1

INTRODUCTION TO ASYMPTOTIC CONVERGENCE

1.1 INTRODUCTION

The first and major proportion of this book is concerned with both asymptotic theory and empirical evidence for the convergence of estimators. The author has contributed here in more than just one article. Mostly, the relevant contributions have been in the field of M-estimators, and it is the purpose of this book to make some ideas that have necessarily been couched in deep theory of continuity and also differentiability of functions more easily accessible and understandable. We do this by illustration of convergence concepts which begin with the strong law of large numbers (SLLN) and then eventually make use of the central limit theorem (CLT) in its most basic form. These two results are central to providing a straightforward discussion of M-estimation theory and its applications. The aim is not to give the most general results on the theory of M-estimation nor the related L-estimation discussed in the latter half of the book, for these have been adequately displayed in other books including Jurečková et al. (2012), Maronna et al. (2006), Hampel et al. (1986), Huber and Ronchetti (2009), and, Staudte and Sheather (1990). Rather, motivation for the results of consistency and asymptotic normality of M-estimators is explained in a way which highlights what to do when one has more than one root to the estimating equations. We should not shy

Robustness Theory and Application, First Edition. Brenton R. Clarke.

Companion website: www.wiley.com/go/clarke/robustnesstheoryandapplication

away from this problem since it tends to be a recurrent theme whenever models become quite complex having multiple parameters and consequently multiple simultaneous potentially nonlinear equations to solve.

1.2 PROBABILITY SPACES AND DISTRIBUTION FUNCTIONS

We begin with some basic terminology in order to set the scene for the study of convergence concepts which we then apply to the study of estimating equations and also loss functions. So we assume that there is an "Observation Space" denoted by $\tilde{R}$, which can be a subset of a separable metric space R. In published papers by the author, it was assumed typically that $\tilde{R}$ was a separable metric space, which did not allow for discrete data observed on, say, the nonnegative integers (such as Poisson distributed data). However, it is enough to consider now $\tilde{R} \subset R$ since the arguments follow through easily enough. So the generality of the discussion includes data that are either continuous or discrete, and either univariate or multivariate, or say defined on the positive real line such as in lifetime data.

For instance, if the data are k-dimensional continuous multivariate data, we have that $R \equiv \mathcal{E}^k$. Then we let $\mathcal{B}$ be the smallest σ-field containing the class of open sets on R generated by the metric on R. These are called the k-dimensional Borel sets, and $\mathcal{B}$ is known as the Borel σ-field. See Problem 1.1 for the definition of a σ-field.

A distribution on R is a nonnegative and countably additive set function, μ, on $\mathcal{B}$, for which $\mu\{R\} = 1$, and it is well known say that on $R = \mathcal{E}$ there corresponds a unique right continuous function F whose limits are zero and one at $-\infty$ and $+\infty$, defined by $F(x) = \mu\{(-\infty, x]\}$.

We shall denote an abstract probability space to be $(\Omega, \tilde{\mathcal{A}}, \mathcal{P})$. This formulation assumes $\tilde{\mathcal{A}}$ to be a σ-field of subsets of Ω, with $\mathcal{P}$ a probability measure on $\tilde{\mathcal{A}}$. Ω is thought of as a sample space and elements of Ω, denoted by ω, are the outcomes. Then a sequence of random variables on Ω is defined via

$$\boldsymbol{X}(\omega) = X_1(\omega), X_2(\omega), \ \ldots\ , X_n(\omega), \ \ldots\ , \tag{1.1}$$

taking values in the infinite product space $(R^\infty, \mathcal{B}^\infty)$. The observed sample of size n is then written as

$$(X_1(\omega), \ \ldots\ , X_n(\omega)) = \pi^{(n)} o \boldsymbol{X}(\omega),$$

while the nth random variable is given by $X_n(\omega) = \pi_n o \boldsymbol{X}(\omega)$. Both $\pi^{(n)}$ and π_n are then what are termed measurable maps with respect to $\mathcal{B}^\infty$. They induce distributions $G^{(n)}$ and G_n on $(R^n, \mathcal{B}^n)$ and $(R, \mathcal{B})$, respectively. A useful reference on equivalent representations of infinite sequences of random variables and probability measures is that of Chung (2001), (now 3rd ed.) (1st ed., pp. 54–58).

We use the symbol $\mathcal{G}$ to denote the space of distributions on $(R, \mathcal{B})$. Consider an arbitrary set $A^{(n)} = (A_1 \times A_2 \times \ldots \times A_n) \in \mathcal{B}^n$. Denote by $G^{(n)}$ the product measure on $(R^n, \mathcal{B}^n)$ that gives

$$\mathcal{P}(\pi^{(n)}\boldsymbol{X}(\omega) \in A_1 \times A_2 \times \ldots \times A_n) = \int_{A^{(n)}} \mathrm{d}G^{(n)}.$$

Then we say that the sequence $\boldsymbol{X}$ is independent identically distributed (*i.i.d.*) if there exists a $G \in \mathcal{G}$ such that for every $A^{(n)}$ in the form above

$$\int_{A^{(n)}} \mathrm{d}G^{(n)} = \int_{A_1} \mathrm{d}G \times \ldots \times \int_{A_n} \mathrm{d}G.$$

1.3 LAWS OF LARGE NUMBERS

The law of large numbers (LLN) is a result that describes what will happen if we repeat an experiment a large number of times. To couch it in simple terms, it is when the average of the results/experiments obtained from a large number of trials should be close to its expected value. How we measure closeness and relate that to a limit of the number of trials, tending to infinity say, requires us to introduce two possible modes of convergence of random variables.

1.3.1 Convergence in Probability and Almost Sure

For a sequence $\{X_n\}$ of random variables on $(\Omega, \tilde{\mathcal{A}}, \mathcal{P})$, we can define two modes of convergence to a random variable X on that same probability space. The first and weaker mode of convergence, convergence in probability, is defined as follows: Let d be the metric distance on R. Then we say, provided the random variables are defined on the same probability space, X_n converges in probability to X if for every $\epsilon > 0$

$$\mathcal{P}\{\omega \in \Omega | d(X_n, X) < \epsilon\} \to 1 \quad \text{as} \quad n \to \infty. \tag{1.2}$$

We say

$$X_n \to_p X. \tag{1.3}$$

Should X be a constant and say $\tilde{R} \subset \mathcal{E}^k$, then say $\mathcal{P}(X = \text{const}) = 1$ the definition can be modified to say $X_n \to_p \text{const}$ if for every $\epsilon > 0$

$$\mathcal{P}\{|X_n - \text{const}| < \epsilon\} \to 1 \quad \text{as} \quad n \to \infty.$$

A stronger form of convergence of say X_n to the random variable X is convergence with probability one or almost sure convergence

$$\mathcal{P}\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1.$$

It is a fact that this form of convergence implies convergence in probability. Hence, statements of almost sure convergence can be preferred to statements made only in probability.

1.3.2 Expectation and Variance

In order to relate the LLN, we need a formal concept of the expected value of a random variable. For a random variable on $(\Omega, \tilde{\mathcal{A}}, \mathcal{P})$, its expected value is written $E[X] = \int_\Omega X(\omega)\mathrm{d}\mathcal{P}(\omega) = \int_R x\mathrm{d}G(x)$, where G is the induced distribution function on the full space R that contains the observation space $\tilde{R}$. Thus, for observations on a subset of the real line, that is where $R \subset \mathcal{E}$ we have then, denoting $\mu = E[X]$,

$$\mu = \int_{-\infty}^{+\infty} x\mathrm{d}G(x).$$

This is the mean value of the random variable X. Similarly, we may define the variance assuming it exists, via

$$\mathrm{var}[X] = E[(X-\mu)^2] = \int_\Omega (X(\omega)-\mu)^2\mathrm{d}\mathcal{P}(\omega) = \int_R (x-\mu)^2\mathrm{d}G(x).$$

As is typically done, we denote the variance $\sigma^2 = \mathrm{var}(X)$. The variance measures the spread of the observations about the expected value. Since the variance is involving squared deviations from the mean, it is not in the same units of measure as the original variable. Hence, what is more often used as a measure of spread is the standard deviation which is the square root of the variance, that is $\sigma = \sqrt{\sigma^2}$. Nevertheless, the variance plays an important part in convergence concepts and indeed makes proofs, even of the LLN in its weaker form, much easier when it is known to exist, that is when it is finite.

