the exponential family, results that we can prove for the exponential family automatically apply to these special distributions, such as normal, binomial, Gamma, etc. This spares us repetition of efforts; in addition, the generalization adds structure to our theory. That is why a study of exponential families on its own is done in Chapter 5.

2.1.3 Location and Scale Parameters

Location and scale parameters let us generate parametric families with one or at most two parameters by starting with a completely fixed density (or mass function) $f_0$. This is a bit like generating new functions from a given function in a calculus class. What is interesting is that although the model is so parsimonious in the number of parameters, just one or two, still, such parametric families can be very useful as models. We start with an example that will motivate this general approach.

Example 2.13. Let $Z \sim N(0, 1)$ with the density function $f_0(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$. Now, take a real number $\mu$ and define a new random variable $X$ as $X = Z + \mu$. Then, $X \sim N(\mu, 1)$ and its density is

$$f(x | \mu) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}, -\infty < x < \infty = f_0(x - \mu).$$

Thus, we took the standard normal density $f_0$, which was centered at zero, and simply shifted the density so that $f$ is centered at $\mu$. Since $\mu$ can be any real number, by starting with a single density $f_0$, we have generated a family of densities $f(x | \mu)$. Also, since $\mu$ is performing the job of choosing the location of the center of $f$, $\mu$ is called a location or shift parameter.
**Definition 2.6.** A real valued continuous random variable $X$ is said to have a location parameter density if it has a density of the form $f(x|\mu) = f_0(x - \mu), -\infty < \mu < \infty$, for some fixed density function $f_0$ on the real line.

A continuous random vector $X_{p \times 1}$ is said to have a location parameter density if it has a density of the form $f(x|\mu) = f_0(x - \mu), \mu = (\mu_1, \cdots, \mu_p) \in \mathcal{R}^p$, for some fixed density function $f_0$ on $\mathcal{R}^p$.

Let us see a few more examples.

**Example 2.14.** Let $X \sim C(\mu, 1)$ (Cauchy density centered at $\mu$), with the density $f(x|\mu) = \frac{1}{\pi(1+(x-\mu)^2)}, -\infty < x < \infty$. Clearly, $f(x|\mu) = f_0(x - \mu)$ with $f_0(z) = \frac{1}{\pi(1+z^2)}, -\infty < z < \infty$. Therefore, the family of $C(\mu, 1)$ densities is a location parameter family of densities on the real line.

**Example 2.15.** Let $X$ have the one parameter density $f(x|\mu) = e^{-(x-\mu)}, x > \mu$. Then, $f(x|\mu) = f_0(x - \mu)$ with $f_0(z) = e^{-z}, z > 0$. Therefore, the family of densities $\{f(x|\mu)\}$ is a location parameter family of densities.

**Example 2.16.** Let $X \sim U[\mu, \mu + 1]$ with the density $f(x|\mu) = 1_{\mu \leq x \leq \mu + 1}$. Then, $f(x|\mu) = f_0(x - \mu)$ with $f_0(z) = 1_{0 \leq z \leq 1}$ (the density of $U[0, 1]$), and so the family of $U[\mu, \mu + 1]$ densities is a location parameter family of densities.

**Example 2.17.** Let the $p$-dimensional random vector $X$ have a multivariate normal distribution, $X \sim N_p(\mu, I_p)$.

\[
f(x|\mu) = \frac{1}{(2\pi)^{p/2}} e^{-(x-\mu)'(x-\mu)/2}, x \in \mathcal{R}^p
\]

$= f_0(z - \mu)$, where $f_0(z) = \frac{1}{(2\pi)^{p/2}} e^{-z'z/2}, z \in \mathcal{R}^p$ (the density of $N_p(0, I_p)$). Therefore, the family of $N_p(\mu, I_p)$ densities is a location parameter family of densities on $\mathcal{R}^p$.

As much as shifting, scaling data is also a very natural idea. Shifting gives us location parameters, and scaling gives scale parameters. Once again, we want to see an illustrative example first.

**Example 2.18.** Let $Z \sim N(0, 1)$ with density $f_0(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$. Now take any positive real number $\sigma$ and define $X = \sigma Z$. Then, $X \sim N(0, \sigma^2)$ and $X$ has the density

\[
f(x|\sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-x^2/2\sigma^2}, -\infty < x < \infty = \frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right).
\]

Thus, we started with the standard normal density $f_0$, and simply rescaled the density to generate a whole family of densities $\{f(x|\sigma)\}$. Since $\sigma$ is performing the task of rescaling the initial variable $Z$, it is called a scale parameter.
Definition 2.7. A real valued continuous random variable \( X \) is said to have a scale parameter density if it has a density of the form \( f(x|\sigma) = \frac{1}{\sigma} f_0(\frac{x}{\sigma}), \sigma > 0 \), for some fixed density \( f_0 \) on the real line.

