

## Chapter 2

# Background: Probability

### 2.1 Summary

We review the fundamental tools used to establish the inferential basis for our models. Results are stated as theorems, lemmas and corollaries. Most of the key proofs are provided in Chapter 16 although, sometimes, when useful to the general development, proofs are given within the text itself. The main ideas of stochastic processes, in particular Brownian motion and functions of Brownian motion, are explained in non-measure-theoretic terms. The background to this, i.e., distribution theory and large sample results, is recalled. Rank invariance is an important concept, i.e., the ability to transform some variable, usually time, via monotonic increasing transformations without having an impact on inference. These ideas hinge on the theory of order statistics and the basic notions of this theory are recalled. An outline of the theory of counting processes and martingales is presented without leaning upon measure-theoretic constructions. The important concepts of explained variation and explained randomness are outlined in elementary terms, i.e., only with reference to random variables and, at least initially, making no explicit appeal to any particular model. This is important since the concepts are hardly any less fundamental than a concept such as variance itself. They ought therefore stand alone, and not require derivation as a particular feature of some model. In practice, of course, we may need estimate conditional distributions and making an appeal to a model at this point is quite natural.

## 2.2 Motivation

The last few decades have seen the topic of survival analysis become increasingly specialized, having a supporting structure based on large numbers of theorems and results which appear to have little application outside of the field. Many recently trained specialists, lacking a good enough grasp of how the field relates to many others, are left with little option but to push this specialization yet further. The result is a field which is becoming largely inaccessible to statisticians from other areas. A key motivation of this work, and this chapter in particular, is to put some brakes on this trend by leaning on classical results. Most of these are well known, others less so, and in this chapter we cover the main techniques from probability and statistics which we will need. Results are not simply presented and the aim is to motivate them from elementary principles known to those with a rudimentary background in calculus.

## 2.3 Integration and measure

The reader is assumed to have some elementary knowledge of set theory and calculus. We do not recall here any of the basic notions concerning limits, continuity, differentiability, convergence of infinite series, Taylor series and so on and the rusty reader may want to refer to any of the many standard calculus texts when necessary. One central result which is frequently called upon is the mean value theorem. This can be deduced as an immediate consequence to the following result known as Rolle's theorem.

**Theorem 2.1** *If  $f(x)$  is continuously differentiable at all interior points of the interval  $[a, b]$  and  $f(a) = f(b)$ , then there exists a real number  $\xi \in (a, b)$  such that  $f'(\xi) = 0$ .*

A simple sketch would back up our intuition that the theorem would be correct. Simple though the result appears to be, it has many powerful implications including;

**Theorem 2.2** *If  $f(x)$  is continuously differentiable on the interval  $[a, b]$ , then there exists a real number  $\xi \in (a, b)$  such that*

$$f(b) = f(a) + (b - a)f'(\xi).$$

When  $f(x)$  is monotone then  $\xi$  is unique. This elementary theorem can form the basis for approximation theory and series expansions such as the Edgeworth and Cornish-Fisher (see Section 2.9). For example, a further immediate corollary to the above theorem obtains by expanding in turn  $f'(\xi)$  about  $f'(a)$  whereby:

**Corollary 2.1** *If  $f(x)$  is at least twice differentiable on the interval  $[a, b]$  then there exists a real number  $\xi \in (a, b)$  such that*

$$f(b) = f(a) + (b - a)f'(a) + \frac{(b - a)^2}{2}f''(\xi).$$

The  $\xi$  of the theorems and corollary would not typically be the same and we can clearly continue the process, resulting in an expansion of  $m + 1$  terms, the last term being the  $m$ th derivative of  $f(x)$ , evaluated at some point  $\xi \in (a, b)$  and multiplied by  $(b - a)^m/m!$ . An understanding of Riemann integrals as limits of sums, definite and indefinite integrals, is mostly all that is required to follow the text. It is enough to know that we can often interchange the limiting processes of integration and differentiation. The precise conditions for this to be valid are not emphasized. Indeed, we almost entirely avoid the tools of real analysis. The Lebesgue theory of measure and integration is on occasion referred to, but a lack of knowledge of this will not hinder the reader. Likewise we will not dig deeply into the measure-theoretic aspects of the Riemann-Stieltjes integral apart from the following extremely useful construction:

**Definition 2.1** *The Riemann integral of the function  $f(x)$  with respect to  $x$ , on the interval  $[a, b]$ , is the limit of a sum  $\sum \Delta_i f(x_{i-1})$ , where  $\Delta_i = x_i - x_{i-1} > 0$ , for an increasing partition of  $[a, b]$  in which  $\max \Delta_i$  goes to zero.*

The limit is written  $\int_a^b f(x)dx$  and can be seen to be the area under the curve  $f(x)$  between  $a$  and  $b$ . If  $b = \infty$  then we understand the integral to exist if the limit exists for any  $b > 0$ , the result itself converging to a limit as  $b \rightarrow \infty$ . Similarly for  $a = -\infty$ . Now, instead of only considering small increments in  $x$ , i.e., integrating with respect to  $x$ , we can make use of a more general definition. We have:

**Definition 2.2** *The Riemann-Stieltjes integral of the function  $f(x)$  with respect to  $g(x)$  is the limit of a sum  $\sum \{g(x_i) - g(x_{i-1})\}f(x_{i-1})$ , for an increasing partition of  $[a, b]$  in which, once again,  $\max \Delta_i$  goes to zero.*

The limit is written  $\int_a^b f(x)dg(x)$  and, in the special case where  $g(x) = x$ , reduces to the usual Riemann integral. For functions, necessarily continuous, whereby  $g(x)$  is an antiderivative of, say,  $h(x)$  and can be written  $g(x) = \int_{-\infty}^x h(u)du$  then the Stieltjes integral coincides with the Riemann integral  $\int f(x)h(x)dx$ . On the other hand whenever  $g(x)$  is a step function with a finite or a countable number of discontinuities then  $\int f(x)dg(x)$  reduces to a sum, the only contributions arising at the discontinuities themselves. This is of great importance in statistical applications where step functions naturally arise as estimators of key functions. A clear example of a step function of central importance is the empirical distribution function,  $F_n(x)$  (this is discussed in detail in Chapter 3). We can then write the sample mean  $\bar{x} = \int u dF_n(u)$  and the population mean  $\mu = \int u dF(u)$ , highlighting an important concept, that fluctuations in the sample mean can be considered a consequence of fluctuations in  $F_n(x)$  as an estimate of  $F(x)$ . Consider the following theorem, somewhat out of sequence in the text but worth seeing here for its motivational value. The reader may wish to take a glance ahead at Sections 2.4 and 3.5.

**Theorem 2.3** *For every bounded continuous function  $h(x)$ , if  $F_n(x)$  converges in distribution to  $F(x)$ , then  $\int h(x)dF_n(x)$  converges in distribution to  $\int h(x)dF(x)$ .*

This is the Helly-Bray theorem. The theorem will also hold (see the Exercises) when  $h(x)$  is unbounded provided that some broad conditions are met. A deep study of  $F_n(x)$  as an estimator of  $F(x)$  is then all that is needed to obtain insight into the sample behavior of the empirical mean, the empirical variance and many other quantities. Of particular importance for the applications of interest to us here, and developed, albeit very briefly, in Section 2.12, is the fact that, letting  $M(x) = F_n(x) - F(x)$ , then

$$E \left\{ \int h(x)dM(x) \right\} = \int h(x)dF(x) - \int h(x)dF(x) = 0, \quad (2.1)$$

a seemingly somewhat innocuous result until we interchange the order of integration (expectation, denoted by  $E$  being an integral operator) and, under some very mild conditions on  $h(x)$  described in Section 2.12, we obtain a formulation of great generality and into which can be fit many statistical problems arising in the context of stochastic processes (see Section 2.12).

## 2.4 Random variables and probability measure

The possible outcomes of any experiment are called events where any event represents some subset of the sample space. The sample space is the collection of all events, in particular the set of elementary events. A random variable  $X$  is a function from the set of outcomes to the real line. A probability measure is a function on some subset of the real line to the interval  $[0,1]$ . Kolmogorov (1933) provided axioms which enable us to identify any measure as being a probability measure. These axioms appear very reasonable and almost self-evident, apart from the last, which concerns assigning probability measure to infinite collections of events. There is, in a well defined sense, many more members in the set of all subsets of any infinite set than in the original set itself, an example being the set of all subsets of the positive integers which has as many members as the real line. This fact would have hampered the development of probability without the inclusion of Kolmogorov's third axiom which, broadly says that the random variable is measurable, or, in other words, that the sample space upon which the probability function is defined is restricted in such a way that the probability we associate with the sum of an infinite collection of mutually exclusive events is the same as the sum of the probabilities associated with each composing event.

A great deal of modern probability theory is based on measure-theoretic questions, questions that essentially arise from the applicability or otherwise of Kolmogorov's third axiom in any given context. This is an area that is highly technical and relatively inaccessible to non-mathematicians, or even to mathematicians lacking a firm grounding in real analysis. The influence of measure theory has been strongly felt in the area of survival analysis over the last 20 or so years and much modern work is now of a very technical nature. Even so, none of the main statistical ideas, or any of the needed demonstrations in this text, require such knowledge. We can therefore largely avoid measure-theoretic arguments, although some of the key ideas that underpin important concepts in stochastic processes are touched upon whenever necessary. The reader is expected to understand the meaning of the term *random variable* on some level.

Observations or outcomes as random variables and, via models, the probabilities we will associate with them are all part of a theoretical, and therefore artificial, construction. The hope is that these probabilities will throw light on real applied problems and it is useful to keep in

mind that, in given contexts, there may be more than one way to set things up. Conditional expectation is a recurring central topic but can arise in ways that we did not originally anticipate. We may naturally think of the conditional expected survival time given that a subject begins the study under, say, some treatment. It may be less natural to think of the conditional expectation of the random variable we use as a treatment indicator given some value of time after the beginning of treatment. Yet, this latter conditional expectation, as we shall see, turns out to be the more relevant for many situations.

### *Convergence for random variables*

Simple geometrical constructions (intervals, balls) are all that are necessary to formalize the concept of convergence of a sequence in real and complex analysis. For random variables there are a number of different kinds of convergence, depending upon which aspect of the random variable we are looking at. Consider any real value  $Z$  and the sequence  $U_n = Z/n$ . We can easily show that  $U_n \rightarrow 0$  as  $n \rightarrow \infty$ . Now let  $U_n$  be defined as before except for values of  $n$  that are prime. Whenever  $n$  is a prime number then  $U_n = 1$ . Even though, as  $n$  becomes large,  $U_n$  is almost always arbitrarily close to zero, a simple definition of convergence would not be adequate and we need consider more carefully the sizes of the relevant sets in order to accurately describe this. Now, suppose that  $Z$  is a uniform random variable on the interval  $(0,1)$ . We can readily calculate the probability that the distance between  $U_n$  and 0 is greater than any arbitrarily small positive number  $\epsilon$  and this number goes to zero with  $n$ . We have convergence in probability. Nonetheless there is something slightly erratic about such convergence, large deviations occurring each time that  $n$  is prime. When possible, we usually prefer a stronger type of convergence. If, for all integer values  $m$  greater than  $n$  and as  $n$  becomes large, we can assert that the probability of the distance between  $U_m$  and 0 being greater than some arbitrarily small positive number goes to zero, then such a mode of convergence is called strong convergence. This stronger convergence is also called convergence with probability one or almost sure convergence. Consider also  $(n+3)U_n$ . This random variable will converge almost surely to the random variable  $Z$ . But, also, we can say that the distribution of  $\log_e(n+3)U_n$ , at all point of continuity  $z$ , becomes arbitrarily close to that of a standard exponential distribution. This is called convergence in distribution. The three modes of convergence are related by:

**Theorem 2.4** *Convergence with probability one implies convergence in probability. Convergence in probability implies convergence in distribution.*

Also, for a sequence that converges in probability, there exists a subsequence that converges with probability one. This latter result requires the tools of measure theory and is not of wide practical applicability since we may not have any obvious way of identifying such a subsequence. In theoretical work it can sometimes be easier to obtain results for weak rather than strong convergence. However, in practical applications, we usually need strong (almost sure, “with probability one”) convergence since this corresponds in a more abstract language to the important idea that, as our information increases, our inferences become more precise.

#### *Convergence of functions of random variables*

In constructing models and establishing inference for them we will frequently appeal to two other sets of results relating to convergence. The first of these is that, for a continuous function  $g(z)$ , if  $Z_n$  converges in probability to  $c$ , then  $g(Z_n)$  converges in probability to  $g(c)$  and, if  $Z_n$  converges in distribution to  $Z$ , then  $g(Z_n)$  converges in distribution to  $g(Z)$ . The second set, Slutsky’s theorem (a proof is given in Randles and Wolf 1979), enables us to combine modes of convergence. In particular, for modeling purposes, if a convergence in distribution result holds when the parameters are known, then it will continue to hold when those same parameters are replaced by consistent estimators. This has great practical value.

## 2.5 Distributions and densities

We anticipate that most readers will have some familiarity with the basic ideas of a distribution function  $F(t) = \Pr(T < t)$ , a density function  $f(t) = dF(t)/dt$ , expectation and conditional expectation, the moments of a random variable and other basic tools. Nonetheless we will go over these elementary notions in the context of survival in the next chapter. We write

$$E \psi(T) = \int \psi(t)f(t)dt = \int \psi(t)dF(t)$$

for the expected value of the function  $\psi(T)$ . Such an expression leaves much unsaid, that  $\psi(t)$  is a function of  $t$  and therefore  $\psi(T)$  itself random, that the integrals exist, the domain of definition of the function being left implicit, and that the density  $f(t)$  is an anti-derivative of the cumulative distribution  $F(t)$  (in fact, a slightly weaker mathematical construct, absolute continuity, is enough but we do not feel the stronger assumption has any significant cost attached to it). There is a wealth of solid references for the rusty reader on these topics, among which Billingsley (1968), Rao (1973), and Serfling (1980) are particularly outstanding. It is very common to wish to consider some transformation of a random variable, the simplest situation being that of a change in origin or scale. The distribution of sums of random variables arises by extension to the bivariate and multivariate cases.