1.3.3 Statements of the Law of Large Numbers

The LLN in either its weaker or stronger form state that – with almost certainty - the sample average

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

converges to the expected value (assumed finite) so that

$$\overline{X}_n \to \mu \quad \text{for } n \to \infty.$$

Here $X_1, X_2, \ldots$ is assumed *i.i.d.* and consequently $E[X_1] = \ldots = E[X_n] = \mu$. It is not necessary that $\text{var}[X_1] = \ldots = \text{var}[X_n] = \sigma^2 < +\infty$. The LLN will hold anyway. Large or infinite variance makes convergence "slower." Finite variance makes proofs easier and shorter. Other ways that the theorem can be extended is by relaxing assumptions of random variables being independent and/or identically distributed. Such considerations may become important when allowing for the incorporation of ideas such as often promulgated in the robustness literature, where it is usually the case that model assumptions can and often do fail. These ideas will become more apparent once we travel on to statistical inference.

Statement of the Weak Law of Large Numbers

The weak law of large numbers (WLLN) states that

$$\overline{X}_n \rightarrow_p \mu \quad \text{as} \quad n \rightarrow \infty$$

Equivalently, for any $\epsilon > 0$

$$\lim_{n \rightarrow +\infty} \mathcal{P}\{|\overline{X}_n - \mu| > \epsilon\} = 0.$$

Interpretation of the WLLN is that for any, however, small nonzero deviation from the mean (governed by setting a value for ϵ say), with a suitably large sample size, there is a high probability that the sample average is close to the expected value, that is within that deviation from the mean.

The WLLN involves convergence in probability and there are examples where a random variable converges in probability but does not converge almost surely. Hence, this is the reason for it to be the weak LLN.

Statement of the SLLN

The SLLN states that

$$\mathcal{P}\{\omega | \lim_{n \rightarrow +\infty} \overline{X}_n(\omega) = \mu\} = 1$$

or put more succinctly,

$$\overline{X}_n \rightarrow_{a.s.} \mu.$$

As noted above, convergence almost surely implies convergence in probability, whence the SLLN implies the WLLN.

1.3.4 Some History and an Example

Probability theory, both modern and historical, has been inspired by the need to analyze games of chance. See for instance that the Italian mathematician Geralamo Cardono, who was well versed in games of chance, is credited by Mlodinow (2008) with writing down a version of the law of large numbers, stating without proof that "the accuracies of empirical statistics tend to improve with the number of trials." Indeed, Jakob Bernoulli (1713) examined a special form of the LLN

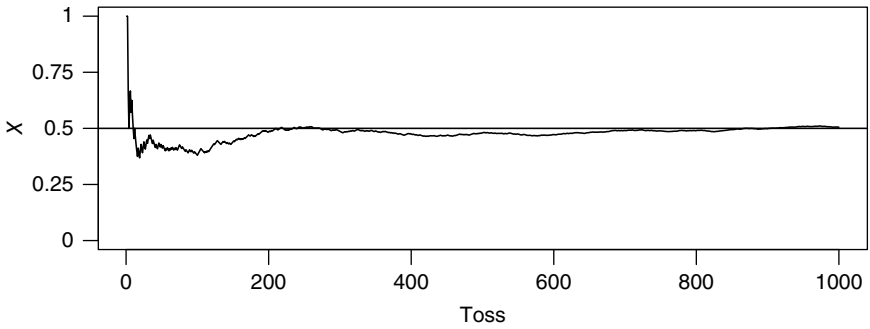

FIGURE 1.1 Illustration of the SLLN in action by plotting the average of Bernoulli random variables with probability $p = 0.5$ as $n =$ number of tosses becomes large.

for a binary random variable and is credited with the first proof. To illustrate the LLN in practice, we consider a sequence of Bernoulli random variables, where with probability one half the outcome of a random variable in the sequence is one, whereas with probability one half the outcome of the random variable is zero. For instance, we may be watching a sequence of independent coin tosses where the random variable is denoted a one if a head occurs or a zero if a tail occurs. Constructing such a sequence is known to be fraught with difficulty as can be gained from reading Diaconis et al. (2007). Of course Bernoulli random variables can be assigned more generally if we had "biased" coins where the probability of a head was some p where $0 < p < 1$, not necessarily $p = \frac{1}{2}$, and the consequent probability of a tail is $q = 1 - p$. In terms of a sample space and possible sequences of random variables, we can consider a random number generator on a computer starting an infinite sequence, the random seed to start that sequence being generated by the time of day. Assuming an infinite number of possible starting points if we choose just one starting point, that is a particular $\omega \in \Omega$ with probability one, we will have chosen a sequence for which the average of those observed coin tosses converges to the expected value of the random variable. Basic maths can show that the expected value of the Bernoulli random variable is a weighted average of the outcomes, $\mu = 1 \times p + 0 \times (1 - p) = p$, whence in the limit if we set $p = 1/2$ if we observe, with probability one, just one sequence for long enough we should see the average of the random variable converges to 0.5. A randomly selected graph is given in Figure 1.1, which demonstrates the convergence by plotting the sample average for up to one thousand "tosses" of the Bernoulli random number generator.

1.3.5 Some More Asymptotic Theory and Application

Consider now a sequence of *i.i.d.* random variables $X_1, X_2, \ldots$. For clarity of exposition, we give the following formulation that helps to relate statements

about events to almost sure convergence. *Definition 1* The sequence of statements $A_1(X_1), A_2(X_1, X_2), \ldots$ are said to hold *for all sufficiently large n*, (*f.a.s.l.n*) if

$$\mathcal{P}\{\omega | \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n(X_1(\omega), \ldots, X_n(\omega))\} = 1.$$

Using this definition it is clear that a sequence of measurable maps $\{T_n(X_1, \ldots, X_n)\}$ from the domain space R^n to the codomain $\mathcal{E}^r$ converges almost surely to T if and only if for every $\eta > 0$ the sequence of statements

$$|T_n(X_1, \ldots, X_n) - T| < \eta \quad n = 1, 2, \ldots$$

holds *f.a.s.l.n*. For an account of the above definition, see Foutz and Srivastava (1979). Interestingly, we have that if we have two sequences of statements that are both known to hold *f.a.s.l.n* say

$$A_1(X_1), A_2(X_1, X_2), \ldots$$

and

$$B_1(X_1), B_2(X_1, X_2), \ldots$$

then the sequence of statements

$$A_1(X_1) \cap B_1(X_1), A_2(X_1, X_2) \cap B_2(X_1, X_2), \ldots$$

also holds *f.a.s.l.n*. This is for the simple reason that since both

$$A = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n(X_1(\omega), \ldots, X_m(\omega))$$

and

$$B = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} B_n(X_1(\omega), \ldots, X_m(\omega))$$

are both sets of probability one whence $\mathcal{P}\{A \cap B\} = 1$, and since also we have the identity

$$A \cap B = \{\cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n\} \cap \{\cup_{m=1}^{\infty} \cap_{n=m}^{\infty} B_n\} = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} (A_n \cap B_n)$$

The implications of this result are that to check any finite number of statements that hold *f.a.s.l.n*. they then hold simultaneously together *f.a.s.l.n*. Thus, automatically if say one set of statements is that there is a root of a random function in say an interval $(\mu - \eta, \mu + \eta)$ and also it is known that the random function is monotonic in the same interval and both statements hold *f.a.s.l.n*, then this allows us to conclude there is a unique zero in that interval *f.a.s.l.n*, for example.

Returning to the SLLN, if there is any real valued measurable function $f(\cdot)$ from the space R to the space $\mathcal{E}$, and as before $X_1, X_2, \ldots$ are *i.i.d.* variables, then

$f(X_1), f(X_2), \ldots$, forms an independent sequence of random variables on $\mathcal{E}$. If in addition $E[|f(X)|] < +\infty$, then by the SLLN

$$\frac{f(X_1) + f(X_2) + \ \ldots \ + f(X_n)}{n} - E[f(X)] \rightarrow_{a.s.} 0. \tag{1.4}$$

The property (1.4) is known to hold for more general sequences than *i.i.d.* sequences and is known as an "ergodic property." Ergodic sequences include stationary sequences satisfying appropriate mixing conditions, but suffice to say the discussion in this book relates in the main to *i.i.d.* random variables.

Now it is possible that one may be interested in setting up a class of functions $\mathcal{A}$ which could be infinite in number, yet one can achieve the following result

$$\sup_{f \in \mathcal{A}} \left| \frac{f(X_1) + f(X_2) + \ \ldots \ + f(X_n)}{n} - E[f(X)] \right| \rightarrow_{a.s.} 0. \tag{1.5}$$

Here at the very least we may require for each $f \in \mathcal{A}$ to be bounded in modulus above by some g where $E[|g(X)|] < +\infty$. In addition, we may also require the functions f to satisfy certain equicontinuity or maybe even stronger continuity conditions which we shall elaborate later. For instance as in Rao (1962), $\mathcal{A}$ is equicontinuous at each $x \in X$, i.e. for each $\epsilon > 0$ there exists a neighborhood N of x such that $|f(y) - f(x)| < \epsilon$, for all $y \in N$ and all $f \in \mathcal{A}$.