A continuous random vector \( X_{p \times 1} \) is said to have a scale parameter density with parameter matrix \( \Sigma_{p \times p} \) assumed to be positive definite, if it has a density of the form \( f(x|\Sigma) = \frac{1}{|\Sigma|^{1/2}} f_0(\Sigma^{-1/2} x) \), for some fixed density \( f_0 \) on \( \mathbb{R}^p \).

Here are some examples.

Example 2.19. Let \( X \sim C(0, \sigma) \) with the density \( f(x|\sigma) = \frac{1}{\pi \sigma (1 + x^2)} \), \( -\infty < x < \infty \). We recognize that \( f(x|\sigma) = \frac{1}{\sigma} f_0(\frac{x}{\sigma}) \), where \( f_0(z) = \frac{1}{\pi(1+z^2)} \), \( -\infty < z < \infty \), the standard Cauchy density. Therefore, the family of \( C(0, \sigma) \) densities is a scale parameter family of densities on the real line.

Example 2.20. Let \( X \sim Exp(\lambda) \) with the density \( f(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}, x > 0 \). Again, we recognize that \( f(x|\lambda) = \frac{1}{\lambda} f_0(\frac{x}{\lambda}) \), where \( f_0(z) = e^{-z}, z > 0 \), the standard exponential density. Therefore, the family of exponential densities with mean \( \lambda \) is a scale parameter family of densities on the real line.

Example 2.21. Let \( X \sim U[0, \theta], \theta > 0 \) with density \( f(x|\theta) = \frac{1}{\theta} I_{0 \leq x \leq \theta} = \frac{1}{\theta} f_0(\frac{x}{\theta}) \), where \( f_0(z) = I_{0 \leq z \leq 1} \), the \( U[0, 1] \) density. Therefore, the family of \( U[0, \theta] \) densities is a scale parameter family of densities on the real line.

Example 2.22. Let \( X \sim N_p(0, \Sigma) \), the multivariate normal with mean vector zero and variance-covariance matrix \( \Sigma \), assumed to be positive definite. Then, \( X \) has density

\[
f(x|\Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{x' \Sigma^{-1} x}{2}}, \quad x \in \mathbb{R}^p.
\]

We recognize that if we let \( f_0(z) = \frac{1}{(2\pi)^{p/2}} e^{-z'/z}, z \in \mathbb{R}^p \), then

\[
f(x|\Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{(\Sigma^{1/2} x)'(\Sigma^{1/2} x)/2}{2}} = \frac{1}{|\Sigma|^{1/2}} f_0(\Sigma^{1/2} x),
\]

\( x \in \mathbb{R}^p \). Therefore, the family of \( N_p(0, \Sigma) \) densities is a scale parameter family of densities on \( \mathbb{R}^p \).

What we have seen above is that by starting with a fixed real valued random variable \( Z \), the transformation \( X = \mu + Z \) generates location parameter families, and the transformation \( X = \sigma Z \) generates scale parameter families. It is a simple leap from here to combine these transformations to generate location-scale parameter families.

Definition 2.8. A continuous real valued random variable \( X \) is said to have a location-scale parameter density if it has a density of the form \( f(x|\mu, \sigma) = \frac{1}{\sigma} f_0(\frac{x-\mu}{\sigma}), -\infty < x < \infty \), for some fixed density function \( f_0 \) on the real line. Here, \( \mu \) can take any value in the real line.
A continuous random vector \( X_{p \times 1} \) is said to have a *location-scale parameter density* with parameters \( \mu, \Sigma \) if it has a density of the form

\[
 f(x|\mu, \Sigma) = \frac{1}{|\Sigma^{1/2}|} f_0(\Sigma^{1/2}(x - \mu)), \quad x \in \mathbb{R}^p,
\]

for some fixed density function \( f_0 \) on \( \mathbb{R}^p \). Here, \( \mu \) can be any point in \( \mathbb{R}^p \) and \( \Sigma_{p \times p} \) is positive definite.

In terms of random variables, a real valued random variable \( X \) with a location-scale parameter density arises as \( X = \mu + \sigma Z \) for some fixed random variable \( Z \); a random vector \( X \) with a location-scale parameter density arises as \( X = \mu + \Sigma^{1/2}Z \) for some fixed random vector \( Z \).

Here is a quick example.

**Example 2.23.** Each of the following is a location-scale parameter density on the real line:

\[
 N(\mu, \sigma^2); \quad C(\mu, \sigma); \quad U[\mu - \sigma, \mu + \sigma]; \quad \text{Double Exponential}(\mu, \sigma).
\]

The corresponding \( Z \) (the fixed random variable you start with) in the four cases are \( N(0, 1), C(0, 1), U[-1, 1] \) and Double Exponential\((0, 1)\).

On the other hand, \( G(\alpha, \lambda) \) (the general gamma densities) is *not a location-scale parameter density*, \( \lambda \) is indeed a scale parameter, but \( \alpha \) is not a location parameter (usually, we refer to \( \alpha \) as a shape parameter).