**Theorem 2.5** *Suppose that the distribution of  $X$  is  $F(x)$  and that  $F'(x) = f(x)$ . Suppose that  $y = \phi(x)$  is a monotonic function of  $x$  and that  $\phi^{-1}(y) = x$ . Then, if the distribution of  $Y$  is  $G(y)$  and  $G'(y) = g(y)$ ,*

$$G(y) = F\{\phi^{-1}(y)\}; \quad g(y) = f\{\phi^{-1}(y)\} \left| \frac{d\phi(x)}{dx} \right|_{x=\phi^{-1}(y)}^{-1} \quad (2.2)$$

**Theorem 2.6** *Let  $X$  and  $Y$  have joint density  $f(x, y)$ . Then the density  $g(w)$  of  $W = X + Y$  is given by*

$$g(w) = \int_{-\infty}^{\infty} f(x, w-x) dx = \int_{-\infty}^{\infty} f(w-y, y) dy. \quad (2.3)$$

A result for  $W = X - Y$  follows immediately and, in the case of  $X$  and  $Y$  being independent, the corresponding expression can also be written down readily as a product of the two respective densities. Similar results hold for the product or ratio of random variables (see Rohatgi 1984, Section 8.4) but, since we have no call for them in this work, we do not write them down here. An immediate corollary that can give an angle on small sample behavior of statistics that are written as sums is;

**Corollary 2.2** *Let  $X_1, \dots, X_n$  be independent, not always identically distributed, continuous random variables with densities  $f_1(x)$  to  $f_n(s)$  respectively. Let  $S_n = \sum_{j=1}^n X_j$ . Then the density,  $g_n(s)$ , of  $S_n$  is given by*

$$g_n(s) = \int_{-\infty}^{\infty} g_{n-1}(s-x) f_n(x) dx.$$



This result can be used iteratively building up successive solutions by carrying out the integration. The integration itself will mostly be not particularly tractable and can be evaluated using numerical routines. Note the difference between making a large sample statistical approximation to the sum and that of a numerical approximation to the integral. The integral expression itself is an exact result.

### *Normal distribution*

A random variable  $X$  is taken to be a normal variate with parameters  $\mu$  and  $\sigma$  when we write  $X \sim \mathcal{N}(\mu, \sigma^2)$ . The parameters  $\mu$  and  $\sigma^2$  are the mean and variance respectively, so that  $\sigma^{-1}(X - \mu) \sim \mathcal{N}(0, 1)$ . The distribution  $\mathcal{N}(0, 1)$  is called the standard normal. The density of the standard normal variate, that is, having mean zero and variance one, is typically denoted  $\phi(x)$  and the cumulative distribution  $\Phi(x)$ . The density  $f(x)$ , for  $x \in (-\infty, \infty)$  is given by

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right].$$

For stochastic processes described below, Brownian motion relates to a Gaussian process, that is, it has been standardized, in an analogous way that the standard normal relates to any other normal distribution. For the normal distribution, all cumulants greater than 2 are equal to zero. Simple calculations (Johnson and Kotz, 1970) show that, for  $X \sim \mathcal{N}(0, 1)$ , then  $E(X^r) = (r - 1)(r - 3) \dots 3.1$ . Thus, all odd moments are equal to zero and all even moments are expressible in terms of the variance. The normal distribution is of very great interest in view of it frequently being the large sample limiting distribution for sums of random variables. These arise naturally via simple estimating equations. These topics are looked at in greater detail below.

The multivariate normal can be characterized in various ways. If and only if all marginal distributions and all conditional distributions are normal then we have multivariate normality. If and only if all linear combinations are univariate normal then we have multivariate normality. It is only necessary to be able to evaluate the standard normal integral,  $\Phi(x) = 1 - \int_x^\infty \phi(x)dx$ , since any other normal distribution,  $f(x)$ , can be put in this form via the linear transformation  $(X - \mu)/\sigma$ . Tables, calculator, and computer routines can approximate the numerical integral. Otherwise, it is worth bearing in mind the following;

**Lemma 2.1** *Upper and lower bounds for the normal integral can be obtained from*

$$\frac{x}{1+x^2} e^{-x^2/2} < \int_x^\infty e^{-u^2/2} du < \frac{1}{x} e^{-x^2/2}.$$

The lemma tells us that we expect  $1 - \Phi(x)$  to behave like  $\phi(x)/x$  as  $x$  increases. The ratio  $\phi(x)/x$  is known as Mill's ratio. Approximate calculations are then possible without the need to resort to sophisticated algorithms, although, in modern statistical analysis, it is now so commonplace to routinely use computers that the value of the lemma is rather limited. The normal distribution plays an important role in view of the central limit theorem described below but also note the interesting theorem of Cramer (1937) whereby, if a finite sum of independent random variables is normal, then each variable itself is normal. Cramer's theorem might be contrasted with central limit theorems whereby sums of random variables, under broad conditions, approach the normal as the sum becomes infinitely large. These limit results are looked at later. The normal distribution is important since it provides the basis to Brownian motion and this is the key tool that we will use for inference throughout this text.

#### *Uniform distribution and the probability integral transform*

For the standard uniform distribution in which  $u \in [0, 1]$ ,  $f(u) = 1$  and  $F(u) = u$ . Uniform distributions on the interval  $[a, b]$  correspond to the density  $f(u) = 1/(b - a)$  but much more important is the fact that for any continuous distribution,  $G(t)$ , we can say:

**Theorem 2.7** *For the random variable  $T$ , having distribution  $G(t)$ , letting  $U_1 = G(T)$  and  $U_2 = 1 - G(T)$ , then both  $U_1$  and  $U_2$  have a standard uniform distribution.*

This central result, underpinning a substantial body of work on simulation and re-sampling, is known as the probability integral transform. Whenever we can invert the function  $G$ , denoted  $G^{-1}$ , then, from a single uniform variate  $U$  we obtain the two variates  $G^{-1}(U)$  and  $G^{-1}(1 - U)$  which have the distribution  $G$ . The two variates are of course not independent but, in view of the strong linearity property of expectation (the expectation of a linear function of random variables is the same linear function of the expectations), we can often use this to our advantage to improve precision when simulating. Another inter-

esting consequence of the probability integral transform is that there exists a transformation of a variate  $T$ , with any given distribution, into a variate having any other chosen distribution. Specifically, we have:

**Corollary 2.3** *For any given continuously invertible distribution function  $H$ , and continuous distribution  $G(t)$ , the variate  $H^{-1}\{G(T)\}$  has distribution  $H$ .*

In particular, it is interesting to consider the transformation  $\Phi^{-1}\{G_n(T)\}$  where  $G_n$  is the empirical estimate (discussed below) of  $G$ . This transformation, which preserves the ordering, makes the observed distribution of observations as close to normal as possible. Note that since the ordering is preserved, use of the transformation makes subsequent procedures nonparametric in as much as the original distribution of  $T$  has no impact. For the problems of interest to us in survival analysis we can use this in one of two ways: firstly, to transform the response variable time in order to eliminate the impact of its distribution and, secondly, in the context of regression problems, to transform the distribution of regressors as a way to obtain greater robustness by reducing the impact of outliers.

#### *Exponential distribution and cumulative hazard transformation*

The standard exponential distribution is defined on the positive real line  $(0, \infty)$ . We have, for  $u \in (0, \infty)$ ,  $f(u) = \exp(-u)$  and  $F(u) = 1 - \exp(-u)$ . An exponential distribution with mean  $1/\alpha$  and variance  $1/\alpha^2$  has density  $f(u) = \alpha \exp(-\alpha u)$  and cumulative distribution  $F(u) = 1 - \exp(-\alpha u)$ . The density of a sum of  $m$  independent exponential variates having mean  $1/\alpha$ , is an Erlang density whereby  $f(u) = \alpha(\alpha u)^{m-1} \exp(-\alpha u) / \Gamma(m)$  and where  $\Gamma(m) = \int_0^\infty \exp(-u) u^{m-1} du$ . The gamma distribution has the same form as the Erlang although, for the gamma, the parameter  $m$  can be any real positive number and is not restricted to being an integer. An exponential variate  $U$  can be characterized as a power transformation on a Weibull variate in which  $F(t) = 1 - \exp[(-\alpha t)^k]$ . Finally, we have the important result:

**Theorem 2.8** *For any continuous positive random variable  $T$ , with distribution function  $F(t)$ , the variate  $U = \int_0^T f(u)/[1 - F(u)] du$  has a standard exponential distribution.*

This result is important in survival modeling and we return to it later. The function  $f(t)/[1 - F(t)]$  is known as the hazard function and

$\int_0^t f(u)/[1 - F(u)]du$  as the cumulative hazard function. The transformation is called the cumulative hazard transformation.

## 2.6 Expectation

It is worth saying a word or two more about expectation as a fundamental aspect of studies in probability. Indeed it is possible for the whole theory to be constructed with expectation as a starting point rather than the now classical axiomatic structure to probability. For a function of a random variable  $T$ ,  $\psi(T)$  say, as stated at the beginning of the previous section, we write,  $E(\psi(T))$  of this function via

$$E\psi(T) = \int \psi(t)f(t)dt = \int \psi(t)dF(t),$$

where the integrals, viewed as limiting processes, are all assumed to converge. The normal distribution function for a random variable  $X$  is completely specified by  $E(X)$  and  $E(X^2)$ . In more general situations we can assume a unique correspondence between the moments of  $X$ ,  $E(X^r)$ ,  $r = 1, 2, \dots$ , and the distribution functions as long as these moments all exist. While it is true that the distribution function determines the moments the converse is not always true. However, it is almost always true (Stuart and Ord 1994, page 111) and, for all the distributions of interest to us here, the assumption can be made without risk. It can then be helpful to view each moment, beginning with  $E(X)$ , as providing information about  $F(x)$ . This information typically diminishes quickly with increasing  $r$ . We can use this idea to improve inference for small samples when large sample approximations may not be sufficiently accurate. Moments can be obtained from the moment generating function,  $M(t) = E\{\exp(tX)\}$  since we have:

**Lemma 2.2** *If  $\int \exp(tx)f(x)dx < \infty$  then*

$$E(X^r) = \left\{ \frac{\partial^r M(t)}{\partial t^r} \right\}_{t=0}, \text{ for all } r.$$

In Section 2.8 we consider the variance function which is also an expectation and is of particular interest to one of our central goals here, that of constructing useful measures of the predictive strength of any model. At the root of the construction lie two important inequalities, the Chebyshev-Bienaymé inequality (described in Section 2.8 and Jensen's inequality described below. For this we first need:

**Definition 2.3** *The real-valued function  $w(x)$  is called “convex” on some interval  $I$  (an infinite set and not just a point) whenever, for  $x_1, x_2 \in I$  and for  $0 \leq \lambda \leq 1$ , we have*

$$w[\lambda x_1 + (1 - \lambda)x_2] \leq \lambda w(x_1) + (1 - \lambda)w(x_2).$$

It is usually sufficient to take convexity to mean that  $w'(x)$  and  $w''(x)$  are greater than or equal to zero at all interior points of  $I$  since this is a consequence of the definition. We have (Jensen’s inequality):

**Lemma 2.3** *If  $w$  is convex on  $I$  then, assuming expectations exist on this interval,  $w[E(X)] \leq E[w(X)]$ . If  $w$  is linear in  $X$  throughout  $I$ , that is,  $w''(x) = 0$  when twice differentiable, then equality holds.*

For the variance function we see that  $w(x) = x^2$  is a convex function and so the variance is always positive. The further away from the mean, on average, the observations are to be found, then the greater the variance. We return to this in Section 2.8. Although very useful, the moment-generating function,  $M(t) = E\{\exp(tX)\}$  has a theoretical weakness in that the integrals may not always converge. It is for this, mainly theoretical, reason that it is common to study instead the characteristic function, which has an almost identical definition, the only difference being the introduction of complex numbers into the setting. The characteristic function, denoted by  $\phi(t)$ , always exists and is defined as:

$$\phi(t) = M(it) = \int_{-\infty}^{\infty} \exp(itx) dF(x), \quad i^2 = -1.$$

Note that the contour integral in the complex plane is restricted to the whole real axis. Analogous to the above lemma concerning the moment-generating function we have

$$E(X^r) = (-i)^r \left\{ \frac{\partial^r \phi(t)}{\partial t^r} \right\}_{t=0}, \quad \text{for all } r.$$

This is important in that it allows us to anticipate the cumulative generating function which turns out to be of particular importance in obtaining improved approximations to those provided by assuming normality. We return to this below in Section 2.9. If we expand the exponential function then we can write;

$$\phi(t) = \int_{-\infty}^{\infty} \exp(itx) dF(x) = \exp \left\{ \sum_{r=1}^{\infty} \kappa_r(it)^r / r! \right\}$$

and, identifying  $\kappa_r$  as the coefficient of  $(it)^r/r!$  in the expansion of  $\log \phi(t)$ . The function  $\psi(t) = \log \phi(t)$  is called the cumulative generating function. When this function can be found then the density  $f(x)$  can be defined in terms of it. We have the important relation

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt, \quad \phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

It is possible to approximate the density  $f(x)$  by working with i.i.d. observations  $X_1, \dots, X_n$  and the empirical characteristic function  $\phi(t) = n^{-1} \sum_{i=1}^n \exp(itx_i)$  which can then be inverted. It is also possible to approximate the integral using a method of numerical analysis, the so-called method of steepest descent, to obtain a saddlepoint approximation (Daniels, 1954). We return to this approximation below in Section 2.9.