1.4 THE MODUS OPERANDI RELATED BY LOCATION ESTIMATION

To illustrate why we would concern ourselves with these abstract concepts, we shall examine the class of functions associated with what is known as the Tukey bisquare function. This is governed by the formula

$$\psi_{\mathrm{BS}}(x) = \begin{cases} x(1 - x^2/c^2)^2 & -c < x < c \\ 0 & \text{Otherwise} \end{cases} \tag{1.6}$$

The constant $c > 0$ may be chosen to be fixed, and for illustration purposes we choose it to be the value 4.685. See Figure 1.2. This function, which is an odd function, is used in many discussions of what are called M-estimators. It was first introduced by Beaton and Tukey (1974) and is prominent in the discussion of M-estimators of location parameters, particularly when it is possible that one has potential contamination in the tails of one's data set. To set the scene, we can imagine a normal parametric location model, that is our sequence of data is generated from a normal distribution with mean μ and the model parametric family is

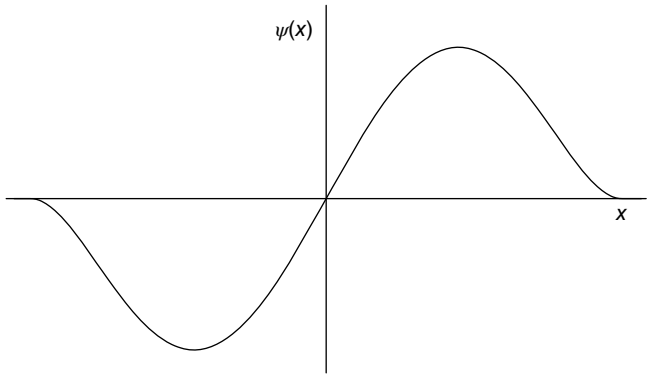

FIGURE 1.2 Illustration of the Tukey bisquare psi function.

$\{F_\tau(x) = \Phi(x-\tau) : \tau \in \mathcal{E}\}$, where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\mathrm{d}u$. To estimate the parameter μ given our sequence of data, we solve the equations

$$K_n(\tau) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \tau) = 0 \tag{1.7}$$

For any fixed τ the normalized sum converges by the SLLN to $K_G(\tau) = \int \psi(x-\tau)\phi(x-\mu)\mathrm{d}x$, where $G = F_\mu$ and $\phi(x) = \Phi'(x)$ is the normal density. The equations are such that $K_G(\mu)$=0, hence a good candidate for an estimate of μ is a zero or root of the empirical equations (1.7). Our class of functions then is $\mathcal{A} = \{\psi(\cdot - \tau) : \tau \in \mathcal{E}\}$. Without loss of generality we set $\mu = 0$ and examine the empirical functions $K_n(\tau)$ that you see for instance in generating several data sets. See for instance Figure 1.3. For such a small sample size, it is clear that the SLLN is only just beginning to work. The 10 plots exhibit a great deal of variability; however, it can be noticed that the zeros of the functions $K_n(\tau)$ that are in the middle of the graphs are centered approximately around the true mean of $\mu = 0$. Clearly also there are infinitely many zeros in the tails of these redescending curves, which carry little or no information about the central roots of the curves. They are an artifact of the redescending nature of ψ_{BS}. Increasing the sample size to $n = 50$ in Figure 1.4 shows the fairly uniform convergence of the empirical curves $K_n(\tau)$ to the asymptotic curve, in this case $K_\Phi(\tau)$, and this convergence is even more graphic in the plots of Figures 1.5 and 1.6 when $n = 100$ and 500, respectively.

Interestingly, the roots about the central value of zero are asymptotically unique. To support this claim, we can examine the derivative of the function $K_n(\tau)$, call it $K_n'(\tau)$. See Figure 1.7. For a sample size of $n = 100$, it is clear that for $\tau \in [-1, 1]$ the function derivative has a value which is almost certainly less

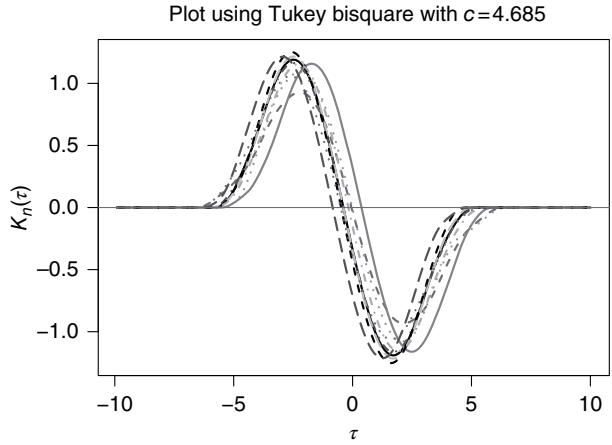

FIGURE 1.3 Illustration of SLLN for the family of functions involving the Tukey bisquare; sample size $n = 5$. Plots are for 10 independent samples.

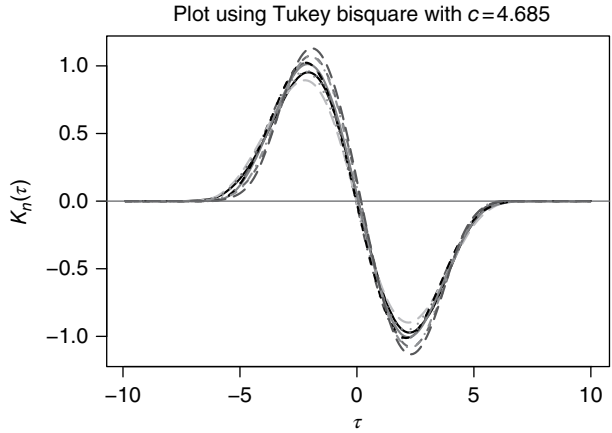

FIGURE 1.4 Illustration of SLLN for the family of functions involving the Tukey bisquare; sample size $n = 50$. Plots are for 10 independent samples.

than say -0.2. For large n it is also clear that the zero of $K_n(\tau)$ is say inside $[-1, 1]$. We therefore conclude it is a unique zero in this region since the empirical curve is almost certainly monotonically decreasing.

The appearance of the Tukey bisquare function came after the introduction of redescending M-estimators of location, initially proposed in Andrews et al. (1972) and further related in Hampel (1974). Due to the existence of multiple roots of the estimating equations, Hampel's recommendation was to choose the root closest to the median. Since the median is consistent to the true $\mu = 0$, it is therefore clear that the root that is closest to the median should be the asymptotically unique

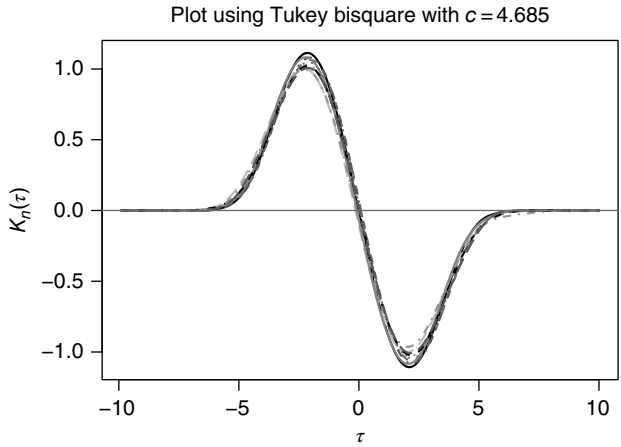

FIGURE 1.5 Illustration of SLLN for the family of functions involving the Tukey bisquare; sample size $n = 100$. Plots are for 10 independent samples.