In \( p \) dimensions, the multivariate normal distributions \( N_p(\mu, \Sigma) \) are the most well known location-scale parameter densities. Of course, there are many others!

### 2.2 Departures from Normality

#### 2.2.1 Heavy Tails

Because of the fortunate fact that exact optimal estimates and other types of optimal methods are generally possible to obtain for normal distributions, and often not so for nonnormal distributions, it is common practice to assume that a random variable of interest has a normal distribution. We do this both in practice and in proving theorems. However, the reality is that real data actually often show departures from normality. Such departures can manifest in many ways, e.g., heavier tails than normal; normal densities are quite tightly concentrated around the mean \( \mu \). The entire area of robustness was materially inspired by concerns about departures from normality in the heaviness of tails. We present below a brief exposure to these issues within the context of modelling. We start with heavy tails.

What is a heavy tailed distribution on the real line? There is no universally agreed definition for it. But people agree on some general properties that heavy tailed distributions should have. A distribution can be heavy tailed only in the upper tail, or only in the lower
tail, or both. Because our discussion here is brief, we limit ourselves to symmetric heavy tailed distributions; in the symmetric case, if a distribution is heavy tailed on one side, it is automatically heavy tailed on the other side. This is conceptually a simpler case to address.

Here is one possible definition of a symmetric heavy tailed distribution.

**Definition 2.9.** A continuous real valued random variable $Z$ with a symmetric density $f_0(z)$ and CDF $F_0(z)$ is said to be **heavy tailed** if for some $\alpha, 0 < \alpha < \infty$, $1 - F_0(z) = P(Z > z) \sim \frac{1}{z^\alpha}$ as $z \to \infty$. Here, the notation $\sim$ means that $\frac{1-F_0(z)}{z^\alpha}$ has a finite nonzero limit as $z \to \infty$.

The definition adopted above leads to good examples. Here is one example.

**Example 2.24.** Let $Z$ have a $t$ distribution with $\alpha$ degrees of freedom, $\alpha > 0$, with the density

$$f_0(z) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{(1 + x^2/\alpha)^{(\alpha+1)/2}}, -\infty < z < \infty.$$  

It is possible to write a complicated exact formula for the CDF $F_0(z)$. But, it is easy to see directly that

$$1 - F_0(z) = \text{constant} \times \int_z^\infty \frac{1}{(1 + x^2/\alpha)^{(\alpha+1)/2}} dx \sim \int_z^\infty \frac{1}{x^{\alpha+1}} dx \sim z^{-\alpha},$$

as $z \to \infty$. Thus, all $t$ distributions are heavy tailed, according to the definition adopted above.

Note that it makes sense to call $t$ distributions heavy tailed. A distribution with heavy tails will carry substantial probabilities of attaining large values. But such distributions
will have a hard time to make its moments, at least the high order moments, finite. Indeed, it can be proved easily that if $Z$ has the $t$ distribution as above, $E(|Z|^k) < \infty$ if and only if $k < \alpha$. An important special case is $C(0, 1)$, the standard Cauchy, which is the same as a $t$ distribution with $\alpha = 1$ degree of freedom, and for a $C(0, 1)$ random variable, $E(|Z|) = \infty$.

You need to be aware that an assumption of normality can be very dangerous for inference if in fact the random variable in question has a heavy tailed distribution.

### 2.2.2 Skewed Densities

Another common and important manifestation of departure from normality is skewness. Skewness is quite common for random variables which, physically, can take only positive values. A notoriously famous example of a positive random variable with a skewed distribution is income.

If a random variable $X$ with mean $\mu$ has three finite moments, and if the distribution of $X$ is symmetric about $\mu$, then it is easily proved that $E(X - \mu)^3 = 0$. In general, $E(X - \mu)^3$ is of course not zero, and positive values of $E(X - \mu)^3$ usually occur for random variables with a density that has a long right tail. Statisticians conventionally use a scaled version of $E(X - \mu)^3$, defined as $\beta = \frac{E(X - \mu)^3}{\sigma^3}$, to measure the extent and direction of skewness of a distribution; it is called the coefficient of skewness. In the above, $\sigma$ is the standard deviation of $X$.

Skewed densities can be generated in many ways. There are some standard parametric families of densities which result in skewed densities; separately, skewed densities can be manufactured by cleverly manipulating symmetric densities. Here, we will limit ourselves to particular examples.

#### Example 2.25. (Lognormal and Inverse Gaussian)

Two particular parametric families of skewed densities are the lognormal and the inverse Gaussian. $X \geq 0$ has a lognormal density with parameters $\mu$ and $\sigma$ if (and only if) $\log X \sim N(\mu, \sigma^2)$. The coefficient of skewness is $\beta = (2 + e^{\sigma^2})\sqrt{e^{\sigma^2} - 1}$, which can be arbitrarily large. The inverse Gaussian distribution with parameters $\theta$ and $\lambda$ has the density

$$f(x|\theta, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi x^3/2}} e^{-\lambda(x-\mu)^2/(2\mu^2x)}, x > 0.$$  

The coefficient of skewness is $\beta = \frac{3\sqrt{\pi}}{\sqrt{\lambda}}$, and once again, it can be arbitrarily large.