## 2.7 Order statistics and their expectations

The normal distribution and other parametric distributions described in the next chapter play a major role in survival modeling. However, robustness of any inferential technique to particular parametric assumptions is always a concern. Hopefully, inference is relatively insensitive to departures from parametric assumptions or is applicable to whole families of parametric assumptions. The most common way to ensure this latter property is via the theory of order statistics which we recall here. Consider the  $n$  independent identically distributed (i.i.d.) random variables:  $X_1, X_2, \dots, X_n$  and a single realization of these that we can order from the smallest to the largest:  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ . Since the  $X_i$  are random, so also are the  $X_{(i)}$ , and the interesting question concerns what we can say about the probability structure of the  $X_{(i)}$  on the basis of knowledge of the parent distribution of  $X_i$ . In fact, we can readily obtain many useful results which, although often cumbersome to write down, are in fact straightforward. Firstly we have:

**Theorem 2.9** Taking  $P(x) = \Pr(X \leq x)$  and  $F_r(x) = \Pr(X_{(r)} \leq x)$  then:

$$F_r(x) = \sum_{i=r}^n \binom{n}{i} P^i(x) [1 - P(x)]^{n-i}. \quad (2.4)$$

This important result has two immediate and well known corollaries dealing with the maximum and minimum of a sample of size  $n$ .

**Corollary 2.4**

$$F_n(x) = P^n(x), \quad F_1(x) = 1 - [1 - P(x)]^n \quad (2.5)$$

In practice, in order to evaluate  $F_r(x)$  for other than very small  $n$ , we exploit the equivalence between partial binomial sums and the incomplete beta function. Thus, if, for  $a > 0$ ,  $b > 0$ ,  $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$  and  $I_\pi(a, b) = \int_0^\pi t^{a-1}(1-t)^{b-1}dt/B(a, b)$ , then putting  $P(x) = \pi$ , we have that  $F_r(x) = I_\pi(r, n-r+1)$ . These functions are widely tabulated and also available via numerical algorithms to a high level of approximation. An alternative, although less satisfying, approximation would be to use the DeMoivre-Laplace normal approximation to the binomial sums. Differentiation of (2.4) provides the density which can be written as

$$f_r(x) = \frac{1}{B(r, n-r+1)} P^{r-1}(x) [1 - P(x)]^{n-r} p(x). \quad (2.6)$$

Since we have a relatively straightforward expression for the distribution function itself, then this expression for the density is not often needed. It can come in handy in cases where we need to condition and apply the law of total probability. Expressions for  $f_1(x)$  and  $f_n(x)$  are particularly simple and we have

**Corollary 2.5**

$$f_1(x) = n[1 - P(x)]^{n-1}p(x), \quad f_n(x) = nP^{n-1}(x)p(x). \quad (2.7)$$

More generally it is also straightforward to obtain

**Theorem 2.10** *For any subset of the  $n$  order statistics:  $X_{n_1}, X_{n_2}, \dots, X_{n_k}$ ,  $1 \leq n_1 \leq \dots \leq n_k$ , the joint distribution  $f(x_1, \dots, x_k)$  is expressed as*

$$f(x_1, \dots, x_k) = n! \left[ \prod_{j=1}^k p(x_j) \right] \prod_{j=0}^k \left\{ \frac{[P(x_{j+1}) - P(x_j)]^{n_{j+1} - n_j - 1}}{(n_{j+1} - n_j - 1)!} \right\} \quad (2.8)$$

in which  $p(x) = P'(x)$ . This rather involved expression leads to many useful results including the following corollaries:

**Corollary 2.6** *The joint distribution of  $X_{(r)}$  and  $X_{(s)}$  is*

$$F_{rs}(x, y) = \sum_{j=s}^n \sum_{i=r}^j \frac{n!}{i!(j-i)!(n-j)!} P^i(x) [P(y) - P(x)]^{j-i} [1 - P(y)]^{n-j}.$$

The joint distribution of  $X_{(r)}$  and  $X_{(s)}$  is useful in establishing a number of practical results such as the distribution of the range, the distribution of the interquartile range and an estimate for the median among others. Using the result (Section 2.5) for the distribution of a difference, a simple integration then leads to the following:

**Corollary 2.7** *Letting  $W_{rs} = X_{(s)} - X_{(r)}$  then: in the special case of a parent uniform distribution we have*

$$f(w_{rs}) = \frac{1}{B(s-r, n-s+r+1)} w_{rs}^{s-r-1} (1-w_{rs})^{n-s+r}. \quad (2.9)$$

Taking  $s = n$  and  $r = 1$ , recalling that  $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$  and that  $\Gamma(n) = n!$ , then we have the distribution of the range for the uniform.

**Corollary 2.8** *Letting  $w = U_{(n)} - U_{(1)}$  be the range for a random sample of size  $n$  from the standard uniform distribution, then the cumulative distribution is given by*

$$F_U(w) = nw^{n-1} - (n-1)w^n. \quad (2.10)$$

Straightforward differentiation gives  $f_U(w) = n(n-1)w^{n-2}(1-w)$ , a simple and useful result. For an arbitrary distribution,  $F(\cdot)$  we can either carry out the same kind of calculations from scratch or, making use once more of the probability integral transform (see Section 2.4), use the above result for the uniform and transform into arbitrary  $F$ . Even this is not that straightforward since, for some fixed interval  $(w_1, w_2)$ , corresponding to  $w = w_2 - w_1$  from the uniform, the corresponding  $F^{-1}(w_2) - F^{-1}(w_1)$  depends not only on  $w_2 - w_1$  but on  $w_1$  itself. Again we can appeal to the law of total probability, integrating over all values of  $w_1$  from 0 to  $1-w$ . In practice, it may be good enough to divide the interval  $(0, 1-w)$  into a number of equally spaced points, ten would suffice, and simply take the average. Interval estimates for any given quantile, defined by  $P(\xi_\alpha) = \alpha$ , follow from the basic result and we have:



**Corollary 2.9** *In the continuous case, for  $r < s$ , the pair  $(X_{(r)}, X_{(s)})$  covers  $\xi_\alpha$  with probability given by  $I_\pi(r, n - r + 1) - I_\pi(r, n - s + 1)$ .*

**Theorem 2.11** *For the special case in which  $n_1 = 1$ ,  $n_2 = 2$ , ...  $n_n = n$ , then*

$$f(x_1, \dots, x_n) = n! \prod_{j=1}^n p(x_j). \quad (2.11)$$

*A characterization of order statistics: Markov property*

The particularly simple results for the exponential distribution lead to a very useful and powerful characterization of order statistics. If  $Z_1, \dots, Z_n$  are i.i.d. exponential variates with parameter  $\lambda$ , then an application of Corollary 2.4 shows that the minimum of  $Z_1$  to  $Z_n$  has itself an exponential distribution with parameter  $n\lambda$ . We can define the random variable  $Y_1$  to be the gap time between 0 and the first observation,  $Z_{(1)}$ . The distribution of  $Y_1$  (equivalently  $Z_{(1)}$ ) is exponential with parameter  $n\lambda$ . Next, we can define  $Y_2$  to be the gap  $Z_{(2)} - Z_{(1)}$ . In view of the lack of memory property of the exponential distribution, once  $Z_{(1)}$  is observed, the conditional distribution of each of the remaining  $(n - 1)$  variables, given that they are all greater than the observed time  $Z_{(1)}$ , remains exponential with parameter  $\lambda$ . The variable  $Y_2$  is then the minimum of  $(n - 1)$  i.i.d. exponential variates with parameter  $\lambda$ . The distribution of  $Y_2$  is therefore, once again, exponential, this time with parameter  $(n - 1)\lambda$ . More generally we have the following lemma:

**Lemma 2.4** *If  $Z_{(1)}, \dots, Z_{(n)}$  are the order statistics from a sample of size  $n$  of standard exponential variates, then, defining  $Z_{(0)} = 0$ ,*

$$Y_i = Z_{(i)} - Z_{(i-1)}, \quad i = 1, \dots, n$$

*are  $n$  independent exponential variates in which  $E(Y_i) = 1/(n - i + 1)$ .*

This elementary result is very important in that it relates the order statistics directly to sums of simple independent random variables which are not themselves order statistics. Specifically we can write

$$Z_{(r)} = \sum_{i=1}^r \{Z_{(i)} - Z_{(i-1)}\} = \sum_{i=1}^r Y_i,$$

leading to the immediate further lemma:

**Lemma 2.5** *For a sample of size  $n$  from the standard exponential distribution and letting  $\alpha_i = 1/(n - i + 1)$ , we have:*

$$E[Z_{(r)}] = \sum_{i=1}^r E(Y_i) = \sum_{i=1}^r \alpha_i, \quad \text{Var}[Z_{(r)}] = \sum_{i=1}^r \text{Var}(Y_i) = \sum_{i=1}^r \alpha_i^2.$$

The general flavor of the above result applies more generally than just to the exponential and, applying the probability integral transform (Section 2.5), we have:

**Lemma 2.6** *For an i.i.d. sample of size  $n$  from an arbitrary distribution,  $G(x)$ , the  $r$ th largest order statistic,  $X_{(r)}$  can be written*

$$X_{(r)} = G^{-1}\{1 - \exp(-Y_1 - Y_2 - \cdots - Y_r)\},$$

where the  $Y_i$  are independent exponential variates in which  $E(Y_i) = 1/(n - i + 1)$ .

One immediate conclusion that we can make from the above expression is that the order statistics from an arbitrary distribution form a Markov chain. The conditional distribution of  $X_{(r+1)}$  given  $X_{(1)}, X_{(2)}, \dots, X_{(r)}$  depends only on the observed value of  $X_{(r)}$  and the distribution of  $Y_{r+1}$ . This conditional distribution is clearly the same as that for  $X_{(r+1)}$  given  $X_{(r)}$  alone, hence the Markov property. If needed we can obtain the joint density,  $f_{rs}$ , of  $X_{(r)}$  and  $X_{(s)}$ , ( $1 \leq r < s \leq n$ ) by a simple application of Theorem 2.10. We then write:

$$f_{rs}(x, y) = \frac{n! P^{r-1}(x)p(x)p(y)[P(y) - P(x)]^{s-r-1}[1 - P(y)]^{n-s}}{(r-1)!(s-r-1)!(n-s)!}.$$

From this we can immediately deduce the conditional distribution of  $X_{(s)}$  given that  $X_{(r)} = x$  as:

$$f_{s|r}(y|x) = \frac{(n-r)!}{(s-r-1)!(n-s)!} \frac{p(y)[P(y) - P(x)]^{s-r-1}[1 - P(y)]^{n-s}}{[1 - P(x)]^{n-r}}.$$

A simple visual inspection of this formula confirms again the Markov property. Given that  $X_{(r)} = x$  we can view the distribution of the remaining  $(n - r)$  order statistics as an ordered sample of size  $(n - r)$  from the conditional distribution  $P(u|u > x)$ .

*Expected values of order statistics*

Given the distribution of any given order statistic we can, at least in principle, calculate any moments, in particular the mean, by applying the basic definition. In practice, this may be involved and there may be no explicit analytic solution. Integrals can be evaluated numerically but, in the majority of applications, it can be good enough to work with accurate approximations. The results of the above subsection, together with some elementary approximation techniques are all that we need. Denoting the distribution of  $X$  as  $P(x)$ , then the probability integral transform (Section 2.5) provides that  $U = P(X)$  has a uniform distribution. The moments of the order statistics from a uniform distribution are particularly simple so that  $E\{U_{(r)}\} = p_r = r/(n+1)$ . Denoting the inverse transformation by  $Q = P^{-1}$ , then

$$X_{(r)} = P^{-1}\{U_{(r)}\} = Q\{U_{(r)}\}.$$

Next, we can use a Taylor series development of the function  $X_{(r)}$  about the  $p_r$  so that

$$X_{(r)} = Q(p_r) + \{U_{(r)} - p_r\}Q'(p_r) + \{U_{(r)} - p_r\}^2Q''(p_r)/2 + \dots$$

and, taking expectations, term by term, we have

$$\begin{aligned} E\{X_{(r)}\} &\approx Q(p_r) + \frac{p_r q_r}{2(n+2)}Q''(p_r) \\ &\quad + \frac{p_r q_r}{(n+2)^2} \left\{ \frac{1}{3}(q_r - p_r)Q'''(p_r) + \frac{1}{8}p_r q_r Q''''(p_r) \right\} \end{aligned}$$

and

$$\begin{aligned} \text{Var}\{X_{(r)}\} &= \frac{p_r q_r}{2(n+2)}[Q'(p_r)]^2 + \frac{p_r q_r}{(n+2)^2} \{2(q_r - p_r)Q'(p_r)Q''(p_r) \\ &\quad + p_r q_r (Q'(p_r)Q'''(p_r) + [Q''(p_r)]^2)\}. \end{aligned}$$

It is straightforward to establish some relationships between the moments of the order statistics and the moments from the parent distribution. Firstly note that

$$E \left\{ \sum_{r=1}^n X_{(r)}^k \right\}^m = E \left\{ \sum_{r=1}^n X_r^k \right\}^m,$$

so that, if  $\mu$  and  $\sigma^2$  are the mean and variance in the parent population, then  $\sum_{r=1}^n \mu_r = n\mu$  and  $\sum_{r=1}^n E\{X_{(r)}^2\} = nE(X^2) = n(\mu^2 + \sigma^2)$ .