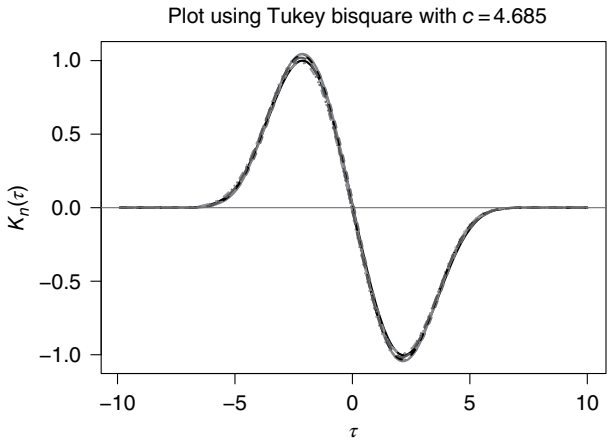

FIGURE 1.6 Illustration of SLLN for the family of functions involving the Tukey bisquare; sample size $n = 500$. Plots are for 10 independent samples.

consistent root of the estimating equations, given our discussion above. Of course numerically finding a unique consistent root of the equations relies on the choice of algorithm. By far the most well-known algorithm is that of Newton–Raphson, where the iterations are

$$\tau_{\nu+1}^{(n)} = \tau_{\nu}^{(n)} - \frac{K_n(\tau_{\nu}^{(n)})}{K_n'(\tau_{\nu}^{(n)})}, \quad \nu = 0, 1, 2, \ldots \tag{1.8}$$

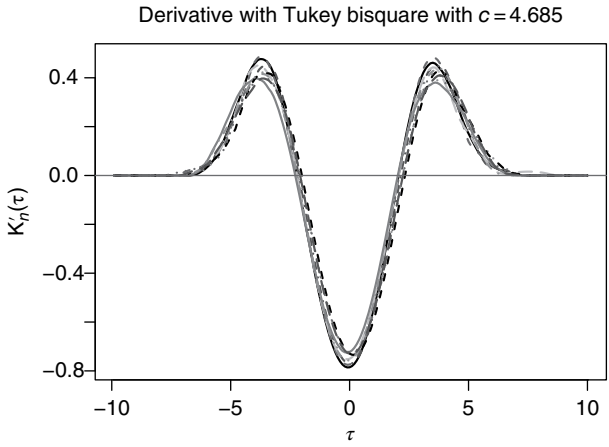

FIGURE 1.7 Illustration of SLLN for the family of functions involving the derivative of the Tukey bisquare; sample size $n = 100$. Plots are for 10 independent samples.

and where the superscript (n) indicates dependence on the random equations (1.7). Convergence of the iteration to a unique consistent root ξ_n, starting with a suitable initial, possibly random, estimate, relies on almost sure uniform convergence of the sequence $K_n(\tau)$ and its derivative $K_n'(\tau)$ to $K_G(\tau)$ and $K_G'(\tau)$, respectively. That is, for any given $\delta > 0$ and arbitrary compact sets C the sequence of statements

$$\sup_{\tau \in C} |K_n(\tau) - K_G(\tau)| < \delta, \quad \sup_{\tau \in C} |K_n'(\tau) - K_G'(\tau)| < \delta, \quad n = 1, 2, \ldots \tag{1.9}$$

hold *f.a.s.l.n.* Here sup denotes the least upper bound (or supremum) of the set. These conditions are easily derived in the case of the Tukey bisquare, ψ_{BS}, as it is a bounded smooth function with continuous and bounded derivatives. See the discussion and proof of Theorem 1 and Lemma 5.1 in Clarke (1986a). Note that the convergence given there is shown to be uniform over the whole real line, replacing C by $\mathcal{E}$. Thus, the figures above reflect this asymptotic convergence seen in Eq. (1.9).

Indeed the paper of Clarke (1986a) goes on to discuss the region from which the Newton–Raphson iteration converges to the central root of the estimating equations. Such a region is asymptotically governed by the curve $K_G(\tau)$ and so an upper limit on such a region is a solution μ^* of the equation

$$S_G(\tau) = 2(\tau - \mu) - \frac{K_G(\tau)}{K_G'(\tau)} = 0.$$

This equation gives the boundary values for the Newton–Raphson domain of attraction in the asymptotic equation $K_G(\tau) = 0$ with respect to the root μ. Without losing generality setting $\mu = 0$ the smallest positive root when $G = \Phi$

is μ_m^*=1.5737. Starting from anywhere in the interval (−1.5737, 1.5737), the Newton–Raphson algorithm applied to the asymptotic equation $K_\Phi(\tau) = 0$ will converge to $\mu = 0$. Obviously, for a finite sample there will be adjustments to make, which are detailed in that paper, but starting from a consistent estimate of μ, such as the median, one can see that the Newton–Raphson algorithm will converge to the root of interest.

It remains then to describe the properties of the root of interest, especially since one not only wishes for a point estimate but rather an idea of the variability of the point estimate. To explain this, we introduce the Lindeberg–Lévy CLT. Consider a sequence of random variables which are independent and identically distributed with mean μ and variance $\sigma^2 < +\infty$ and suppose we are interested in the sample average

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

The law of large numbers gives the asymptotic almost sure limit for the average as being μ. The classical CLT describes the way the averages fluctuate about the asymptotic limit of μ as they converge.

Theorem 1.1: *Lindeberg–Lévy CLT. Suppose* $\{X_1, X_2, \ldots\}$ *is a sequence of independent identically distributed random variables with* $E[X_i] = \mu$ *and* $\text{var}[X_i] = \sigma^2 < +\infty$. *Then as* n *approaches* ∞, *the random variable* $\sqrt{n}(\overline{X}_n - \mu)$ *converges in distribution to a* $N(0, \sigma^2)$ *random variable.*

$$\sqrt{n}(\overline{X}_n - \mu) \rightarrow_d N(0, \sigma^2).$$

For $0 < \sigma < +\infty$, *convergence in distribution means that the cumulative distribution functions of* $\sqrt{n}(\overline{X}_n - \mu)$ *converge pointwise in the sense that*

$$\lim_{n \rightarrow +\infty} \mathcal{P}\{\sqrt{n}\left(\overline{X}_n - \mu\right) \leq z\} = \Phi\left(\frac{z}{\sigma}\right)$$

The convergence is uniform in z where

$$\lim_{n \rightarrow +\infty} \sup_{z \in \mathcal{E}} |\mathcal{P}\{\sqrt{n}(\overline{X}_n - \mu) \leq z\} - \Phi\left(\frac{z}{\sigma}\right)| = 0$$

It is now possible to apply this result to the central root of the estimating equations (1.7). Visualize or denote $\{\hat{\mu}_n\}$ to be the central root of the equations, for example as obtained by a Newton–Raphson iteration limit starting from say a median. To be explicit, we give the existence of an asymptotically unique consistent estimator here, since the proof illustrates the sort of continuity arguments in play and which can be extended to multivariate parameters in Chapter 2. Since not all ψ-functions for M-estimators are continuously differentiable (even though

ψ_{BS} is) we will relax the conditions accordingly and adapt the notation to this chapter.

Lemma 1.1: *(Clarke, 1986a, Lemma 8.1) Assume ψ is continuous. Denote by $K_n'(\tau)$ a left-hand derivative of K_n and suppose for any compact C Eq. (1.9) holds f.a.s.l.n. Let $K_G(\mu)$ be continuously differentiable, $K_G(\mu)$=0 and $K_G'(\mu) < 0$. Then there exists an interval of some radius $\delta > 0$ about μ, such that*

(a) there exists a consistent sequence $\{\mu_n\}$ of zeros of $\{K_n(\tau)\}$ in $(\mu - \delta, \mu + \delta)$, and

(b) for any other sequence $\{\tilde{\theta}_n\}$ of zeros of $\{K_n(\tau)\}$ in $(\mu - \delta, \mu + \delta)$, $\tilde{\theta} = \mu_n$ f.a.s.l.n.

Proof. By continuity let $\delta > 0$ so that $K_G'(\tau) < \frac{1}{2}K_G'(\mu) = y_0$ say, for $\tau \in (\mu - \delta, \mu + \delta)$. From Eq. (1.9),

$$K_n'(\tau) < \frac{1}{2}y_0 \text{ uniformly in } \tau \in [\mu - \delta, \mu + \delta] \tag{1.10}$$

Then by Eq. (1.9) and the mean value theorem the statement

$$K_n(\mu - \delta) > K_G(\mu - \delta) + y_0\delta > 0 = K_G(\mu) > K_G(\mu + \delta) - y_0\delta > K_n(\mu + \delta)$$

holds f.a.s.l.n. Continuity of $K_n(\tau)$ implies existence of a zero $\mu_n \in (\mu - \delta, \mu + \delta)$ f.a.s.l.n. and by Eq. (1.10) this zero must be unique in the given interval. Since δ is arbitrary $\mu_n \rightarrow_{a.s.} \mu$. □

Having established the existence of an asymptotically unique consistent sequence of roots, we find from the mean value theorem

$$0 = K_n(\hat{\mu}_n) = K_n(\mu) + K_n'(\xi_n)(\hat{\mu}_n - \mu), \tag{1.11}$$

where ξ_n is a point lying somewhere between $\hat{\mu}_n$ and μ. Recalling Eq. (1.9) and the facts that $K_G'(\cdot)$ is continuous and also necessarily $\xi_n \rightarrow_{a.s.} \mu$, it is clear that $K_n'(\xi_n) \rightarrow_{a.s.} K_G'(\mu) = \int_{-\infty}^{+\infty} \psi_{\mathrm{BS}}'(x)\phi(x)\mathrm{d}x$, which also implies

$$K_n'(\xi_n) \rightarrow_p K_G'(\mu) = -\int_{-\infty}^{+\infty} \psi_{\mathrm{BS}}'(x)\phi(x)\mathrm{d}x.$$