#### Example 2.26. (Skew Normal Distributions)

Skew normal distributions originated in the work of O'Hagan and Leonhard (1976), and have become a popular parametric model or modelling mildly skewed distributions with support on the whole real line. The standard skew normal density on $(-\infty, \infty)$ is $f_0(x|\alpha) = 2\phi(x)\Phi(\alpha x)$, where $\phi, \Phi$ are the
standard normal density and CDF, and \( \alpha \) is a general real number. It is interesting that whatever be \( \alpha, f_0(x|\alpha) \) integrates to one, namely, \( \int_{-\infty}^{\infty} f_0(x|\alpha) \, dx = 1 \). The general skew normal density is obtained from the standard skew normal density by introducing a location parameter \( \mu \) and a scale parameter \( \sigma \):

\[
f(x|\alpha, \mu, \sigma) = \frac{1}{\sigma} f_0(x - \frac{\mu}{\sigma}), -\infty < x < \infty.
\]

The case \( \alpha = 0 \) corresponds to the \( N(\mu, \sigma^2) \) density, which, of course, is symmetric. For \( \alpha > 0 \), the density is skewed to the right, and likewise, for \( \alpha < 0 \), it is skewed to the left.

### 2.2.3 Multimodality and Mixture Models

Nearly all standard parametric distributions that we know by name are unimodal. Foremost among them are normal distributions. However, whenever the population from which one samples has a heterogeneous mixed character, variables may have a bimodal or even a multimodal distribution. We had casually discussed this in Section 1.2.1 while trying to model heights of people in a population.

A convenient method for generating bimodal or multimodal densities is to start with two or in general some \( k \) uniimodal densities, and take a mixture of these densities. For example, if \( f_1, f_2 \) are two densities, then \(.5f_1 + .5f_2 \) is a 50 – 50 mixture of \( f_1 \) and \( f_2 \). If the component densities \( f_1, f_2 \) are well separated in the location of their centers, then the mixture density has a good chance of producing a bimodal or multimodal shape. This, of course, is a rough statement, but it can be made precise. We give a brief introduction to mixture densities here, so you will be aware of them and their potential uses in practice.
Definition 2.10. Given $k$ density functions $f_1, f_2, \cdots, f_k$, and constants $p_1, p_2, \cdots, p_k > 0$ such that $\sum_{i=1}^{k} p_i = 1$, the density function $f(x) = \sum_{i=1}^{k} p_i f_i(x)$ is called a mixture of $f_1, f_2, \cdots, f_k$ with mixing proportions $p_1, p_2, \cdots, p_k$.

The most common mixture densities that we use are normal mixtures. Normal mixtures do produce nonunimodal shapes unless the means of the component normal densities are close to each other. For example, a 50–50 mixture of $N(\mu_1, 1)$ and $N(\mu_2, 1)$ is bimodal if $|\mu_1 - \mu_2| > 1$; plot of such a normal mixture nicely shows the bimodality.

We mention a pleasant theorem on the unimodality behavior of normal mixtures; it is a practically useful result. The result is a little hard to comprehend at first. But it says something intuitive; it says that the mixture density is bimodal if the means $\mu_1, \mu_2$ are sufficiently far apart, and if neither of the two component normal densities dominates the overall mixture density. A proof may be seen in Robertson and Fryer (1969).

Theorem 2.1. Let $f_i(x), i = 1, 2$ be the densities of $N(\mu_i, \sigma_i^2), i = 1, 2, \mu_1 \leq \mu_2$, and $f(x)$ the general two component univariate normal mixture density $f(x) = pf_1(x) + (1-p)f_2(x)$.

Define

$$\tau = \frac{\sigma_2}{\sigma_1}, \delta = \frac{\mu_2 - \mu_1}{\sigma_1}, \delta_0 = \left[2(\tau^4 - \tau^2 + 1)^{3/2} - (2\tau^6 - 3\tau^4 - 3\tau^2 + 2)\right]^{1/2}/\tau,$$

and let $y_1 < y_2$ be the two roots of the equation

$$(\tau^2 - 1)y^3 - \delta(\tau^2 - 2)y^2 - \delta^2 y + \delta^2 \tau^2 = 0$$
in the interval $0 < y < \delta$. Define also $p_1, p_2$ by the equations

$$p_i^{-1} = 1 + \frac{\tau^3 y_i}{\delta - y_i} e^{-y_i^2/2 + (y_i - \delta)^2/2\tau^2}, i = 1, 2.$$ 

Then, the mixture density $f$ is bimodal if the following two conditions are satisfied:

$$1) \delta > \delta_0; \quad 2) p_1 < p < p_2.$$ 

A neat corollary of this theorem is the following.