*Normal parent distribution*

For the case of a normal parent the expected values can be evaluated precisely for small samples and the approximations themselves are relatively tractable for larger sample sizes. One approach to data analysis in which it may be desirable to have a marginal normal distribution in at least one of the variables under study is to replace the observations by the expectations of the order statistics. These are sometimes called normal scores, typically denoted by  $\xi_{rn} = E(X_{(r)})$  for a random sample of size  $n$  from a standard normal parent with distribution function  $\Phi(x)$  and density  $\phi(x)$ . For a random sample of size  $n$  from a normal distribution with mean  $\mu$  and variance  $\sigma^2$  we can reduce everything to the standard case since  $E(X_{(r)}) = \mu + \xi_{rn}\sigma$ . Note that, if  $n$  is odd, then, by symmetry, it is immediately clear that  $E(X_{(r)}) = 0$  for all  $r$  that are odd. We can see that  $E(X_{(r)}) = -E(X_{(n-r+1)})$ . For  $n$  as small as, say, 5 we can use integration by parts to evaluate  $\xi_{r5}$  for different values of  $r$ . For example,  $\xi_{55} = 5 \int 4\Phi^3(x)\phi^2(x)dx$  which then simplifies to:  $\xi_{55} = 5\pi^{-1/2}/4 + 15\pi^{-3/2}\sin^{-1}(1/3)/2 = 1.16296$ . Also,  $\xi_{45} = 5\pi^{-1/2}/2 - 15\pi^{-3/2}\sin^{-1}(1/3) = 0.49502$  and  $\xi_{35} = 0$ . Finally,  $\xi_{15} = -1.16296$  and  $\xi_{25} = -0.49502$ . For larger sample sizes in which the integration becomes too fastidious we can appeal to the above approximations using the fact that

$$Q'(p_r) = \frac{1}{\phi(Q)}, \quad Q''(p_r) = \frac{Q}{\phi^2(Q)}, \quad Q'''(p_r) = \frac{1 + 2Q^2}{\phi^3(Q)},$$

$$Q''''(p_r) = \frac{Q(7 + 6Q^2)}{\phi^4(Q)}.$$

The above results arise from straightforward differentiation. Analogous calculations can be used to obtain exact or approximate expressions for  $\text{Cov}\{X_{(r)}, X_{(s)}\}$ .

## 2.8 Entropy and variance

In view of the mathematical equivalence of the density, distribution function and the hazard, we can be satisfied knowing any one of these functions for a variable  $T$  of interest. In the majority of areas of application of statistics, theoretical physics, and, possibly, biophysics being potential exceptions, we cannot really know much about these functions. Our usual strategy will be to collect data that enables the estimation of one or more of the functions, with any additional plausible

assumptions about the nature of these functions making this task that much easier. Paucity of data, or a need to only know the most important features of a distribution, will often lead us to restricting our attention to some simple summary measures. The most common summary measures are those of location and variance. For a measure of location we usually take the mean  $\mu$  or the median  $\xi_{0.5}$ . They tell us something about where the most likely values of  $T$  occur. An idea of just how “likely” these “likely” values are, in other words how concentrated is the distribution around the location measure, is most often provided by the variance or the square root of this, the standard deviation. The variance  $\sigma^2$  is defined by

$$\sigma^2 = E\{T - E(T)\}^2 = \int (t - \mu)^2 f(t) dt = \int (t - \mu)^2 dF(t). \quad (2.12)$$

An important insight into just why  $\sigma^2$  provides a good measure of precision, in other terms predictability, is given by:

**Theorem 2.12** *For every positive constant  $a$*

$$\Pr \{|T - \mu| \geq a\sigma\} \leq 1/a^2. \quad (2.13)$$

This famous inequality, known as the Bienaymé-Chebyshev inequality, underlines the fact that the smaller  $\sigma^2$  the better we can predict. A lesser used, although equally useful, measure of concentration is the so-called entropy of the distribution. Apart from a negative sign, this is also called the information of the distribution which is defined by  $V(f, f)$  where

$$V(g, h) = E \log g(T) = \int \log g(t) h(t) dt. \quad (2.14)$$

The entropy is just  $-V(g, h)$ . Note that the integral operator  $E$  in  $E \log g(T)$  is with respect to the density  $h(t)$ , this added generality being needed in the regression context. For univariate study the information is simply  $V(f, f)$  and would be written  $V$  since the arguments are implicit. Our intuition is good for  $\sigma^2$ , since it is clear that the further away, on average, are the values of  $T$ , then the larger will be  $\sigma^2$ . The same is true, although less obvious, for  $V$ . As  $T$  becomes concentrated around its mode (value of  $t$ , taken to be unique, at which  $f(t)$  assumes its greatest value), then, since  $\int f(t) dt$ , the area under the curve, is fixed at one,  $f(t)$  itself becomes larger at and around the mode. In the limit, as all the information becomes concentrated at a single point  $t_0$ , then  $f(t_0)$ , as well as  $E \log f(T)$ , tends to positive

infinity. The more spread out are the values of  $T$  then the closer to zero will tend to be  $E \log f(T)$ . Intermediary values of  $E \log f(T)$  then can be taken to correspond to different degrees of dispersion. Consider also the following which is true for any number of random variables and which, for the purposes of illustration, we limit to  $X_1$ ,  $X_2$  and  $X_3$ . We have

$$E \log f(X_1, X_2, X_3) = E \log f(X_3|X_2, X_1) + E \log f(X_2|X_1) \\ + E \log f(X_1),$$

so that the total information can be decomposed into sequential orthogonal contributions, each adding to the total amount of information so far. Note also, since we can interchange the  $X_i$ , the order in which the total information is put together has no impact on the final result. This is of course a desirable property. The information measure, as an indicator of precision, is well known in communication theory (Shannon and Weaver 1949) and statistical ecology, but is not so well known in biostatistics. It is also worth considering the fact that the most commonly used estimating technique, maximum likelihood, is best viewed as an empirical version of information. This follows since the usual log-likelihood divided by the sample size (which can be taken as a fixed constant) provides a consistent estimate of the information. Both the variance and the information are of particular interest when we condition on some other variable  $Z$ , possibly a vector. This is the regression setting where we focus on the impact of explanatory variables on some response variable of interest. The information gain would consider the distance between the distribution  $f(t)$  and  $f(t|z)$ . In the above construction the function  $g(t)$  is first equated with  $f(t)$  and subsequently to  $f(t|z)$ , whereas  $h(t)$  remains fixed at  $f(t, z)$ . Note also, that in this case, the integral is over the space of  $T$  and  $Z$ . This enables the construction of a simple and powerful measure of predictability. The amount by which the variance, or information, changes following such conditioning provides a direct quantification of the predictive strength of  $Z$ . We look at this more closely in the following subsection.

### *Explained randomness and explained variation*

Any models we work with are simply tools to enable us to efficiently construct conditional distributions. Validity of our models is an important issue, upon which we dwell later, but, for now, let us suppose our models are good enough to accurately reproduce the conditional

distributions of  $T$  given  $Z$  where  $Z$  may be a vector. The improvement in our predictive ability, given  $Z$ , can be quantified in view of the above Bienaymé-Chebyshev inequality and the variance decomposition

$$\text{Var}(T) = \text{Var} E(T|Z) + E \text{Var}(T|Z). \quad (2.15)$$

The total variance,  $\text{Var}(T)$ , breaks down into two parts, one of which we can interpret as the signal,  $\text{Var} E(T|Z)$ , and one as the pure noise  $E \text{Var}(T|Z)$ . The percentage of  $\text{Var}(T)$  that is taken up by  $\text{Var} E(T|Z)$  is the amount of the total variance that can be explained by  $Z$ . This translates directly the predictive power of  $Z$  so that the percentage of explained variance, is then quite central to efforts at quantifying how well our models do. We define it as

$$\Omega^2 = \frac{\text{Var} E(T|Z)}{\text{Var}(T)} = \frac{\text{Var}(T) - E \text{Var}(T|Z)}{\text{Var}(T)}. \quad (2.16)$$

The quantity  $\Omega^2$  in its own right is not well developed in the literature and we devote Section 3.9 to studying its importance. Following Draper (1984), there have been a number of challenges to  $\Omega^2$  as a useful concept (Healy 1984: Kvalseth 1985: Scott and Wild 1991: Willett and Singer 1988). However Draper's paper of 1984 was flawed and its conclusions did not hold up (Draper 1985). As a result, this subsequent work, having taken Draper's 1984 paper as its starting point, inherits the same logical errors.

Explained randomness, as opposed to explained variation, arises from a less transparent construction. We can use a monotonic transform of the expected information (expectation taken with respect to the distribution of  $Z$ ) and, taking  $D(T) = \exp -2E V\{f(t), f(t|Z)\}$ :  $D(T|Z) = \exp -2E V\{f(t|Z), f(t|Z)\}$ , we define the explained randomness  $\rho^2$  to be

$$\rho^2 = \frac{D(T) - D(T|Z)}{D(T)}. \quad (2.17)$$

We interpret  $\rho^2$  as the proportion of explained randomness in  $T$  attributable to  $Z$ . We also have the following important lemma that could, in its own right, be taken as a reason for studying explained randomness, but which, in any event, underlines a useful relationship between explained variation and explained randomness:

**Lemma 2.7** *If the pair  $(T, Z)$  are bivariate normal then  $\Omega^2 = \rho^2$ .*

The lemma provides further motivation for being interested in  $\rho^2$ , in that, for the more familiar classic regression case of a bivariate normal,

we obtain the same results by considering explained randomness that we obtain by considering explained variation. For other distributions, where variance itself may not be the best measure of dispersion, the concept explained randomness, based on entropy, might be viewed as having more generality. Our own experience in practical data analysis suggests that, as far as hierarchical model building or the quantification of partial or multiple effects is concerned, there does not appear to be anything to really choose between the two measures. For operational purposes we can take the two measures to be essentially equivalent, the use of one rather than the other being more a question of taste rather than one based on any real advantages or disadvantages.

## 2.9 Approximations

*Approximations to means and variances for functions of  $T$  ( $\delta$ -method)*

Consider some differentiable monotonic function of  $X$ , say  $\psi(X)$ . Our particular concern often relates to parameter estimates in which case the random variable  $X$  would be some function of the  $n$  i.i.d. data values, say  $\theta_n$  as an estimator of the parameter  $\theta$ . In the cases of interest,  $\theta_n$  converges with probability one to  $\theta$  and so also does  $\psi(\theta_n)$  to  $\psi(\theta)$ . Although  $\theta_n$  may not be unbiased for  $\theta$ , for large samples, the sequence  $E(\theta_n)$  converges to  $E(\theta) = \theta$ . Similarly  $E[\psi(\theta_n)]$  converges to  $\psi(\theta)$ . The mean value theorem (Section 2.2) enables us to write

$$\psi(\theta_n) = \psi(\theta) + (\theta_n - \theta)\psi'(\theta) + \frac{(\theta_n - \theta)^2}{2}\psi''(\xi) \quad (2.18)$$

for  $\xi \in (\theta \pm \theta_n)$ . Rearranging this expression, ignoring the third term on the right hand side, and taking expectations we obtain

$$\text{Var}\{\psi(\theta_n)\} \approx E\{\psi(\theta_n) - \psi(\theta)\}^2 \approx \{\psi'(\theta)\}^2 \text{Var}(\theta_n) \approx \{\psi'(\theta_n)\}^2 \text{Var}(\theta_n)$$

as an approximation to the variance. The approximation, once obtained in any given setting, is best studied on a case-by-case basis. It is an exact result for linear functions. For these, the second derivative is equal to zero and, more generally, the smaller the absolute value of this second derivative, the better we might anticipate the approximation to be. For  $\theta_n$  close to  $\theta$  the squared term will be small in absolute value when compared with the linear term, an additional motivation



to neglecting the third term. For the mean, the second term of Equation (2.18) is zero when  $\theta_n$  is unbiased, otherwise close to zero and, this time, ignoring this second term, we obtain

$$E\{\psi(\theta_n)\} \approx \psi(\theta_n) + \frac{1}{2} \text{Var}(\theta_n) \psi''(\theta_n) \quad (2.19)$$

as an improvement over the rougher approximation based on the first term alone of the above expression. Extensions of these expressions to the case of a consistent estimator  $\psi(\theta_n) = \psi(\theta_{1n}, \dots, \theta_{pn})$  of  $\psi(\theta)$  proceeds in the very same way, only this time based on a multivariate version of Taylor's theorem. These are:

$$\begin{aligned} \text{Var}\{\psi(\theta_n)\} &\approx \sum_{j=1}^p \sum_{m \geq j}^p \frac{\partial \psi(\theta)}{\partial \theta_j} \frac{\partial \psi(\theta)}{\partial \theta_m} \text{Cov}(\theta_{jn}, \theta_{mn}), \\ E\{\psi(\theta_n)\} &\approx \psi(\theta_{1n}, \dots, \theta_{pn}) + \frac{1}{2} \sum_j \sum_m \frac{\partial^2 \psi(\theta_n)}{\partial \theta_j \partial \theta_m} \text{Cov}(\theta_{jn}, \theta_{mn}). \end{aligned}$$

When  $p = 1$  then the previous expressions are recovered as special cases. Again, the result is an exact one in the case where  $\psi(\cdot)$  is a linear combination of the components  $\theta_j$  and this helps guide us in situations where the purpose is that of confidence interval construction. If, for example, our interest is on  $\psi$  and some strictly monotonic transformation of this, say  $\psi^*$ , is either linear or close to linear in the  $\theta_j$ , then it may well pay, in terms of accuracy of interval coverage, to use the delta-method on  $\psi^*$ , obtaining the end points of the confidence interval for  $\psi^*$  and subsequently inverting these, knowing the relationship between  $\psi$  and  $\psi^*$ , in order to obtain the interval of interest for  $\psi$ . Since  $\psi$  and  $\psi^*$  are related by one-to-one transformations then the coverage properties of an interval for  $\psi^*$  will be identical to those of its image for  $\psi$ . Examples in this book include confidence intervals for the conditional survivorship function, given covariate information, based on a proportional hazards model as well as confidence intervals for indices of predictability and multiple coefficients of explained variation.