We now intend to invoke Slutsky's theorem. Consider two sequences $\{U_n\}$ and $\{V_n\}$, where U_n converges in distribution to a random variable U and V_n converges in probability to a constant k. Then $U_n/V_n \rightarrow_d U/k$ provided k is invertible. Rewriting Eq. (1.11) we have

$$\sqrt{n}(\hat{\mu}_n - \mu) = \frac{-\sqrt{n}K_n(\mu)}{K_n'(\xi_n)}. \tag{1.12}$$

It is now easy to see in virtue of the CLT that in the case where $G = \Phi$, that is the normal model holds, then

$$\sqrt{n}K_n(\mu) = \frac{1}{\sqrt{n}} \sum \psi_{\mathrm{BS}}(X_i - \mu) \tag{1.13}$$

$$\rightarrow_d N(0, \int_{-\infty}^{+\infty} \psi_{\mathrm{BS}}^2(x)\phi(x)\mathrm{d}x) \tag{1.14}$$

here noting that the sequence of variables $\{\psi_{\mathrm{BS}}(X_i - \mu)\}_{i=1}^{\infty}$ are independent with zero mean (the curve $\psi_{\mathrm{BS}}(\cdot)$ is odd) and variance $E[\psi_{\mathrm{BS}}^2(Z)]$, where Z is a standard normal random variable. Now employing Slutsky's theorem we find

$$\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow_d N\left(0, \sigma_{\mathrm{BS}}^2 = \frac{E[\psi_{\mathrm{BS}}^2(Z)]}{E[\psi_{\mathrm{BS}}'(Z)]^2}\right).$$

As a reflection of this asymptotic result, and what it may tell us about the location of the true mean, we refer to Figure 1.8 which is a histogram of 10 000 location estimates, found by starting at the median and doing five iterations of the Newton–Raphson algorithm. These are obtained using the Tukey bisquare as detailed earlier with sample size $n = 100$. Notice that the histogram is bell shaped, reflecting asymptotic normality of the estimates. Another way of viewing these results is

$$\mathcal{P}\{\sqrt{n}\left(\frac{\hat{\mu}_n - \mu}{\sigma_{BS}}\right) \leq z\} \rightarrow \Phi(z)$$

which suggests that a 95% confidence interval for the true mean given an estimate from any one sample is $(\hat{\mu}_n - 1.96 \times \sigma_{\mathrm{BS}}/\sqrt{n},\ \hat{\mu}_n + 1.96 \times \sigma_{\mathrm{BS}}/\sqrt{n})$. In this example with $c = 4.685$, the value of $\sigma_{\mathrm{BS}} = 1.026$ whereupon approximately 95% of estimates from the simulation will lie between ± 0.201 from the true mean which is zero in the illustration.

The primary aim of this section has been to introduce ideas of asymptotic convergence that can be illustrated, in this instance, using the vehicle of the redescending Tukey bisquare function, heavily used in the theory of robustness. As yet we

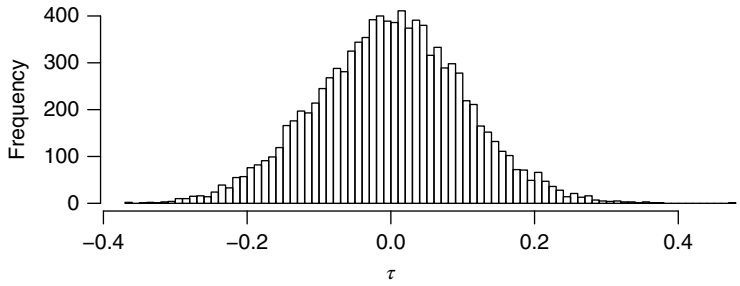

FIGURE 1.8 Histogram of the Tukey bisquare psi function estimates.

have not delved into what is a robust estimator? This is the subject of the next subsection and next chapter. The above analysis makes use of the uniform convergence over a class of functions, essentially gained through the action of the SLLN at each parameter value.

One of the first and very important results involving uniform convergence is called the Glivenko–Cantelli theorem, which asserts that

$$\mathcal{P}\{\omega | \lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x,\omega) - G(x)| = 0\} = 1. \tag{1.15}$$

Here $F_n(x,\omega)$ is the *empirical distribution function* which is the distribution (random) that assigns atomic mass $1/n$ to each point of the sample $\pi^{(n)}o\boldsymbol{X}(\omega)$. The proof of Eq. (1.15) stems from the SLLN applied to the sequence of indicator function values $\{I_{(-\infty,x]}(X_i)\}_{i=1}^{\infty}$. Extensions of the Glivenko–Cantelli theorem to $\mathcal{E}^k$ space and more general sets have been carried out by Wolfowitz (1954), Rao (1962), Topsoe (1970), and Elker et al. (1979). All of these consider the theorem for closed half spaces in $\mathcal{E}^k$. A more general theory is available in the work of Vapnik and Chervonenkis (2004) which also has applications in machine learning but suffice to say so long as your functions are smooth in some sense as in the form of the Tukey bisquare we do not need the abstractness of the most general theory. See also Pollard (1990) for further results on empirical processes.

Interestingly, the estimator $\hat{\mu}_n$ which is a solution to Eq. (1.7) can be thought of as an implicitly defined estimator that is a function of the empirical distribution function. For example,

$$K_n(\tau) = \int_{-\infty}^{+\infty} \psi(x-\tau)\mathrm{d}F_n(x),$$

so that we can think of $K_n(\tau) \equiv K_{F_n}(\tau)$. Defining the median to be $F_n^{-1}(1/2) = \inf\{u | F_n(u) \geq 1/2\}$ and the functional solution such that if we consider the set of solutions

$$S(\psi_{\mathrm{BS}}, F_n) = \{\tau | K_n(\tau) = 0\} \tag{1.16}$$

then the estimator can be thought of as a function of the empirical distribution function by letting $\rho(F_n, \tau) = |\tau - F_n^{-1}(1/2)|$ and defining the estimator via

$$\inf_{\tau \in S(\psi_{\mathrm{BS}}, F_n)} \rho(F_n, \tau) = \rho(F_n, T[F_n]). \tag{1.17}$$

The estimator $T[F_n]$ is then both a function of the empirical distribution function and the function ψ_{BS} and the "selection functional" ρ. The idea of a selection functional was first coined in Clarke (1983a) and elaborated on in Clarke (1990). Then $T[F_n] \equiv T[\psi_{\mathrm{BS}}, \rho, F_n]$. In the unlikely event that there are two solutions to Eq. (1.17) minimizing the distance from the median, one can choose the one on the left, effectively introducing a third tier of selection, though this is generally not needed. The functional ρ need not be the same as the objective functional used to derive the estimating equations, though this does not exclude such functionals.

In summary we have seen an explicit example illustrating the action of the SLLN. We have quoted from the literature the Glivenko–Cantelli result which yields one avenue regarding the convergence of the empirical distribution, F_n to the underlying distribution which can be any $G \in \mathcal{G}$. For a specific example, we consider corresponding uniform convergence of $K_{F_n}(\tau)$ to $K_\Phi(\tau)$ and also $K'_{F_n}(\tau)$ to $K'_\Phi(\tau)$. For the central asymptotic unique consistent root that results, we find that so long as $K'_\Phi(\mu) \neq 0$, the resulting central root behaves asymptotically normal, with a well-defined asymptotic variance. None of these results are new, but they prove an interesting backdrop to the results that shall later come and demonstrate clearly the connection between empirical convergence concepts and the behavior of some useful estimators. If we were to retain the asymptotic distribution as being normal, the efficient estimator $\overline{X}$ would narrowly beat the estimator that results from employing the Tukey bisquare with tuning constant $c = 4.685$, the asymptotic variance ratio being 0.95. However, as discussed graphically in the works by Tukey (1960) and Huber (1964), the dominance of the classical estimates at the model is soon overshadowed if we consider underlying distributions in ϵ contaminated distributions.