**Corollary 2.1.** If $\sigma_1 = \sigma_2 = \sigma$ and $p = \frac{1}{2}$, then the mixture density $f$ is bimodal if and only if $|\mu_1 - \mu_2| / \sigma > 1$. If a random variable $X$ has a bimodal or multimodal density, and we mistakenly assume it to be unimodal, say normal, then there would be some negative effect on the quality of inference. There has not been nearly as much research on it, as there is on effect of heavy tails or skewness on the quality of inference that is based on an erroneous assumption of normality. Bi or multimodality, when present, typically tells us something about the heterogeneous composition of the population. If we miss it in our modeling stage, we lose that potentially useful information. Campbell (1984) and Eisenberger (1964) are two easy to read references on uses of mixtures and multimodality; Lindsay and Basak (1993) is a simple yet useful proposal for identification of the mixture structure by using moments.

### 2.3 Linear and Generalized Linear Models

The word *model* actually means much more than a distribution for a random variable. A model can mean very different things in different contexts. Let us see a simple example. The concept of *regression* was invented by Francis Galton in the context of a famous dataset on heights of parents and children. We mentioned this dataset in passing in Section 1.2.4. The general idea of regression is easy to understand. We have two variables, say $X$ and $Y$, and in some sense $X$ has a causal effect on $Y$, rather than the other way around. For example, between the two variables, height and weight, we tend to think of height affecting weight, rather than weight affecting height. Suppose $X = x$, a specific value. Knowing this, what can we say about $Y$? Obviously, we cannot say anything about $Y$ with total certainty just because we know that $X = x$. But, perhaps, we can say something about the average value of $Y$ when $X = x$, namely, $E(Y|X = x)$. If we denote $m(x) = E(Y|X = x)$, then $m(x)$ is called the *regression function* of $Y$ on $X$, and writing a functional form for $m(x)$ would correspond to choosing a model in this context. You can see that the word *model* means an entirely different thing now!

The simplest type of regression function is a linear function, $m(x) = \beta_0 + \beta_1 x$ for some parameters $\beta_0, \beta_1$. A slightly more complex regression function would be a quadratic
function, \( m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \); but, this is also a linear model, if we simply call \( x = x_1 \) and \( x^2 = x_2 \). With this notation, we get a linear model again, \( m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \). Most generally, a linear model of this kind can be \( m(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \); we call \( x_1, x_2, \ldots, x_p \) covariates, and the \( p + 1 \) parameters \( \beta_i, 0 \leq i \leq p \) regression coefficients. By writing \( x = (1, x_1, \ldots, x_p)' \) and \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)' \), we can write this in the equivalent vector notation \( m(x) = x'\beta \). Because we have modelled the regression function \( m(x) \) this way by using a finite number of parameters, a linear model is considered to be a parametric model for a regression function.

The obvious next question is, is there something more general than a straight linear model for a regression function. As you may have anticipated, the answer is yes. A very well known and useful generalization is to say that \( m(x) = G(x'\beta) \) for some function \( G: \mathbb{R} \to \mathbb{R} \). Of course, the linear model corresponds to the special function \( G(u) = u \). The model \( m(x) = G(x'\beta) \) for some function \( G \) is called a generalized linear model; the function \( G \) is called the link function and is chosen depending on the particular context. For example, a popular link function in some regression problems with categorical response variables is the logit link function \( G(u) = \frac{e^u}{1+e^u} \). Notice that for given \( G \), this is still a parametric model. A classic introductory reference is McCullagh and Nelder (1989).

### 2.4 Nonparametric and Semiparametric Models

In Example 2.6, we saw an early example of what the word nonparametric is supposed to mean. A nontechnical way to understand the difference between parametric and nonparametric models is this: if the unknown parts of the chosen model in your particular problem can be completely described in terms of a finite number of unknown parameters, then the model is viewed as a parametric model for that problem. Otherwise, the model is viewed as an infinite parametric or nonparametric model. It is difficult to give a clearcut and universal definition of a nonparametric model that you can apply in every situation; but given a problem, and a model, usually it is easy to characterize the model as parametric, nonparametric, or something in between, such as semiparametric. Examples will help us understand this terminology.