### *Cornish-Fisher approximations*

In the construction of confidence intervals, the  $\delta$ -method makes a normality approximation to the unknown distribution and then replaces the first two moments by local linearization. A different approach, while still working with a normal density  $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ ,

in a way somewhat analogous to the construction of a Taylor series, is to express the density of interest,  $f(x)$ , in terms of a linear combination of  $\phi(x)$  and derivatives of  $\phi(x)$ . Normal distributions with nonzero means and variances not equal to one are obtained by the usual simple linear transformation and, in practical work, the simplest approach is to standardize the random variable  $X$  so that the mean and variance corresponding to the density  $f(x)$  are zero and one, respectively.

The derivatives of  $\phi(x)$  are well known, arising in many fields of mathematical physics and numerical approximations. Since  $\phi(x)$  is simply a constant multiplying an exponential term it follows immediately that all derivatives of  $\phi(x)$  are of the form of a polynomial that multiplies  $\phi(x)$  itself. These polynomials (apart from an alternating sign coefficient  $(-1)^i$ ) are the Hermite polynomials,  $H_i(x)$ ,  $i = 0, 1, \dots$ , and we have

$$H_0 = 1, \quad H_1 = x, \quad H_2 = x^2 - 1, \quad H_3 = x^3 - 3x, \quad H_4 = x^4 - 6x^2 + 3,$$

with  $H_5$  and higher terms being calculated by simple differentiation. The polynomials are of importance in their own right, belonging to the class of orthogonal polynomials and useful in numerical integration. Indeed, we have that

$$\int_{-\infty}^{\infty} H_i^2(x) \phi(x) dx = i!, \quad i = 0, \dots : \int_{-\infty}^{\infty} H_i(x) H_j(x) \phi(x) dx = 0, \quad i \neq j.$$

This orthogonality property is exploited in order for us to obtain explicit expressions for the coefficients in our expansion. Returning to our original problem we wish to determine the coefficients  $c_i$  in the expansion

$$f(x) = \sum_{i=0}^{\infty} c_i H_i(x) \phi(x) \quad (2.20)$$

and, in order to achieve this we multiply both sides of equation (2.20) by  $H_j(x)$ , subsequently integrating to obtain the coefficients

$$c_j = \frac{1}{j!} \int_{-\infty}^{\infty} f(x) H_j(x) dx. \quad (2.21)$$

Note that the polynomial  $H_j(x)$  is of order  $j$  so that the right-hand side of equation (2.21) is a linear combination of the moments, (up to the  $j$ th), of the random variable  $X$  having associated density  $f(x)$ .

These can be calculated step-by-step. For many standard densities several of the lower-order moments have been worked out and are available. Thus, it is relatively straightforward to approximate some given density  $f(x)$  in terms of a linear combination of  $\phi(x)$ .

The expansion of Equation (2.20) can be used in theoretical investigations as a means to study the impact of ignoring higher-order terms when we make a normal approximation to the density of  $X$ . We will use the expansion in an attempt to obtain more accurate inference for proportional hazards models fitted using small samples. Here the large sample normal assumption may not be sufficiently accurate and the approximating equation is used to motivate potential improvements obtained by taking into account moments of higher order than just the first and second. When dealing with actual data, the performance of any such adjustments needs to be evaluated on a case-by-case basis. This is because theoretical moments will have to be replaced by observed moments and the statistical error involved in that can be of the same order, or greater, than the error involved in the initial normal approximation. If we know or are able to calculate the moments of the distribution, then the  $c_i$  are immediately obtained. When the mean is zero we can write down the first four terms as

$$c_0 = 1, \quad c_1 = 0, \quad c_2 = (\mu_2 - 1)/2, \quad c_3 = \mu_3/6, \quad c_4 = (\mu_4 - 6\mu_2 + 3)/24,$$

from which we can write down an expansion in terms of  $\phi(x)$  as

$$f(x) = \phi(x) \{ 1 + (\mu_2 - 1)H_2(x)/2 + \mu_3 H_3(x)/6 \\ + (\mu_4 - 6\mu_2 + 3)H_4(x)/24 + \cdots \}.$$

This series is known as the Gram-Charlier series, and stopping the development at the fourth term corresponds to making corrections for skewness and kurtosis. In our later development of the properties of estimators in the proportional hazards model we will see that making corrections for skewness can help make inference more accurate, whereas, at least in that particular application, corrections for kurtosis appear to have little impact (Chapter 11).

#### *Saddlepoint approximations*

A different, although quite closely related, approach to the above uses saddlepoint approximations. Theoretical and practical work on these approximations indicate them to be surprisingly accurate for the tails

of a distribution. We work with the inversion formula for the cumulant generating function, a function that is defined in the complex plane, and in this two-dimensional plane, around the point of interest (which is typically a mean or a parameter estimate) the function looks like a minimum in one direction and a maximum in an orthogonal direction: hence the name “saddlepoint.” Referring back to Section 2.6 recall that we identified  $\kappa_r$  as the coefficient of  $(it)^r/r!$  in the expansion of the cumulant generating function  $K(t) = \log \phi(t)$  where  $\phi(t)$  is the characteristic function. We can exploit the relationship between  $\phi(t)$  and  $f(x)$ ; that is,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt, \quad \phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

to approximate  $f(x)$  by approximating the integral. The numerical technique that enables this approximation to be carried out is called the method of steepest descent and is described in Daniels (1954). The approximation to  $f(x)$  is simply denoted as  $f_s(x)$  and, carrying through the calculations, we find that

$$f_s(x) = \left\{ \frac{n}{2\pi K''(\lambda_x)} \right\}^{1/2} \exp[n\{K(\lambda_x) - x\lambda_x\}] \quad (2.22)$$

in which the solution to the differential equation in  $\lambda$ ,  $K'(\lambda) = x$  is given by  $\lambda_x$ . Our notation here of  $x$  as a realization of some random variable  $X$  is not specifically referring to our usual use of  $X$  as the minimum of survival time  $T$  and the censoring time  $C$ . It is simply the variable of interest and that variable, in our context, will be the score statistic (Chapter 11). For now, we assume the score to be composed of  $n$  contributions so that we view  $x$  as a mean based on  $n$  observations. Since, mostly, we are interested in the tails of the distribution, it can often help to approximate the cumulative distribution directly rather than make a subsequent appeal to numerical integration. Denoting the saddlepoint approximation to the cumulative distribution by  $F_s(x)$ , we write

$$F_s(x) = \Phi(u_x) + \phi(u_x)(u_x^{-1} + v_x^{-1}) \quad (2.23)$$

where  $\phi(x)$  indicates the standard normal density,  $\Phi(x) = \int_{-\infty}^x \phi(u) du$ , the cumulative normal,  $u_x = [2n\{x\lambda_x - K(\lambda_x)\}]^{1/2} \text{sgn}(\lambda_x)$ , and  $v_x = \lambda_x \{nK''(\lambda_x)\}^{1/2}$ . Since we are only concerned with tail probabilities

we need not pay attention to what occurs around the mean. If we do wish to consider  $F_s(x)$ , evaluated at the mean, the approximation is slightly modified and the reader is referred to Daniels (1987).

## 2.10 Stochastic processes

We define a stochastic process to be a collection of random variables indexed by  $t \in T$ . We write these as  $X(t)$  and take  $t$  to be fixed. If the set  $T$  has only a finite or a countably infinite number of elements then  $X(t)$  is referred to as a discrete-time process. We will be most interested in continuous-time processes. In applications we can standardize by the greatest value of  $t$  in the set  $T$  that can be observed, and so we usually take  $\sup\{t : t \in T\} = 1$ . We also take  $\inf\{t : t \in T\} = 0$ . We will be especially interested in observations on any given process between 0 and  $t$ . We call this the sample path.

### *Independent increments and stationarity*

Consider some partition of  $(0,1)$  in which  $0 = t_0 < t_1 < t_2 < \dots < t_n = 1$ . If the set of random variables  $X(t_i) - X(t_{i-1})$   $i = 1, \dots, n$  are independent then the stochastic process  $X(t)$  is said to have independent increments. Another important property is that of stationarity. We say that a stochastic process  $X(t)$  has stationary increments if  $X(s+t) - X(s)$  has the same distribution for all values of  $s$ . Stationarity indicates, in as much as probabilistic properties are concerned, that when we look forward, from the point  $s$ , a distance  $t$ , the only relevant quantity is how far forward  $t$  we look. Our starting point itself is irrelevant. As we progress through time, everything that we have learned is summarized by the current position. It can also be of value to consider a process with a slighter weaker property, the so-called second-order stationarity. Rather than insist on a requirement for the whole distribution we limit our attention to the first two moments and the covariance between  $X(s+t)$  and  $X(s)$  which depends only upon  $|t|$ . Our main focus is on Gaussian processes which, when they have the property of second-order stationarity, will in consequence be stationary processes. Also, simple transformations can produce stationary processes from nonstationary ones, an example being the transformation of the Brownian bridge into an Ornstein-Uhlenbeck process.

*Gaussian processes*

If for every partition of  $(0,1)$ ,  $0 = t_0 < t_1 < t_2 < \cdots < t_n = 1$ , the set of random variables  $X(t_1), \dots, X(t_n)$  has a multivariate normal distribution, then the process  $X(t)$  is called a Gaussian process. Brownian motion, described below, can be thought of as simply a standardized Gaussian process. A Gaussian process being uniquely determined by the multivariate means and covariances it follows that such a process will have the property of stationarity if for any pair  $(s, t : t > s)$ ,  $\text{Cov}\{X(s), X(t)\}$  depends only on  $(t - s)$ . In practical studies we will often deal with sums indexed by  $t$  and the usual central limit theorem will often underlie the construction of Gaussian processes.

**2.11 Brownian motion**

Consider a stochastic process  $X(t)$  on  $(0,1)$  with the following three properties:

1.  $X(0) = 0$ , i.e., at time  $t = 0$  the starting value of  $X$  is fixed at 0.
2.  $X(t), t \in (0,1)$  has independent stationary increments.
3. At each  $t \in (0,1)$  the distribution of  $X(t)$  is  $\mathcal{N}(0, t)$ .

This simple set of conditions completely describes a uniquely determined stochastic process called Brownian motion. It is also called the Wiener process or Wiener measure. It has many important properties and is of fundamental interest as a limiting process for a large class of sums of random variables on the interval  $(0,1)$ . An important property is described in Theorem 2.13 below. Firstly we make an attempt to describe just what a single realization of such a process might look like. Later we will recognize the same process as being the limit of a sum of independent random contributions. The process is continuous and so, approximating it by any drawing, there cannot be any gaps. At the same time, in a sense that can be made more mathematically precise, the process is infinitely jumpy. Nowhere does a derivative exist. Figure 2.1 illustrates this via a simulated approximation. The right-hand figure is obtained from the left-hand one by homing in on the small interval  $(0.20, 0.21)$ , subtracting off the value observed at  $t = 0.20$ , and rescaling to the interval  $(0,1)$ . The point we are trying to make is that the resulting process itself looks like (and indeed is) a realization of Brownian motion. Theoretically, this

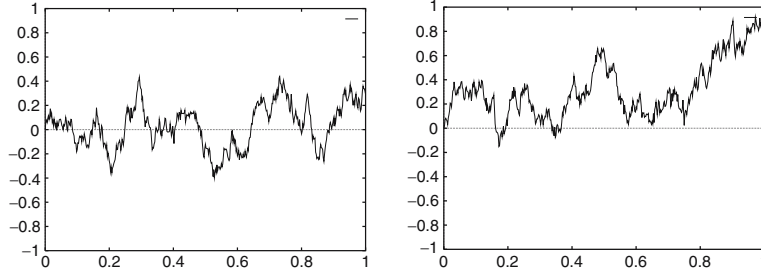


Figure 2.1: Two simulated independent realizations of a Brownian motion process.

could be repeated without limit which allows us to understand in some way how infinitely jumpy is the process. In practical examples we can only ever approximate the process by linearly connecting up adjacent simulated points.

**Theorem 2.13** *Conditioning on a given path we have*

$$\begin{aligned} \Pr \{X(t+s) > x | X(s) = x_s, X(u), 0 \leq u < s\} \\ = \Pr \{X(t+s) > x | X(s) = x_s.\} \end{aligned}$$

So, when looking ahead from time point  $s$  to time point  $t+s$ , the previous history indicating how we arrived at  $s$  is not relevant. The only thing that matters is the point at which we find ourselves at time point  $s$ . This is referred to as the Markov property. The joint density of  $X(t_1), \dots, X(t_n)$  can be written as

$$f(x_1, x_2, \dots, x_n) = f_{t_1}(x_1) f_{t_1-t_2}(x_2 - x_1) \cdots f_{t_n-t_{n-1}}(x_n - x_{n-1})$$

This follows from the independent stationary increment condition. A consequence of the above result is that we can readily evaluate the conditional distribution of  $X(s)$  given some future value  $X(t)$  ( $t > s$ ). Applying the definition for conditional probability we have the following.