1.5 EFFICIENCY OF LOCATION ESTIMATORS

Based on asymptotic efficiency Tukey (1960) compared the estimator $\overline{X}$ with trimmed means and the median at distributions of the form $G(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/3)$. Here Tukey is intending that ϵ is the average proportion of contaminated data entering the sample with an increased variance of nine. The idea of asymptotic efficiency assumes that the best one can do at a parametric distribution is written down in terms of the inverse of Fisher information. This is encapsulated in the Cramér–Rao inequality. For instance, let $T = T(\pi^{(n)} o \boldsymbol{X}(\omega))$ be an unbiased estimator of μ, meaning $E[T] = \mu$. Here $\pi^{(n)} o \boldsymbol{X}(\omega)$ has a probability

function $f_\mu(\pi^{(n)}o\boldsymbol{X})$. Notably, the support of f does not depend on μ and under mild conditions of regularity which hold here.

$$\text{var}(T) \geq \frac{1}{E\left\{\left(\frac{\partial}{\partial\mu}\log f_\mu(\pi^{(n)}o\boldsymbol{X})\right)^2\right\}} \tag{1.18}$$

The denominator in Eq. (1.18) is denoted the *Fisher information* in $\pi^{(n)}o\boldsymbol{X}$ about μ. Thus, if $\pi^{(n)}o\boldsymbol{X}$ is a random sample from $f_\mu(x)$, the information in $\pi^{(n)}o\boldsymbol{X}$ is n times the information in a single observation:

$$\mathcal{I}(\mu|\pi^{(n)}o\boldsymbol{X}) = n\mathcal{I}(\mu|X)$$

whence Eq. (1.18) can be written as

$$\text{var}(T) \geq \frac{1}{nE\left\{\left(\frac{\partial}{\partial\mu}\log f_\mu(X)\right)^2\right\}}.$$

This leads to a natural definition of efficiency

$$\text{eff}(T) = \frac{1/\mathcal{I}(\mu|\pi^{(n)}o\boldsymbol{X})}{\text{var}(T)}$$

and we say that when an estimator T has efficiency one, then it is 100% efficient. For instance, assuming $f_\mu(x) = \phi(x-\mu)$, it is not hard to see $\mathcal{I}(\mu|X) = 1$ and that the sample mean or average carries efficiency of 100% which is the best possible. Consider the Tukey bisquare estimator, call it T_{BS}, having tuning constant $c = 4.685$. The asymptotic efficiency of this estimator at the standard normal distribution is 95%. Indeed the asymptotic efficiency of the sample median is $2/\pi \simeq 0.637$. The interpretation is that in large samples, using the sample mean requires only about 64% as many observations to estimate μ with a specified precision as would be required using the sample median. But to now compare two estimators, say the mean and the Tukey bisquare, when the underlying parametric distribution is $f_\mu(x) = (1-\epsilon)\phi(x-\mu) + \frac{\epsilon}{3}\phi((x-\mu)/3)$ which is the density of a contaminated normal distribution, we can compare their relative efficiency, which is

$$\frac{\text{eff}T_{\text{BS}}}{\text{eff}\overline{X}} = \frac{\text{var}(\overline{X})}{\text{var}(T_{\text{BS}})}. \tag{1.19}$$

Clearly, this negates the need for evaluating the Fisher information. Also to be realistic we need, say, to decide how many iterations of the Newton–Raphson algorithm we have employed in evaluating the Tukey bisquare estimator starting

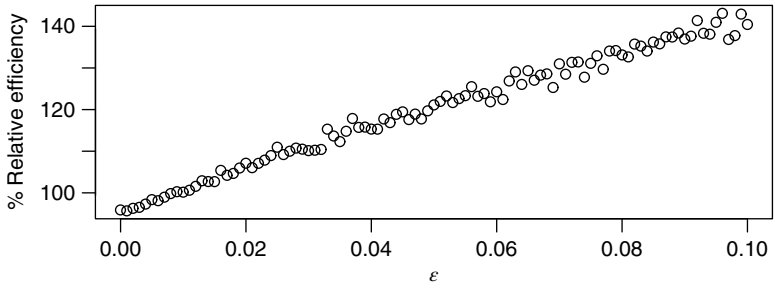

FIGURE 1.9 Estimated relative efficiency of the Tukey bisquare with tuning constant $c = 4.685$ relative to the sample mean based on 10 000 samples of size $n = 100$, with $\nu = 5$ Newton–Raphson iterations beginning from the median.

from the median, for example. We consider the number of Newton–Raphson iterations to be $\nu = 5$. By symmetry we can deduce that both estimators are unbiased for μ. Without loss of generality we can set $\mu = 0$. Since the estimators are unbiased and from any one sample we have an unbiased estimate $\hat{\mu}^2$ of the variance, sampling repeatedly 10 000 samples each of size $n = 100$, an estimate of the variance of the estimator $\hat{\mu}$ is gained from the SLLN, visualize $\hat{\text{var}} = \frac{1}{n}\sum_i \hat{\mu}_i^2$. This is done for epsilon fixed and for both of the estimates respectively from which an estimate of the relative efficiency at that epsilon is afforded. Then one calculates the estimates $\overline{X}$ and T_{BS} with values of epsilon incremented by 0.01 from zero to 0.1, and we obtain empirical measures of relative efficiency as indicated in Figure 1.9. Clearly, as the proportion of contamination increases from zero to 0.1, the efficiency of the Tukey bisquare estimator quickly overtakes that of the sample mean and indeed it appears to be roughly 140% when the proportion of contamination reaches 10%. Ten percent contamination is considered realistic and indeed Hampel et al. (1986) in summarizing their section 1.2c "The Frequency of Gross Errors" write "altogether, 1–10% gross errors in routine data seem to be more the rule rather than the exception."

Huber (1964) considered restricting distributions to be of the form $G(x) = (1 - \epsilon)\Phi(x) + \epsilon H(x)$, where H was only allowed to be a symmetric distribution. Tukey's results show that it is easy to find examples where estimators other than $\overline{X}$ can do better in terms of asymptotic performance at neighboring models to the normal distribution, while Huber includes a minimax solution that describes how to minimize the worst that nature can do to you by his choice of minimax distribution function, albeit a symmetric distribution function.

Much of the discussion in Chapter 2 is devoted to convergence results when G may be in an appropriate neighborhood of a parametric distribution. We explore ideas of convergence in distribution and further relate the generalities associated with convergence of the empirical curve K_{F_n} to K_G.

1.6 ESTIMATION OF LOCATION AND SCALE

Finally in this section we recognize that most practical applications involving M-estimators of location of a normal distribution use an estimate of scale of that distribution. Rarely in estimation do we assume a constant scale for the parametric distribution. The most practical and for that matter easily implementable scale estimate in the scenario where one needs to implement a robust M-estimator, such as the Tukey bisquare, is to choose the median absolute deviation about the median (MAD), defined as

$$\text{MAD}(X_1, \ldots, X_n) = \text{Med}\{|X_i - \text{Med}(X1, \ldots, X_n)|\}.$$

This estimator uses the sample median twice, firstly, to get an estimate of location and then after forming the absolute residuals, finding the median of those absolute residuals. To make MAD consistent for σ, when assuming the data are $N(\mu, \sigma^2)$, we use

$$\text{MADN} = \frac{\text{MAD}}{0.6745}.$$

Then the Tukey bisquare estimate of location is found by replacing Eq. (1.7) by

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{\text{BS}}\left(\frac{X_i - \tau}{\text{MADN}}\right) = 0.$$

Of course the M-estimation approach can also be taken in the estimation of location and scale. The estimating psi-function is a now a vector function $\boldsymbol{\Psi} = (\psi_1, \psi_2)^T$. Here ψ_1 governs the estimator for location and ψ_2 governs the estimator for scale. Both are implemented simultaneously through the vector equation

$$\boldsymbol{K_n}(\tau_1, \tau_2) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}\left(\frac{X_i - \tau_1}{\tau_2}\right) = \boldsymbol{0}. \tag{1.20}$$

A predecessor to the smooth (smooth meaning continuous first derivatives) location estimator of Beaton and Tukey (1974) was the Hampel three-part redescender first discussed in the book by Andrews et al. (1972). See also Hampel (1974). This estimator of location depends on three parameters a, b, c. For instance, the estimating function has the form

$$\psi_{1;a,b,c}(x) = \begin{cases} x & \text{for } 0 \le |x| \le a \\ a\ \text{sign}(x) & \text{for } a \le |x| \le b \\ a\frac{c-|x|}{c-b}\text{sign}(x) & \text{for } b \le |x| \le c \\ 0 & \text{for } c \le |x| \end{cases} \tag{1.21}$$

See the illustration of Figure 1.10.