Why consider nonparametric models at all? To answer this, we need to have a brief discussion about the pros and cons of parametric modelling vis-à-vis nonparametric modelling of an inference problem. Take a simple modelling problem to understand this issue. Suppose the modelling task at hand is to choose a distribution for a real valued random variable \( X \). Guided by tradition and also convenience, you may model it as \( X \sim N(\mu, \sigma^2) \). This is of course a parametric model, and even more, a parametric model with only two parameters. But, that does not mean that it cannot be a useful model for your particular problem. The pros of your normal model are that it is parsimonious,
both parameters have physically meaningful meanings (mean and variance), you already have a lot of known theory in place for normal distributions, and estimation of functions, \( g(\mu, \sigma) \), is relatively straightforward under your parametric model. On the other hand, by choosing a normal model, you have automatically assumed light tails, symmetry, unimodality, and most of all, a rigid functional form for your density. What if some of these assumptions are either false, or not testable from your data? Nonparametric modelling offers to give you some protection from making myriad rigid assumptions. If you forsake normality, and make much weaker assumptions, for example only that your underlying density \( f \) is uniformly bounded and smooth, say once or twice differentiable, then you have significantly loosened the normality assumption. You are no longer assuming any rigid functional form for your density \( f \). This would be a nonparametric model for this problem.

What are the cons of nonparametric modelling? The downsides of nonparametric modelling are that in order to be overcautious about potentially risky parametric modelling, you may have taken too limp a model. It is a bit like no risks, no gains in making investments. In nonparametric models, it requires more data to achieve comparable accuracy to what you can achieve using parametric models, if the parametric model was reasonably accurate. Another difficulty with nonparametric modelling is that in multivariate situations, it is really difficult to carry on with purely nonparametric inference unless you have a tremendous amount of data.

Let us see an illustrative example of some nonparametric models.

Example 2.27. (Parametric vs. Nonparametric Models) Suppose we need to choose a distribution for a positive random variable \( X \). If we assume that \( X \) has an exponential distribution, \( X \sim \text{Exp}(\lambda), \lambda > 0 \) being a parameter, then that is a parametric model. On the other hand, if we assume that \( X \) has a density \( f(x) \) such that \( f \) is monotone decreasing on \([0, \infty)\) and \( f(0) < 0\), then that would be a nonparametric model. We can also choose a model that is less rigid than exponential, but makes more assumptions than our nonparametric model. For example, we may choose a hierarchical model, such as \( X|\lambda \sim \text{Exp}(\lambda) \) and \( \lambda \sim g(\lambda) \). Then, our final model for the density of \( X \) is \( f(x) = \int_0^\infty \frac{1}{\lambda} e^{-x/\lambda} g(\lambda) d\lambda \). We may call this a semiparametric model.

Consider next a regression problem, in which there are two covariates \( X_1, X_2 \), and a single response variable \( Y \). A standard linear model \( m(x_1, x_2) = E(Y | X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \) is a parametric model for the regression function \( m(x_1, x_2) \). On the other hand, if we only assume that \( m(x_1, x_2) \) is a sufficiently smooth function, for example that it has two or three continuous partial derivatives, that would be a nonparametric model for the regression function \( m(x_1, x_2) \). We can choose a model that is somewhere in between. One example is a partially linear model, which sets \( m(x_1, x_2) = \beta_0 + \beta_1 x_1 + g(x_2) \),
where \( g(x_2) \) is a sufficiently smooth function, for example that it is twice continuously differentiable. As you can see, there is a lot of scope for writing models of different degrees of rigidity in a given inference problem. A lucid introduction to nonparametric modelling is Hddotardle et. al (2004).

2.5 At Instructor’s Discretion

2.5.1 Parameters Growing with \( n \)

In the traditional iid parametric model, say \( X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \), the number of parameters stays fixed no matter how large the sample size is. In fact, this is a crucial advantage in the standard parametric modelling; we have accumulating sample information without a growth in the number of parameters. There are, however, some problems in which the number of parameters grows with the sample size; basically, each new sampling unit brings with it one or more new parameters. You can instinctively see that unless these parameters are all somehow similar, we cannot estimate all of them.

Here are two examples of models in which the number of parameters grows with \( n \). One example is the Gaussian sequence model \( X_i \sim N(\mu_i, 1), i = 1, 2, \cdots, n \). We will treat the Gaussian sequence model in detail in Chapter 14. The Gaussian sequence model may arise in the following way. Suppose \( n \) transmissions (say signals) from some source are recorded on an instrument, and the \( i \)th signal has mean \( \mu_i \). Typically, you believe that most of the transmissions are pure noises, i.e., most of the \( \mu_i \) are zero, but you do not know which ones. You want to identify the nonzero \( \mu_i \)'s and estimate them. This is a difficult problem.

Another well known example of a model with a growing number of parameters is the Rasch model, a famous model in psychometrics. The Rasch model is used to model cognitive abilities of individuals in a testing situation. Suppose \( n \) individuals are each taking a true-false examination consisting of \( k \) questions (called items). A correct answer corresponds to a response \( Y = 1 \) and a wrong answer corresponds to \( Y = 0 \). The Rasch model for the probability of a correct response is

\[
P(Y_{ij} = 1) = \frac{e^{\delta_i - \tau_j}}{1 + e^{\delta_i - \tau_j}}, \quad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, k;
\]

\( \delta_i \) measures the ability of the \( i \)th person and \( \tau_j \) measures the difficulty of the \( j \)th question.