**Corollary 2.10** *The conditional distribution of  $X(s)$  given  $X(t)$  ( $t > s$ ) is normal with a mean and a variance given by,*

$$E\{X(s)|X(t) = w\} = ws/t, \quad \text{Var}\{X(s)|X(t) = w\} = s(t-s)/t.$$

This result helps provide insight into another useful process, the Brownian bridge described below. Other important processes arise as simple transformations of Brownian motion. The most obvious to consider is where we have a Gaussian process satisfying conditions (1) and (2) for Brownian motion but where, instead of the variance increasing linearly, i.e.,  $\text{Var } X(t) = t$ , the variance increases either too quickly or too slowly so that  $\text{Var } X(t) = \phi(t)$  where  $\phi(\cdot)$  is some monotonic increasing function of  $t$ . Then we can transform the time axis using  $\phi(\cdot)$  to produce a process satisfying all three conditions for Brownian motion. Consider also the transformation

$$V(t) = \exp(-\alpha t/2)X\{\exp(\alpha t)\}$$

where  $X(t)$  is Brownian motion. This is the Ornstein-Uhlenbeck process. It is readily seen that:

**Corollary 2.11** *The process  $V(t)$  is a Gaussian process in which  $E\{V(t)\} = 0$  and  $\text{Cov}\{V(t), V(s)\} = \exp\{-\alpha(t-s)/2\}$ .*

#### *Time-transformed Brownian motion*

Consider a process,  $X^\psi(t)$ , defined via the following three conditions, for some continuous  $\psi$  such that,  $\psi(t') > \psi(t)$  ( $t' > t$ ); (1)  $X^\psi(0) = 0$  (2)  $X^\psi(t), t \in (0, 1)$  has independent stationary increments; (3) at each  $t \in (0, 1)$  the distribution of  $X^\psi(t)$  is  $\mathcal{N}\{0, \psi(t)\}$ . The usual Brownian motion described above is exactly this process when  $\psi(t) = t$ . However, in view of the continuity and monotonicity of  $\psi$ , there exists an inverse function  $\psi^{-1}$  such that  $\psi^{-1}\{\psi(t)\} = t$ . Clearly, we can transform the process  $X^\psi(t)$  by multiplying, at each  $t$ , by  $\sqrt{t/\psi(t)}$ , and, defining  $\sqrt{0/\psi(0)} = 0$ . The resulting process we can call  $X(t)$  and it is readily seen that this process is standard Brownian motion. Thus, the only crucial assumption in Brownian motion is that of independent increments. Once we can assert this to be the case, it is only a question of scale and location to obtain standard Brownian motion.

#### *Brownian bridge*

Let  $W(t)$  be Brownian motion. We know that  $W(0) = 0$ . We also know that with probability one the process  $W(t)$  will return at some point to the origin. Let's choose a point, and in particular the point  $t = 1$  and consider the conditional process  $W^0(t)$ , defined to be Brownian motion



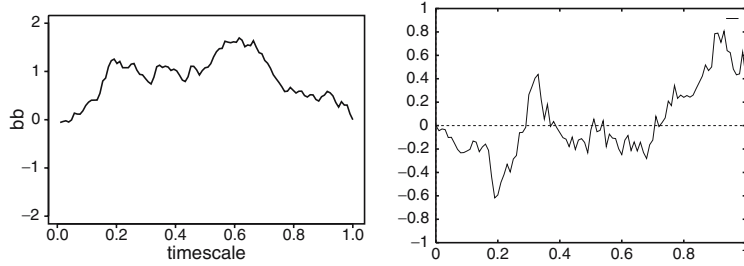


Figure 2.2: Two transformations of simulated Brownian motion by conditioning on  $W(1)$ . The first has  $W(1) = 0$  (Brownian bridge); the second has  $W(1) = 0.5$ .

conditioned by the fact that  $W(1) = 0$ . For small  $t$  this process will look very much like the Brownian motion from which it is derived. As  $t$  goes to one the process is pulled back to the origin since at  $t = 1$  we have that  $W^0(1) = 0$  and  $W(t)$  is continuous. Also  $W^0(0) = W(0) = 0$ . Such a process is called tied down Brownian motion or the Brownian bridge. Figure 2.2 illustrates a realization of a Brownian bridge and a realization of a Brownian motion constrained to assume a value other than zero at  $t = 1$ . We will see below that realizations of a Brownian bridge can be viewed as linearly transformed realizations of Brownian motion itself, and vice versa. From the results of above the section we can investigate the properties of  $W^0(t)$ . The process is a Gaussian process so we only need consider the mean and covariance function for the process to be completely determined. We have

$$E\{W(s)|W(1) = 0\} = 0 \quad \text{for } s < t.$$

This comes immediately from the above result. Next we have:

**Theorem 2.14**

$$\text{Cov}(W(s), W(t)|W(1) = 0) = s(1 - t). \quad (2.24)$$

This provides a simple definition of the Brownian bridge as being a Gaussian process having mean zero and covariance function  $s(1 - t)$ ,  $s < t$ . An alternative way of constructing the Brownian bridge is to consider the process defined as

$$W^0(t) = W(t) - tW(1), \quad 0 \leq t \leq 1.$$

Clearly  $W^0(t)$  is a Gaussian process. We see that

$$E\{W(0)\} = W(0) = E\{W(1)\} = W(1) = E\{W(t)\} = 0$$

so that the only remaining question is the covariance function for the process to be completely and uniquely determined. The following corollary is all we need.

**Corollary 2.12** *The covariance function for the process defined as  $W^0(t)$  is,*

$$\text{Cov}\{W^0(s), W^0(t)\} = s(1-t) \quad s < t.$$

This is the covariance function for the Brownian bridge developed above and, by uniqueness, the process is therefore itself the Brownian bridge. Such a covariance function is characteristic of many observed phenomena. The covariance decreases linearly with distance from  $s$ . As for Brownian motion, should the covariance function decrease monotonically rather than linearly, then a suitable transformation of the time scale enables us to write the covariance in this form. At  $t = s$  we recover the usual binomial expression  $s(1-s)$ .

Notice that not only can we go from Brownian motion to a Brownian bridge via the simple transformation

$$W^0(t) = W(t) - tW(1), \quad 0 \leq t \leq 1,$$

but the converse is also true, i.e., we can recover Brownian motion,  $X(t)$ , from the Brownian bridge,  $Z(t)$ , via the transformation

$$X(t) = (t+1)Z\left(\frac{t}{t+1}\right). \quad (2.25)$$

To see this, first note that, assuming  $Z(t)$  to be a Brownian bridge, then  $X(t)$  is a Gaussian process. It will be completely determined by its covariance process  $\text{Cov}\{X(s), X(t)\}$ . All we then require is the following lemma:

**Lemma 2.8** *For the process defined in (2.25),  $\text{Cov}\{X(s), X(t)\} = s$ .*

The three processes: Brownian motion, the Brownian bridge, and the Ornstein-Uhlenbeck are then closely related and are those used in the majority of applications. Two further related processes are also of use in our particular applications: integrated Brownian motion and reflected Brownian motion.

*Integrated Brownian motion*

The process  $Z(t)$  defined by  $Z(t) = \int_0^t W(u)du$ , where  $W(t)$  is Brownian motion is called integrated Brownian motion. Note that  $dZ(t)/dt = W(t)$  so that, for example, in the context of a model of interest, should we be able to construct a process converging in distribution to a process equivalent to Brownian motion, then the integrated process will converge in distribution to a process equivalent to integrated Brownian motion. We can see (by interchanging limits) that  $Z(t)$  can be viewed as the limit of a sum of Gaussian processes and is therefore Gaussian. Its nature is completely determined by its mean and covariance function. We have that

$$E\{Z(t)\} = E\left\{\int_0^t W(u)du\right\} = \int_0^t E\{W(u)\}du = 0. \quad (2.26)$$

For  $s < t$  we have:

**Lemma 2.9** *The covariance function for  $Z(s)$  and  $Z(t)$  is*

$$\text{Cov}\{Z(s), Z(t)\} = s^2(t/2 - s/6). \quad (2.27)$$

**Lemma 2.10** *The covariance function for  $Z(t)$  and  $W(t)$  is*

$$\text{Cov}\{Z(t), W(t)\} = t^2/2. \quad (2.28)$$

For a model in which inference derives from cumulative sums, this would provide a way of examining how reasonable are the underlying assumptions if repetitions are available. Repetitions can be obtained by bootstrap resampling if only a single observed process is available. Having standardized, a plot of the log-covariance function between the process and the integrated process against log-time ought be linear with slope of two and intercept of minus log 2 assuming that model assumptions hold.

*Reflected Brownian motion*

Suppose we choose some positive value  $r$  and then define the process  $W_r(t)$  as a function of Brownian motion,  $W(t)$ , in the following way: If  $W(t) < r$  then  $W_r(t) = W(t)$ . If  $W(t) \geq r$  then  $W_r(t) = 2r - W(t)$ . We have:

**Lemma 2.11**  *$W_r(t)$  is a Gaussian process,  $EW_r(t) = 0$ ,  $\text{Cov}\{W_r(s), W_r(t)\} = s$  when  $s < t$ .*

Thus,  $W_r(t)$  is also Brownian motion. Choosing  $r$  to be negative and defining  $W_r(t)$  so that, when  $W(t) > r$  then  $W_r(t) = W(t)$ . If  $W(t) \leq r$  then  $W_r(t) = 2r - W(t)$ . accordingly we have the same result. The process  $W_r(t)$  coincides exactly with  $W(t)$  until such a time as a barrier is reached. We can imagine this barrier as a mirror, and beyond the barrier the process  $W_r(t)$  is a simple reflection of  $W(t)$ . The interesting thing is that the resulting process is itself Brownian motion. One way of conceptualizing the idea is to imagine a large number of realizations of a completed Brownian motion process sampled independently. Imagine then these same realizations with a reflection applied. Then, whatever the point of reflection, if we consider the two collected sets of realizations, our overall impression of the behavior of the two processes will be the same. The value of this construction is to be seen in situations where, at some point in time, corresponding to some expected point of reflection under a hypothesis of drift, the drift changes direction. Under the hypothesis of Brownian motion, both Brownian motion, and Brownian motion reflected at some point, will look alike and will obey the same probability laws. Under an alternative hypothesis of drift however (see below), the behaviors will look quite different. This observation enables a simple construction with which to address the problem of crossing hazards.

### *Maximum of a Brownian motion*

A useful further result can be immediately obtained from the preceding one dealing with reflected Brownian motion. Suppose that  $W(t)$  is a Brownian motion. We might wish to consider the process  $M(t) = \sup_{u \in (0,t)} W(u)$ , which is the greatest value obtained by the process  $W(u)$  in the interval  $(0, t)$ . The greatest absolute distance is also of interest but, by symmetry arguments, this can be obtained immediately from the distribution of  $M(t)$ . Another related question, useful in interim analyzes, is the distribution of  $W(t)$  given the maximum  $M(t)$  obtained up until that time point. We have the following:

**Lemma 2.12** *If  $W(t)$  is standard Brownian motion and  $M(t)$  the maximum value attained on the interval  $(0, t)$ , i.e.,  $M(t) = \sup_{u \in (0,t)} W(u)$ , then*

$$\Pr \{M(t) > a\} = 2 \Pr \{W(t) > a\}.$$

This is a simple and elegant result and enables us to make simultaneous inference very readily. Sometimes, when using a Brownian motion

approximation for a process, we may want to, for example, describe an approximate confidence interval for the whole process rather than just a confidence interval at a single point  $t$ . In such a case the above result comes into play immediately. The joint distribution is equally simple and we make use of the following.

**Lemma 2.13** *If  $W(t)$  is standard Brownian motion and  $M(t)$  the maximum value attained on the interval  $(0, t)$ , i.e.,  $M(t) = \sup_{u \in (0, t)} W(u)$ , then*

$$\Pr \{W(t) < a - b, M(t) > a\} = \Pr \{W(t) > a + b\}.$$

The conditional distribution  $\Pr \{W(t) < a - b | M(t) > a\}$  can then be derived immediately by using the results of the two lemmas.

#### *Brownian motion with drift*

We will see that simple Brownian motion provides a good model for describing score statistics, or estimating equations, once standardized. This is because we can visualize these sums as approximating a limiting process arising from summing increments, for which the expected value is equal to zero. The setting in which we study such sums is typically that of evaluating some null hypothesis, often one of some given effect,  $H_0 : \beta = \beta_0$ , but sometimes a less obvious one, in the goodness-of-fit context, for example, whereby we can have,  $H_0 : \beta(t) = \hat{\beta}$ . Almost invariably, when we consider a null hypothesis, we have an alternative in mind, frequently a local or first alternative to the null. For a null hypothesis of Brownian motion, a natural and immediate alternative is that of Brownian motion with drift. Consider then the stochastic process  $X(t)$  defined by

$$X(t) = W(t) + \mu t$$

where  $W(t)$  is Brownian motion. We can immediately see that  $E\{X(t)\} = \mu t$  and  $\text{Var}\{X(t)\} = t$ . As for Brownian motion  $\text{Cov}\{X(s), X(t)\} = s, s < t$ . Alternatively we can define the process in a way analogous to our definition for Brownian motion as a process having the following three properties:

1.  $X(0) = 0$ .
2.  $X(t), t \in (0, 1)$  has independent stationary increments.
3. At each  $t \in (0, 1)$ ,  $X(t)$  is  $\mathcal{N}(\mu t, t)$ .

Clearly, if  $X(t)$  is Brownian motion with drift parameter  $\mu$ , then the process  $X(t) - \mu t$  is standard Brownian motion. Also, for the more common situation in which the mean may change non-linearly with time, provided the increments are independent, we can always construct a standard Brownian motion by first subtracting the mean at time  $t$ , then transforming the timescale in order to achieve a linearly increasing variance.