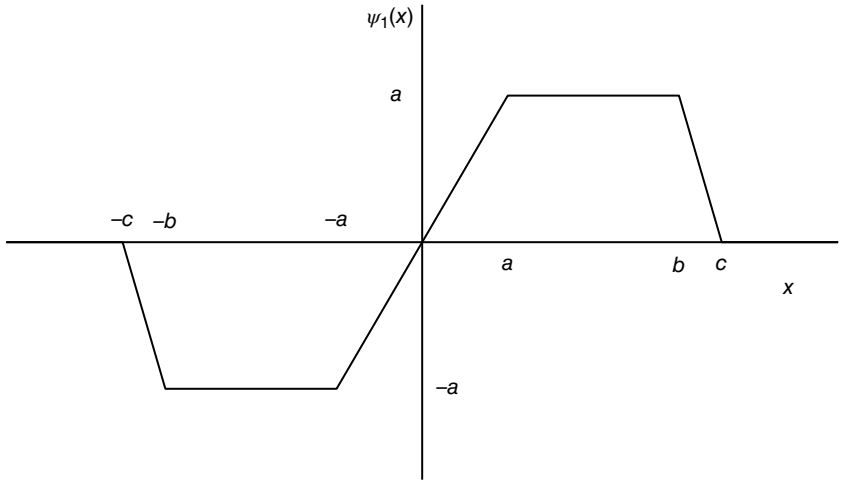

FIGURE 1.10 Plot of Hampel ψ_1 redescender for location.

However, it is also possible to construct an M-estimator of scale, along the lines of a three-part redescender in the following form.

$$\psi_{2;a,b,c}(x) = \begin{cases} x^2 - 1 - p & \text{for } 0 \le |x| \le a \\ a^2 - 1 - p & \text{for } a \le |x| \le b \\ (a^2 - 1 - p)\frac{c-|x|}{c-b} & \text{for } b \le |x| \le c \\ 0 & \text{for } c \le |x| \ , \end{cases} \tag{1.22}$$

where p is defined implicitly from the equation $\int \psi_{2;a,b,c}(x)\mathrm{d}\Phi(x) = 0$. See for example Clarke and Milne (2013). This function is illustrated in Figure 1.11. Then to obtain the estimator one solves Eq. (1.20) with $\Psi = \Psi_{a,b,c} = (\psi_{1;a,b,c}, \psi_{2;a,b,c})^T$.

It can now be noticed that the above functions have sharp corners, and consequently they do not have continuous derivatives. This means that expansions using the mean value theorem as in Eq. (1.11) are not now as straightforward, not to mention the fact that one now has a bivariate set of equations which requires multidimensional calculus. The extra dimension of scale also has to be incorporated. The calculus involved is more difficult, though it can be done using say the nonsmooth analysis discussed in Frank Clarke's book (1983b). This discussion is left till later chapters, but for a reference see Clarke (1986b).

Another important point is that Eq. (1.20) also admits multiple roots. Implementing such an equation requires a further understanding of consistency arguments which we begin in Chapter 2. Since both location and scale are estimated simultaneously, one might define a selection functional that combines both the location choice and a choice for scale. For instance, in estimating

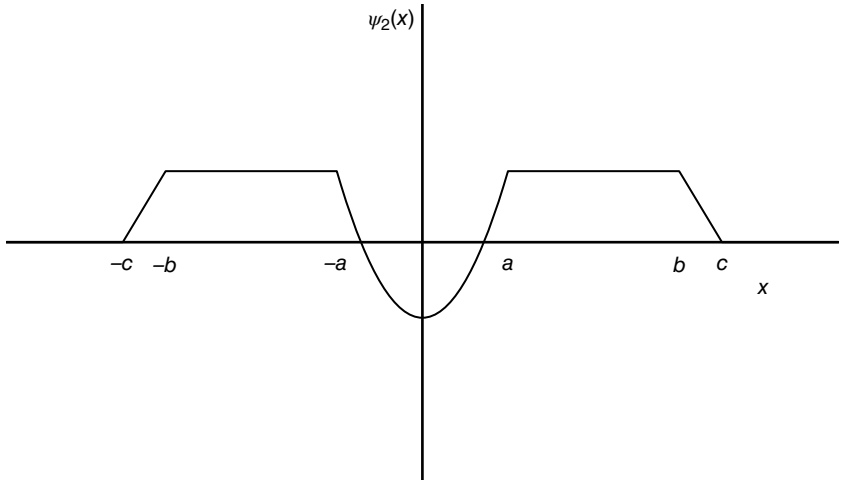

FIGURE 1.11 Plot of ψ_2 redescender for scale.

only scale of a normal distribution, essentially involving the second component of the vector equation (1.20) using the choice (1.22), one can choose to select the scale estimator from possible multiple roots by using $\rho(F_n, \tau_2) = |\{F_n^{-1}(3/4) - F_n^{-1}(1/4)\}/\{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)\} - \tau_2|$. For location and scale parameters estimated simultaneously, a useful selection functional to distinguish the robust estimator from multiple roots of equations (1.20) is for example,

$$\rho(F_n; \tau_1, \tau_2) = (F_n^{-1}(1/2) - \tau_1)^2 + (\{F_n^{-1}(3/4) - F_n^{-1}(1/4)\}/\{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)\} - \tau_2)^2. \tag{1.23}$$

See, for example, section 7 of Clarke (1983a).

This choice of estimator, while in a sense rejects outliers in the tails of the normal distribution, will have to be implemented carefully given possible multiple roots of the equations. An alternative possibility, which has a longer history, is to implement Huber's Proposal 2 estimator [see Huber (1964) and Huber (1981, section 6.4)], where one down-weights but does not exclude outlying values. This can be arrived at by setting $b = c = \infty$ and for purposes of discussion $k = a$. Then the estimator for location and scale looks like $\mathbf{\Psi} = (\psi_1, \psi_2)^T$ where

$$\left.\begin{array}{l} \psi_1(x) = \min(|x|, k)\text{sign}(x) \\ \text{and} \\ \psi_2(x) = \psi_1^2(x) - E_\Phi[\psi_1^2(Z)] \end{array}\right\} \quad -\infty < x < \infty \tag{1.24}$$

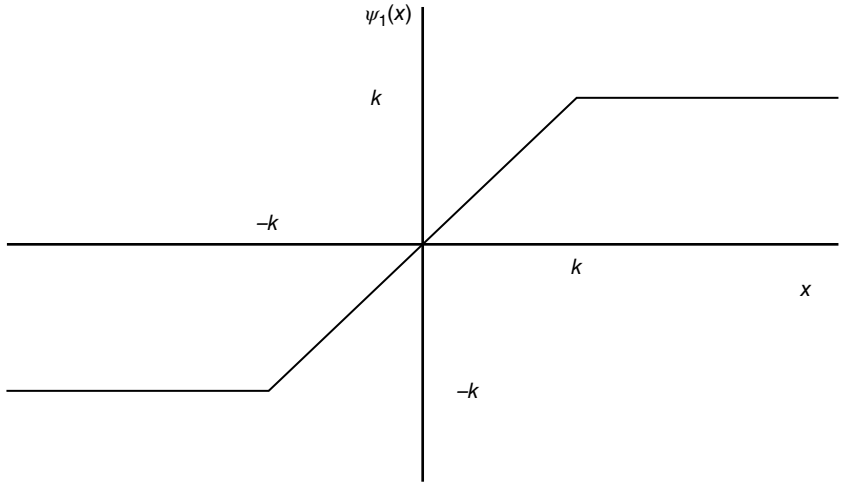

FIGURE 1.12 Plot of Huber's ψ-function for location.

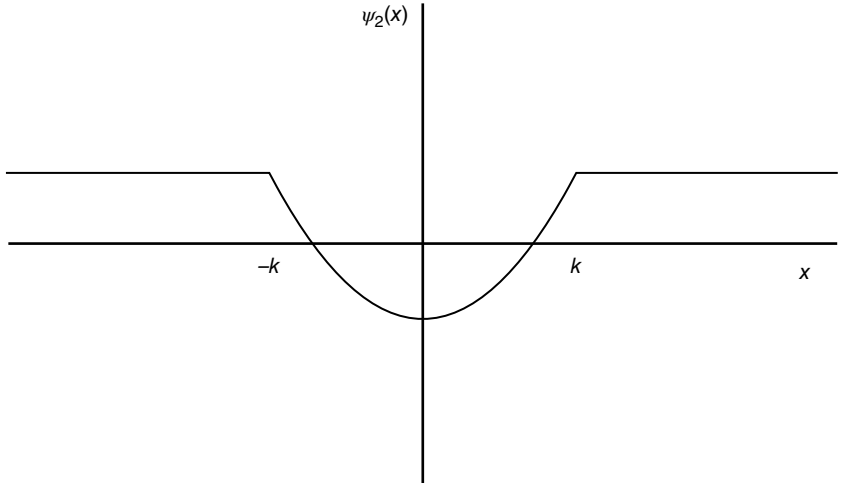

FIGURE 1.13 Plot of Huber's χ-function for estimating scale.