Once again, due to the growing number of parameters, estimation in the Rasch model is difficult; see, for example, Ghosh (1995).

2.5.2 Models for Time Series Data

A sequence of sample observations \( X_1, X_2, \cdots \) are not always independent. In numerous practical problems, the observations have some sort of mutual dependence. This is particularly the case when the sequence of data values have an inherent time character, such
as hourly data on temperature, or closing values of a stock on successive days. It would be foolish to model such data as iid, or independent, even if not iid.

The subject of time series was specifically constructed to model such dependence when the sample observations have a time character. Time series is now a vast area within the subject of statistics and numerous models with associated extensive theory and methods exist. As usual, our intention at this stage is just to get acquainted with a few models.

**Example 2.28. (Time Series Models)** A mathematical formulation for the concept that each sample value depends on the sample value at the previous time in a linear fashion is one of the most basic time series models. An *autoregressive process of order 1*, written as $AR(1)$, is defined through the system of equations $X(t) = \rho X(t-1) + \epsilon(t)$, where $\rho$ is a parameter, and $\epsilon(t)$ are random errors. We have to make suitable assumptions about $\rho$ and the random errors $\epsilon(t)$ in order to make further analysis sensible. Typically, we assume that $|\rho| < 1$, and often we assume that the random errors $\epsilon(t)$ are iid normal with zero means, and a possibly unknown variance $\sigma^2$. Natural generalizations of an $AR(1)$ process are the processes, called $AR(p)$ processes, defined through $X(t) = \sum_{i=1}^{p} \rho_i X(t-i)$. Once again, we must make suitable assumptions about the parameters $\rho_i$ in order to make further analysis sensible, but we will not enter into those details at this stage. For autoregressive processes, the correlation at lag $k$, defined as $\text{corr}(X(t), X(t+k))$ goes to zero at an exponentially fast rate with the lag $k$. Thus, autoregressive processes are popular for modelling *relatively weak dependence*.

Another approach to modelling dependence in time series data is to use the model $X(t) = \sum_{i=0}^{q} \theta_i \epsilon(t-i)$, where the errors $\epsilon(t)$ are latent random variables; we do not observe their values in a practical problem. They are used only to model the mutual dependence among the sequence of data values $X(t)$. Such processes are called $MA(q)$, *moving average of order q*. We must, again, make suitable assumptions about the parameters $\theta_i$, and the latent variables $\epsilon(t)$. Often, we may take the $\epsilon(t)$ to be iid $N(0, \sigma^2)$. Moving average processes have the *killed correlation property* $\text{corr}(X(t), X(t+k)) = 0$ if the lag $k$ is larger than the order $q$ of the moving average process. There are also very popular time series models that incorporate the structure of both autoregressive and moving average processes; they are called *autoregressive moving average processes*, written as $ARMA(p,q)$.

Excellent introductory references include Brockwell and Davis (1991) and Hamilton (1994).

### 2.5.3 Multivariate Models

### 2.6 Exercises

**Exercise 2.1.** Where does a statistical model for a problem come from?
Exercise 2.2. (Writing a Model). For each of the following random variables, choose a parametric family of distributions, mixture distributions included, list all the parameters in your model, and describe the parameter space:
(a) Total cholesterol level of 30 – 45 year old women;
(b) Total cholesterol level of women 30 and above;
(c) Number of cigarettes smoked per day by a male smoker;
(d) Lengths of telephone calls from US to China.

Exercise 2.3. (Basic). Suppose $X_1, \cdots, X_n$ are iid observations from $Poi(\lambda)$. Classify each of the following as a parameter, statistic, or functional:

$$\bar{X} - 2; \frac{s^2}{\lambda}; P(s^2 > \lambda); |\bar{X} - s^2|; E(\bar{X} - s^2).$$

Exercise 2.4. (Conceptual). Is every statistic also automatically a functional?

Exercise 2.5. (Think). Suppose $X_1, \cdots, X_n$ are iid observations from $N(\mu, 1)$. Give an example of a statistic as well as a functional which has zero mean under all values of $\mu$.

Exercise 2.6. Suppose $X \sim N(\mu, 1)$ and suppose $Y = X$ if $X > \mu$ and $-X$ if $X \leq \mu$. Is the distribution of $X$ regular or nonregular?

Exercise 2.7. Suppose $X \sim U[1,h(\theta)]$, where $\theta$ is a parameter and $h$ is a general function. Can the distribution of $X$ ever be regular?

Exercise 2.8. Show that the family of exponential densities, $Exp(\lambda)$ is a one parameter exponential family.

Exercise 2.9. (A Little Tricky). Show that the family of $U[0, \theta]$ densities is not an exponential family.

Exercise 2.10. Suppose a positive random variable $X$ has a distribution in the one parameter exponential family. Prove or disprove that $X^2$ also has a distribution in the one parameter exponential family.

Exercise 2.11. Show that the family of double exponential densities with parameters $\mu, \sigma$ is a location-scale parameter family.