### *Probability results for Brownian motion*

There are a number of well-established and useful results for Brownian motion and related processes. The arcsine law can be helpful in comparing processes. Defining  $X^+(t)$  to be the time elapsed from the origin that the Brownian process remains positive, i.e.,  $\sup\{t : X(s) > 0 : 0 < s < t\}$  then  $\Pr(X^+ < x) = (2/\pi) \sin^{-1} \sqrt{x}$ . This law can be helpful in comparing processes and also in examining underlying hypotheses. For the Brownian bridge the largest distance from the origin in absolute value has a known distribution given in a theorem of Kolmogorov:

$$\Pr \left\{ \sup_t |W_0(t)| \leq \alpha \right\} \rightarrow 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 \alpha^2), \quad \alpha \geq 0. \quad (2.29)$$

The sum can be seen to be convergent since this is an alternating sign series in which the  $k$ th term goes to zero. Furthermore, the error in ignoring all terms higher than the  $n$ th is less, in absolute value, than the size of the  $(n+1)$ th term. Given that the variance of  $W_0(t)$  depends on  $t$  it is also of interest to study the standardized distribution  $B_0(t) = W_0(t)/\sqrt{t(1-t)}$ . This is, in fact, the Ornstein-Uhlenbeck process. Simple results for the supremum of this are not possible since the process becomes unbounded at  $t = 0$  and  $t = 1$ . Nonetheless, if we are prepared to reduce the interval from  $(0, 1)$  to  $(\varepsilon_1, \varepsilon_2)$  where  $\varepsilon_1 > 0$  and  $\varepsilon_2 < 1$  then we have an approximation due to Miller and Siegmund (1982):

$$\Pr \left\{ \sup_t |B_0(t)| \geq \alpha \right\} \approx \frac{4\phi(\alpha)}{\alpha} + \phi(\alpha) \left( \alpha - \frac{1}{\alpha} \right) \log \left\{ \frac{\varepsilon_2(1-\varepsilon_1)}{\varepsilon_1(1-\varepsilon_2)} \right\}, \quad (2.30)$$

where  $\phi(x)$  denotes the standard normal density. This enables us to construct confidence intervals for a bridged process with limits themselves going to zero at the endpoints. To obtain these we use the fact

that  $\Pr\{W_0(t) > \alpha\} = \Pr\{\sqrt{t(1-t)}B_0(t) > \alpha\}$ . For most practical purposes though it is good enough to work with Equation 2.29 and approximate the infinite sum by curtailing summation for values of  $k$  greater than 2.

## 2.12 Counting processes and martingales

Although not automatically our first choice for inference, the use of counting processes and martingales for inference in survival problems currently dominates this subject area. We will look at inference based on counting processes and martingales in Section 3.6, for some general results, and in Chapter 10 for the specific application to the proportional hazards model. In this chapter we aim to provide some understanding to the probability structure upon which the theory is based.

### *Martingales and stochastic integrals*

Recalling the discussion of Section 2.3 and that, for a bounded function  $H(x)$  and the empirical distribution function  $F_n(x)$ , we have, by virtue of the Helly-Bray theorem, that  $\int H(x)dF_n(x)$  converges in distribution to  $\int H(x)dF(x)$ . If we define  $M(x) = F_n(x) - F(x)$  and change the order of integration, i.e., move the expectation operator,  $E$ , outside the integral, then

$$E \left\{ \int H(x)dM(x) \right\} = 0.$$

This expression is worth dwelling upon. We think of  $E$  as being an integral operator or as defining some property of a random variable, specifically a measure of location. The random variable of relevance is not immediately apparent but can be seen to be  $F_n(x)$ , an  $n$ -dimensional function from the observations to the interval  $[0, 1]$ . We can suppose, at least initially, the functions  $F(x)$  and  $H(x)$  to be fixed and known. Our conceptual model allows the possibility of being able to obtain repetitions of the experiment, each time taking  $n$  independent observations. Thus, for some fixed given  $x$ , the value of  $F_n(x)$  will generally vary from one experiment to the next. We view  $x$  as an argument to a function, and  $F_n(x)$  as being random having a distribution studied below in Section 3.3. Recalling Section 2.3 on integration, note that we can rewrite the above equation as:

$$E \lim_{\max \Delta_i \rightarrow 0} \sum \{M(x_i) - M(x_{i-1})\}H(x_{i-1}) = 0, \quad (2.31)$$

where  $\Delta_i = x_i - x_{i-1} > 0$  and where, as described in Section 2.3 the summation is understood to be over an increasing partition in which  $\Delta_i > 0$  and  $\max \Delta_i$  goes to zero. Now, changing the order of taking limits, the above expression becomes

$$\lim_{\max \Delta_i \rightarrow 0} \sum E\{[M(x_i) - M(x_{i-1})]H(x_{i-1})\} = 0, \quad (2.32)$$

a result which looks simple enough but that has a lot of force when each of the infinite number of expectations can be readily evaluated. Let's view Equation 2.32 in a different light, one that highlights the sequential and ordered nature of the partition. Rather than focus on the collection of  $M(x_i)$  and  $H(x_i)$ , we can focus our attention on the increments  $M(x_i) - M(x_{i-1})$  themselves, the increments being multiplied by  $H(x_{i-1})$ , and, rather than work with the overall expectation implied by the operator  $E$ , we will set up a sequence of conditional expectations. Also, for greater clarity, we will omit the term  $\lim_{\max \Delta_i \rightarrow 0}$  altogether. We will put it back when it suits us. This lightens the notation and helps to make certain ideas more transparent. Later, we will equate the effect of adding back in the term  $\lim_{\max \Delta_i \rightarrow 0}$  to that of replacing finite differences by infinitesimal differences. Consider then

$$U = \sum \{M(x_i) - M(x_{i-1})\}H(x_{i-1}) \quad (2.33)$$

and, unlike the preceding two equations, we are able to greatly relax the requirement that  $H(x)$  be a known function or that  $M(x)$  be restricted to being the difference between the empirical distribution function and the distribution function. By sequential conditioning upon  $\mathcal{F}(x_i)$  where  $\mathcal{F}(x_i)$  are increasing sequence of sets denoting observations on  $M(x)$  and  $H(x)$ , for all values of  $x$  less than or equal to  $x_i$ , we can derive results of wide applicability. In particular, we can now take  $M(x)$  and  $H(x)$  to be stochastic processes. Some restrictions are still needed for  $M(x)$ , in particular that the incremental means and variances exist. We will suppose that

$$E\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\} = 0, \quad (2.34)$$

in words, when given  $\mathcal{F}(x_{i-1})$ , the quantity  $M(x_{i-1})$  is fixed and known and the expected size of the increment is zero. This is not a strong requirement and only supposes the existence of the mean since, should the expected size of the increment be other than zero, then we can subtract this difference to recover the desired property. Furthermore,



given  $\mathcal{F}(x)$ , the quantity  $H(x)$  is fixed. The trick is then to exploit the device of double expectation whereby for events,  $\mathcal{A}$  and  $\mathcal{B}$ , it is always true that  $E(\mathcal{A}) = EE(\mathcal{A}|\mathcal{B})$ . In the context of this expression,  $\mathcal{B} = \mathcal{F}(x_{i-1})$ , leading to

$$E(U) = \sum H(x_{i-1})E\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\} = 0 \quad (2.35)$$

and, under the assumption that the increments are uncorrelated we have the variance is the sum of the variance of each component to the sum. Thus

$$\text{Var}(U) = \sum E\{H^2(x_{i-1})[M(x_i) - M(x_{i-1})]^2|\mathcal{F}(x_{i-1})\}. \quad (2.36)$$

In order to keep the presentation uncluttered we use a single operator  $E$  in the above expressions, but there are some subtleties that ought not go unremarked. For instance, in Equation 2.36, the inner expectation is taken with respect to repetitions over all possible outcomes in which the set  $\mathcal{F}(x_{i-1})$  remains unchanged, whereas the outer expectation is taken with respect to all possible repetitions. In Equation 2.35 the outer expectation, taken with respect to the distribution of all potential realizations of all the sets  $\mathcal{F}(x_{i-1})$ , is not written and is necessarily zero since all of the inner expectations are zero. The analogous device to double expectation for the variance is not so simple since  $\text{Var}(Y) = E \text{Var}(Y|Z) + \text{Var} E(Y|Z)$ . Applying this we have

$$\text{Var}\{M(x_i) - M(x_{i-1})\} = E \text{Var}\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\} \quad (2.37)$$

since  $\text{Var} E\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\}$  is equal to zero, this being the case because each term is itself equal to the constant zero. The first term also requires a little thought, the outer expectation indicated by  $E$  being taken with respect to the distribution of  $\mathcal{F}(x_{i-1})$ , i.e., all the conditional distributions  $M(x)$  and  $H(x)$  where  $x \leq x_{i-1}$ . The next key point arises through the sequential nesting. These outer expectations, taken with respect to the distribution of  $\mathcal{F}(x_{i-1})$  are the same as those taken with respect to the distribution of any  $\mathcal{F}(x)$  for which  $x \geq x_{i-1}$ . This is an immediate consequence of the fact that the lower-dimensional distribution results from integrating out all the additional terms in the higher-dimensional distribution. Thus, if  $x_{\max}$  is the greatest value of  $x$  for which observations are made then we can consider that all of these outer expectations are taken with respect to  $\mathcal{F}(x_{\max})$ . Each time that we condition upon  $\mathcal{F}(x_{i-1})$  we will treat  $H(x_{i-1})$  as a fixed constant and so it can be simply squared and moved

outside the inner expectation. It is still governed by the outer expectation which, for all elements of the sum, we will take to be with respect to the distribution of  $\mathcal{F}(x_{\max})$ . Equation 2.36 then follows.

Making a normal approximation for  $U$ , and from the theory of estimating equations, given any set of observations, that  $U$  depends monotonically on some parameter  $\beta$ , then it is very straightforward to set up hypothesis tests for  $\beta = \beta_0$ . Many situations, including that of proportional hazards regression, lead to estimating equations of the form of  $U$ . The above set-up, which is further developed below in a continuous form, i.e., after having “added in” the term  $\lim_{\max \Delta_i \rightarrow 0}$ , applies very broadly. We need the concept of a process, usually indexed by time  $t$ , the conditional means and variances of the increments, given the accumulated information up until time  $t$ .

We have restricted our attention here to the Riemann-Stieltjes definition of the integral. The broader Lebesgue definition allows the inclusion of subsets of  $t$  tolerating serious violations of our conditions such as conditional means and variances not existing. The conditioning sets can be also very much more involved. Only in a very small number of applications has this extra generality been exploited. Given that it considerably obscures the main ideas to all but those well steered in measure theory, it seems preferable to avoid it altogether. Also avoided here is the martingale central limit theorem. This theorem is much quoted in the survival analysis context and, again, since there are so few applications in which the needed large sample normality cannot be obtained via more standard central limit theorems, a lack of knowledge of this theorem will not handicap the reader.

### *Counting processes*

The above discussion started off with some consideration of the empirical cumulative distribution function  $F_n(t)$  which is discussed in much more detail in Section 3.5. Let’s consider the function  $N(t) = \{nF_n(t) : 0 \leq t \leq 1\}$ . We can view this as a stochastic process, indexed by time  $t$  so that, given any  $t$  we can consider  $N(t)$  to be a random variable taking values from 0 to  $n$ . We include here a restriction that we generally make which is that time has some upper limit, without loss of generality, we call this 1. This restriction can easily be avoided but it implies no practical constraint and is often convenient in practical applications. We can broaden the definition of  $N(t)$  beyond that of  $nF_n(t)$  and we have:

**Definition 2.4** *A counting process  $N = \{N(t) : 0 \leq t \leq 1\}$  is a stochastic process that can be thought of as counting the occurrences (as time  $t$  proceeds) of certain type of events. We suppose these events occur singly.*

Very often  $N(t)$  can be expressed as the sum of  $n$  individual counting processes,  $N_i(t)$ , each one counting no more than a single event. In this case  $N_i(t)$  is a simple step function, taking the value zero at  $t = 0$  and jumping to the value one at the time of an event. The realizations of  $N(t)$  are integer-valued step functions with jumps of size  $+1$  only. These functions are right-continuous and  $N(t)$  is the (random) number of events in the time interval  $[0, t]$ . We associate with the stochastic process  $N(t)$  an intensity function  $\alpha(t)$ . The intensity function serves the purpose of standardizing the increments to have zero mean. In order to better grasp what is happening here, the reader might look back to Equation 2.34 and the two sentences following that equation. The mean is not determined in advance but depends upon  $\mathcal{F}_{t-}$  where, in a continuous framework,  $\mathcal{F}_{t-}$  is to  $\mathcal{F}_t$  what  $\mathcal{F}(x_{i-1})$  is to  $\mathcal{F}(x_i)$ . In technical terms:

**Definition 2.5** *A filtration,  $\mathcal{F}_t$ , is an increasing right continuous family of sub-sigma algebras.*

This definition may not be very transparent to those unfamiliar with the requirement of sigma additivity for probability spaces and there is no real need to expand on it here. The requirement is a theoretical one which imposes a mathematical restriction on the size, in an infinite sense, of the set of subsets of  $\mathcal{F}_t$ . The restriction guarantees that the probability we can associate with any infinite sum of disjoint sets is simply the sum of the probabilities associated with those sets composing the sum. For our purposes, the only key idea of importance is that  $\mathcal{F}_{t-}$  is a set containing all the accumulated information (hence “increasing”) on all processes contained in the past up until but not including the time point  $t$  (hence “right continuous”). We write,  $\alpha = \{\alpha(t) : 0 \leq t \leq 1\}$  where

$$\alpha(t)dt = \Pr\{N(t) \text{ jumps in } [t, t+dt) | \mathcal{F}_{t-}\} = E\{dN(t) | \mathcal{F}_{t-}\},$$

the equality being understood in an infinitesimal sense, i.e., the functional part of the left-hand side,  $\alpha(t)$ , is the limit of the right-hand side divided by  $dt > 0$  as  $dt$  goes to zero. In the chapter on survival analysis we will see that the hazard function,  $\lambda(t)$ , expressible as the ratio of the

density,  $f(t)$ , to the survivorship function,  $S(t)$ , i.e.,  $f(t)/S(t)$ , can be expressed in fundamental terms by first letting  $Y(t) = I(T \geq t)$ . Understanding, once again, the equality sign as described in the previous sentences but one, we have

$$\lambda(t)dt = \Pr \{N(t) \text{ jumps in } [t, t + dt) | Y(t) = 1\} = E\{dN(t) | Y(t) = 1\}.$$