and the graphs of ψ_1 and ψ_2 are in Figures 1.12 and 1.13 respectively. The implementation of Huber's Proposal 2 is discussed in Venables and Ripley (2002) for example. In a novel approach to gaining a redescending estimator for location and scale, where the $\boldsymbol{\Psi}$-function is continuously differentiable for both component functions, Bachmaier (2007) in essence provides the following functions:

$$\psi_1(x) = \begin{cases} -\psi_1(-x) & \text{for} \quad x < 0 \\ x & \text{for } 0.0 \le x \le 0.9 \\ -(x-1.4)^2 + 1.15 & \text{for } 0.9 < x < 1.9 \\ -x + 2.8 & \text{for } 1.9 \le x \le 2.3 \\ 0.5(x-3.3)^2 & \text{for } 2.3 < x < 3.3 \\ 0 & \text{for } 3.3 \le x \end{cases} \tag{1.25}$$

$$\psi_2(x) = \begin{cases} x^2 - 0.7 & \text{for} \quad |x| \le 1.0 \\ -(|x|-2)^2 + 1.3 & \text{for } 1.0 \le |x| \le 3.0 \\ (3.3 - (|x|)^2/0.3 & \text{for } 3.0 \le |x| \le 3.3 \\ 0 & \text{for } 3.3 \le |x| \quad . \end{cases} \tag{1.26}$$

Bachmaier (2007) discusses in detail various ways of arriving at a solution that is consistent. With either choice of redescender involving Hampel's choice and the adapted Hampel redescender for scale, or Bachmaier's choice, when one solves Eq. (1.20) one needs to search for solutions using iterative nonlinear equation solving algorithms, which require initial estimates of location and scale (Bachmaier's functions are illustrated in Figures 1.14 and 1.15). One can use Huber's estimates or the estimates based on nonparametric quantities such as $(\text{Median}, \text{MADN})^T$ for the initial estimates. In small samples, it may be worth searching say starting with a grid of initial estimates and then implementing the selection functional to choose the final estimate from possible multiple solutions. While this entails extra

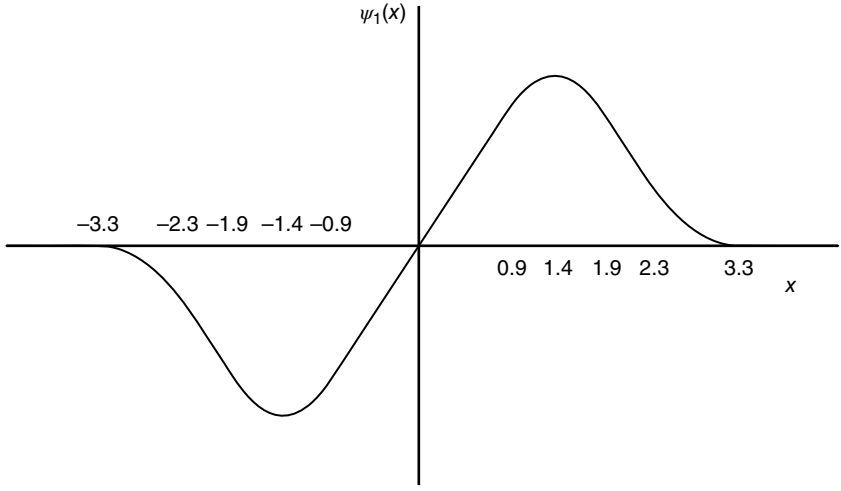

FIGURE 1.14 Plot of Bachmaier's ψ-function for estimating location.

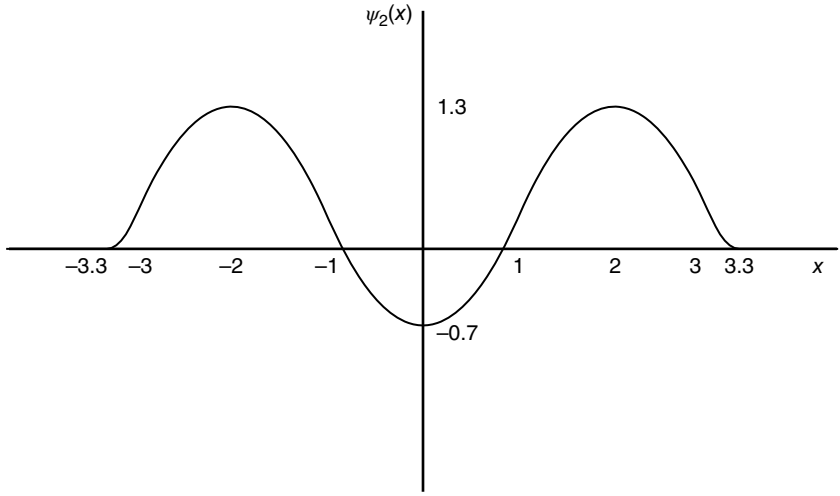

FIGURE 1.15 Plot of Bachmaier's ψ_2-function for estimating scale.

work, it may be philosophically more satisfying in that one gives zero weight to observations that are in the tails of the distribution. Extreme outliers are thus given zero weight in the estimation.

The problems thrown up by the introduction of redescending M-estimates of location and scale are found more widely in parametric estimation and even in the restricted scenario of maximum likelihood estimation. The following are the questions often raised: How do we numerically find a solution? If there are more than one solution to the problem (this may involve more than one solution to the maximizing equations, or indeed in some parametric models different solutions that give equal maxima to the likelihood), how do we distinguish between solutions? In Chapter 2 we give a general theory of consistency, asymptotic normality, and asymptotic efficiency based on a functional approach to estimation.

PROBLEMS

1.1. A collection $\tilde{\mathcal{A}}$ of subsets of Ω is called a σ-field if it satisfies the following conditions:

(i) $\emptyset \in \tilde{\mathcal{A}}$

(ii) if $A_1, A_2, \ldots \in \tilde{\mathcal{A}}$ then $\cup_{i=1}^{\infty} A_i \in \tilde{\mathcal{A}}$

(iii) if $A \in \tilde{\mathcal{A}}$ then $A^c \in \tilde{\mathcal{A}}$, where A^c is the complement of A.

Use elementary set operations to show that if $\tilde{\mathcal{A}}$ is closed under countable intersections: that is, if $A_1, A_2, \ldots$ are in $\tilde{\mathcal{A}}$, then so is $\cap_i A_i$.

1.2. Consider a die that results in six with probability one in six. Consider a sequence of independent tosses of the die where for the ith the random variable X_i is a one if the die is a six and a zero otherwise. The random variables have a common distribution. Show that the mean of this distribution is $\frac{1}{6}$ and conclude what is the limiting value in probability and also almost surely of the sample average $\frac{1}{n}\sum_{i=1}^{n} X_i$ as $n \to \infty$.

1.3. Consider a random variable with the exponential distribution which is defined on the positive real line and has density $f_\theta(x) = \exp(-x/\theta)/\theta$. Find the expectation and variance of this random variable.

1.4. Show that the class of functions $\mathcal{A} = \{\psi_{\mathrm{BS}}(\cdot - \tau) : \tau \in \mathcal{E}\}$ is equicontinuous.

1.5. Consider a sequence of independent coin tosses of a fair coin and let $\overline{X}_n$ the average number of heads. Show that the distribution of the random variable $\sqrt{n}\left(\overline{X}_n - \frac{1}{2}\right)$ converges to a $N\left(0, \frac{1}{4}\right)$ random variable.

1.6. Let $X_1, X_2, \ldots$ be an *i.i.d.* sequence of observations with common distribution F that has a symmetric density f which is located about $\mu = F^{-1}\left(\frac{1}{2}\right)$. Under various regularity conditions the asymptotic variance of the median is

$$\mathrm{var}[\mathrm{Med}(X_1, \ldots, X_n)] \sim \frac{1}{4nf^2\left(F^{-1}\left(\frac{1}{2}\right)\right)}.$$

Show that when f is the normal density with unit scale that the efficiency of the median is $\frac{2}{\pi}$ and this does not depend on μ. Also investigate whether or not the efficiency depends on the scale of the normal density σ say?

1.7. Consider the following data: 1.0, −1.8, −0.5, −1.4, 0.7, 0.9, −0.6, 1.9, −0.1, 3.5. Calculate the median and MADN for these data. Subsequently, write an algorithm using the Newton–Raphson iteration of formula (1.8) by substituting the Tukey bisquare function (1.6) to evaluate the Tukey bisquare estimator when using MADN to estimate the scale.

1.8. For the data of the previous question and using Bachmaier's choice of M-estimator as defined by Eqs. (1.25) and (1.26) use a nonlinear equation solver starting the iteration from (Median, MADN). Investigate whether there are other roots of the equation by forming a grid of initial estimates by choosing the location estimate to vary from −1 to 1 in steps of 0.2 and scale to vary from 0.5 to 2 in steps of 0.5 and iterating to a root. This will involve running the algorithm several times. If there are multiple roots, decide which is the closest to (Median, MADN)?