Exercise 2.12. (Skills). One of the following is a scale parameter family; identify which one.

$N(\theta, \theta); N(\theta, \theta^2); Poi(\theta)$.

Exercise 2.13. (Conceptual). Can a parameter $\theta$ be simultaneously a location and a scale parameter? Either give an example, or prove that it is not possible.
Exercise 2.14. Give an example of a $p$-dimensional density other than the multivariate normal which is a location-scale parameter family.

Exercise 2.15. (Think). Give an example of a location-scale parameter family of densities on the real line which does not have any moments.

Exercise 2.16. Suppose a real valued random variable $X$ has a location-scale parameter density. Prove that $P(\mu - k_1\sigma \leq X \leq \mu + k_2\sigma)$ is a fixed constant for any given $k_1, k_2$.

Exercise 2.17. (Calculation). Derive a formula for the coefficient of skewness of the inverse Gaussian distribution.

Exercise 2.18. (Calculation). Derive a formula for the coefficient of skewness of the Gamma distribution.

Exercise 2.19. (Calculation). Derive a formula for the coefficient of skewness of a general two component normal mixture.

Exercise 2.20. (Surprise). Show that every moment of a lognormal distribution is finite, but that the mgf does not exist in any interval around zero.

Exercise 2.21. (Identifiability). A parametric family of distributions $\{P_\theta\}$ is called identifiable if whenever we take $\theta_1 \neq \theta_2$, the corresponding distributions $P_{\theta_1}$ and $P_{\theta_2}$ are also two different distributions.

Prove that the mixture normal densities $pN(\mu_1, 1) + (1 - p)N(\mu_2, 1)$ form an identifiable family.

Exercise 2.22. (Identifiability). Prove by construction that the family of mixtures of uniform densities $pU[0, \theta] + (1 - p)U[\theta, 1]$ is not an identifiable family.

Exercise 2.23. (Very Tricky). Give an example of a density on $(0, \infty)$ such that $E(X^\alpha) = \infty$ for all $\alpha > 0$.

Exercise 2.24. (Anticipated). Give a proof that a mixture of two arbitrary normal densities cannot have three or more local maxima.

Exercise 2.25. (Anticipated). Give a proof that a mixture of two arbitrary normal densities cannot be a normal density.

Exercise 2.26. Identify all values of $p$ such that $pN(0, 1) + (1 - p)N(3, 1)$ is bimodal.

Exercise 2.27. Give a proof that a mixture of two arbitrary exponential densities is necessarily unimodal.
Exercise 2.28. (Normal Variance Mixtures). Prove that the standard Cauchy density is a mixture of $N(0, \sigma^2)$ densities.
Is this true of all Cauchy densities centered at zero?

Exercise 2.29. (Probit Model). Consider regression of a binary variable $Y$, taking only the values 0, 1, on $p$ covariates $X_1, \ldots, X_p$. Consider the generalized linear model for this problem with the link function $G(u) = \Phi(u)$, the standard normal CDF. Write out explicitly $P(Y = 1|X_1 = x_1, \ldots, X_p = x_p)$.

Exercise 2.30. (GLM for Poisson Data). Consider regression of a Poisson variable $Y$ on $p$ covariates $X_1, \ldots, X_p$. Consider the generalized linear model for this problem, and suggest a link function; keep in mind that Poisson variables are nonnegative, and so the mean is positive.

Exercise 2.31. (Moving Averages). Calculate in analytical form the correlation between $X(t)$ and $X(t + k)$ for an MA(1) process, if $\epsilon(t)$ are iid $N(0,1)$.

Exercise 2.32. (AR(1) Process). Calculate in analytical form the correlation between $X(t)$ and $X(t + k)$ for an AR(1) process, if $\epsilon(t)$ are iid $N(0,1)$.

Exercise 2.33. Suppose $X(t) = \epsilon(t)\epsilon(t - 1)$, where $\epsilon(t)$ are iid $N(0,1)$. Calculate in analytical form the correlation between $X(t)$ and $X(t + k)$; does it depend on $t$?

Exercise 2.34. (a) Plot the following time series:

54, 43, 67, 49, 50, 44, 44, 58, 49, 60, 40

(b) Make an intelligent guess for the correlation between $X(t)$ and $X(t + k)$, $k = 1, 2$, assuming that they do not depend on $t$.

Exercise 2.35. (Conceptual). What will go wrong if $|\rho| > 1$ in the AR(1) model?

Exercise 2.36. (Plotting Simulated Data). (a) Simulate an AR(1) process of length 100 and plot it, by assuming $\rho = 0.6$ and that $\epsilon(t)$ are iid $N(0,1)$.

(b) Make an intelligent guess for the correlation between $X(t)$ and $X(t + 1)$; now, compare it with the theoretical value of the correlation between $X(t)$ and $X(t + 1)$.

2.7 References

York.