It is instructive to compare the above definitions of  $\alpha(t)$  and  $\lambda(t)$ . The first definition is the more general since, choosing the sets  $\mathcal{F}_t$  to be defined from the at-risk function  $Y(t)$  when it takes the value one, enables the first definition to reduce to a definition equivalent to the second. The difference is an important one in that if we do not provide a value for  $I(T \geq t)$  then this is a  $(0, 1)$  random variable and, in consequence,  $\alpha(t)$  is a  $(0, \lambda(t))$  random variable. For this particular case we can express this idea succinctly via the formula

$$\alpha(t)dt = Y(t)\lambda(t)dt. \quad (2.38)$$

Replacing  $Y(t)$  by a more general “at risk” indicator variable will allow for great flexibility, including the ability to obtain a simple expression for the intensity in the presence of censoring as well as the ability to take on-board multistate problems where the transitions are not simply from alive to dead but from, say, state  $j$  to state  $k$  summarized via  $\alpha_{jk}(t)dt = Y_{jk}(t)\lambda_{jk}(t)dt$  in which  $Y_{jk}(t)$  is left continuous and therefore equal to the limit  $Y_{jk}(t - \epsilon)$  as  $\epsilon > 0$  goes to zero through positive values, an indicator variable taking the value one if the subject is in state  $j$  and available to make a transition to state  $k$  at time  $t - \epsilon$  as  $\epsilon \rightarrow 0$ . The hazards  $\lambda_{jk}(t)$  are known in advance, i.e., at  $t = 0$  for all  $t$ , whereas the  $\alpha_{jk}(t)$  are random viewed from time point  $s$  where  $s < t$ , with the subtle condition of left continuity which leads to the notion of “predictability” described below. The idea of sequential standardization, the repeated subtraction of the mean, that leans on the evaluation of intensities, can only work when the mean exists. This requires a further technical property, that of being “adapted.” We say

**Definition 2.6** *A stochastic process  $X(t)$  is said to be adapted to the filtration  $\mathcal{F}_t$  if  $X(t)$  is a random variable with respect to  $\mathcal{F}_t$ .*

Once again the definition is not particularly transparent to nonprobabilists and the reader need not be over-concerned since it will not be referred to here apart from in connection with the important concept of a predictable process. The basic idea is that the relevant quantities

upon which we aim to use the tools of probability modeling should all be contained in  $\mathcal{F}_t$ . If any probability statement we wish to construct concerning  $X(t)$  cannot be made using the set  $\mathcal{F}_t$  but requires the set  $\mathcal{F}_{t+u}$ , where  $u > 0$ , then  $X(t)$  is not adapted to  $\mathcal{F}_t$ . In our context just about all of the stochastic processes that are of interest to us are adapted and so this need not be a concern. A related property, of great importance, and which also will hold for all of those processes we focus attention on, is that of predictability. We have

**Definition 2.7** *A real-valued stochastic process,  $H(t)$ , that is left continuous and adapted to the filtration  $\mathcal{F}_t$  is called a predictable process.*

Since  $H(t)$  is adapted to  $\mathcal{F}_t$  it is a random variable with respect to  $\mathcal{F}_t$ . Since the process is left continuous it is also adapted to  $\mathcal{F}_{t-}$ . Therefore, whenever we condition upon  $\mathcal{F}_{t-}$ ,  $H(t)$  is simply a fixed and known constant. This is the real sense of the term “predictable” and, in practice, the property is a very useful one. It is frequently encountered in the probabilistic context upon which a great number of tests are constructed. Counting processes can be defined in many different ways and such a formulation allows for a great deal of flexibility. Suppose for instance that we have events of type 1 and events of type 2, indicated by  $N_1(t)$  and  $N_2(t)$  respectively. Then  $N(t) = N_1(t) + N_2(t)$  counts the occurrences of events of either type. For this counting process we have

$$\alpha(t)dt = P(N(t) \text{ jumps in } [t, t+dt) | \mathcal{F}_{t-}),$$

i.e., the same as  $P(N_1(t) \text{ or } N_2(t) \text{ jump in } [t, t+dt) | \mathcal{F}_{t-})$  and, if as is reasonable in the great majority of applications, where, we assume to be negligible the probability of seeing events occurring simultaneously compared to seeing them occur singly, then

$$\alpha(t)dt = E\{dN_1(t) + dN_2(t) | \mathcal{F}_{t-}\} = \alpha_1(t) + \alpha_2(t).$$

This highlights a nice linearity property of intensities, not shared by probabilities themselves. For example, if we consider a group of  $n$  subjects and  $n$  individual counting processes  $N_i(t)$ , then the intensity function,  $\alpha(t)$ , for the occurrence of an event, regardless of individual, is simply  $\sum \alpha_i(t)$ . This result does not require independence of the processes, only that we can consider as negligible the intensities we might associate with simultaneous events.

Another counting process of great interest in survival applications concerns competing risks. Suppose there are two types of event but that they cannot both be observed. The most common example of this is right censoring where, once the censoring event has occurred, it is no longer possible to make observations on  $N_i(t)$ . This is discussed more fully in the following chapters and we limit ourselves here to the observation that  $N_i(t)$  depends on more than one variable. In the absence of further assumptions, we are not able to determine the intensity function, but if we are prepared to assume that the censoring mechanism is independent of the failure mechanism, i.e., that  $\Pr(T_i > t | C_i > c) = \Pr(T_i > t)$ , then a simple result is available.

**Theorem 2.15** *Let the counting process,  $N_i(t)$ , depend on two independent and positive random variables,  $T_i$  and  $C_i$  such that  $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ . Let  $X_i = \min(T_i, C_i)$ ,  $Y_i(t) = I(X_i \geq t)$ ; then  $N_i(t)$  has intensity process*

$$\alpha_i(t)dt = Y_i(t)\lambda_i(t)dt. \quad (2.39)$$

The counting process,  $N_i(t)$ , is one of great interest to us since the response variable in most studies will be of such a form, i.e., an observation when the event of interest occurs but an observation that is only possible when the censoring variable is greater than the failure variable. Also, when we study a heterogeneous group, our principal focus in this book, the theorem still holds in a modified form. Thus, if we can assume that  $\Pr(T_i > t | C_i > c, Z = z) = \Pr(T_i > t | Z = z)$ , we then have:

**Theorem 2.16** *Let the counting processes,  $N_i(t)$ , depend on two independent and positive random variables,  $T_i$  and  $C_i$ , as well as  $Z$  such that*

$$N_i(t) = I\{T_i \leq t, T_i \leq C_i, Z = z\}. \quad (2.40)$$

*Then the intensity process for  $N_i(t)$  can be written as  $\alpha_i(t, z)dt = Y_i(t)\lambda_i(t, z)dt$ .*

The assumption needed for Theorem 2.16, known as the conditional independence assumption, is weaker than that needed for 2.15 in that the latter theorem contains the former as a special case. Note that the stochastic processes  $Y_i(t)$  and  $\alpha_i(t)$  are left continuous and adapted to  $\mathcal{F}_t$ . They are therefore predictable stochastic processes, which means that, given  $\mathcal{F}_{t-}$ , we treat  $Y_i(t)$ ,  $\alpha_i(t)$  and, assuming that  $Z(t)$  is predictable,  $\alpha_i(t, z)$  as fixed constants.

## 2.13 Exercises and class projects

1. Use a simple sketch to informally demonstrate the mean value theorem.
2. Newton-Raphson iteration provides sequentially updated estimates to the solution to the equation  $f(x_0) = 0$ . At the  $n$ th step, we write  $x_{n+1} = x_n - f(x_n)/f'(x_n)$  and claim that  $x_n$  converges (in the analytical sense) to  $x_0$ . Use the mean value theorem and, again, a simple sketch to show this. Intuitively, which conditions will lead to convergence and which ones can lead to failure of the algorithm.
3. Let  $g(x)$  take the value 0 for  $-\infty < x \leq 0$  ;  $1/2$  for  $0 < x \leq 1$  ; 1 for  $1 < x \leq 2$  ; and 0 otherwise. Let  $f(x) = x^2 + 2$ . Evaluate the Riemann-Stieltjes integral of  $f(x)$  with respect to  $g(x)$  over the real line.
4. Note that  $\sum_{i=1}^n i = n(n+1)/2$ . Describe a function such that a Riemann-Stieltjes integral of it is equal to  $n(n+1)/2$ . Viewing integration as an area under a curve, conclude that this integral converges to  $n^2$  as  $n$  becomes large.
5. Suppose that in the Helly-Bray theorem for  $\int h(x)dF_n(x)$ , the function  $h(x)$  is unbounded. Break the integral into components over the real line. For regions where  $h(x)$  is bounded the theorem holds. For the other regions obtain conditions that would lead to the result holding generally.
6. Prove the probability integral transformation by finding the moment-generating function of the random variable  $Y = F(X)$  where  $X$  has the continuous cumulative distribution function  $F(x)$  and a moment-generating function that exists.
7. If  $X$  is a continuous random variable with probability density function  $f(x) = 2(1-x)$ ,  $0 < x < 1$ , find that transformation  $Y = \psi(X)$  such that the random variable  $Y$  has the uniform distribution over  $(0,2)$ .
8. The order statistics for a random sample of size  $n$  from a discrete distribution are defined as in the continuous case except that now we have  $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ . Suppose a random sample of size 5 is

taken with replacement from the discrete distribution  $f(x) = 1/6$  for  $x = 1, 2, \dots, 6$ . Find the probability mass function of  $X_{(1)}$ , the smallest order statistic.

9. Ten points are chosen randomly and independently on the interval  $(0,1)$ . Find (a) the probability that the point nearest 1 exceeds 0.8, (b) the number  $c$  such that the probability is 0.4 that the point nearest zero will exceed  $c$ .

10. Find the expected value of the largest order statistic in a random sample of size 3 from (a) the exponential distribution  $f(x) = \exp(-x)$  for  $x > 0$ , (b) the standard normal distribution.

11. Find the probability that the range of a random sample of size  $n$  from the population  $f(x) = 2e^{-2x}$  for  $x \geq 0$  does not exceed the value 4.

12. Approximate the mean and variance of (a) the median of a sample of size 13 from a normal distribution with mean 2 and variance 9, (b) the fifth-order statistic of a random sample of size 15 from the standard exponential distribution.

13. Simulate 100 observations from a uniform distribution. Do the same for an exponential, Weibull and log-logistic distribution with different parameters. Next, generate normal and log-normal variates by summing a small number of uniform variates. Obtain histograms. Do the same for 5000 observations.

14. Obtain the histogram of 100 Weibull observations. Obtain the histogram of the logarithms of these observations. Compare this with the histogram obtained by the empirical transformation to normality.

15. Suppose that  $T_1, \dots, T_n$  are  $n$  exponential variates with parameter  $\lambda$ . Show that, under repeated sampling, the smallest of these also has an exponential distribution. Is the same true for the largest observation? Suppose we are only given the value of the smallest of  $n$  observations from an exponential distribution with parameter  $\lambda$ . How can this observation be used to estimate  $\lambda$ .



16. Suppose that  $X_i$   $i = 1, \dots, n$  are independent exponential variates with parameter  $\lambda$ . Determine, via simple calculation, the variance of  $\min(X_1, \dots, X_n)$ .
17. Having some knowledge of the survival distribution governing observations we are planning to study, how might we determine an interval of time to obtain with high probability a given number of failures? How should we proceed in the presence of censoring?
18. Derive the Bienaymé-Chebyshev inequality. Describe the advantages and drawbacks of using this inequality to construct confidence intervals in a general setting.
19. Suppose that the entropy described in Equation 2.14 depends on a parameter  $\theta$  and is written  $V_\theta(f, f)$ . Consider  $V_\alpha(f, f)$  as a function of  $\alpha$ . Show that this function is maximized when  $\alpha = \theta$ .
20. Using the device of double expectation derive Equation 2.15. Why is this breakdown interpreted as one component corresponding to “signal” and one component corresponding to “noise.”
21. Suppose that  $\theta_n$  converges in probability to  $\theta$  and that the variance of  $\theta_n$  is given by  $\psi(\theta)/n$ . Using Equation 2.19, find a transformation of  $\theta_n$  for which, at least approximately, the variance does not depend on  $\theta$ .
22. Consider a stochastic process  $X(t)$  on the interval  $(2, 7)$  with the following properties: (a)  $X(0) = 2$ , (b)  $X(t), t \in (2, 7)$  has increments such that (c), for each  $t \in (2, 7)$  the distribution of  $X(t)$  is Weibull with mean  $2 + \lambda t^\gamma$ . Can these increments be independent and stationary? Can the process be described using the known results of Brownian motion?
23. For Brownian motion, explain why the conditional distribution of  $X(s)$  given  $X(t)$  ( $t > s$ ) is normal with  $E\{X(s)|X(t) = w\} = ws/t$  and  $\text{Var}\{X(s)|X(t) = w\} = s(t - s)/t$ . Deduce the mean and the covariance process for the Brownian bridge.
24. The Ornstein-Uhlenbeck process can be thought of as transformed Brownian motion in which the variance has been standardized. Explain why this is the case.

25. Reread the subsection headed “Time-transformed Brownian motion” (Section 2.11) and conclude that the only essential characteristic underwriting the construction of Brownian motion is that of independent increments.

26. Find the value of  $t \in (0, 1)$  for which the variance of a Brownian bridge is maximized.

27. Suppose that under  $H_0$ ,  $X(t)$  is Brownian motion. Under  $H_1$ ,  $X(t)$  is Brownian motion with drift, having drift parameter 2 as long as  $X(t) < 1$  and drift parameter minus 2 otherwise. Describe likely paths for reflected Brownian motion under both  $H_0$  and  $H_1$ . As a class exercise simulate ten paths under both hypotheses. Comment on the resulting figures.



<http://www.springer.com/978-0-387-25148-6>

Proportional Hazards Regression

O'Quigley, J.

2008, XVIII, 542 p. 41 illus., Hardcover

ISBN: 978-0-387-25148-6