The Theory of Statistics and Its Applications

By Dennis D. Cox
Rice University

Chapter 3

Asymptotics

In this chapter we undertake the study of asymptotics or large sample theory. Asymptotics plays an important role in Statistical Science, for several reasons. As in applied mathematics, for many complicated problems there are nice asymptotic approximations that are highly accurate and much easier to compute than the exact solution. For instance, "ray theory" is basically an asymptotic solution to a wave equation as the frequency tends to infinity. In Statistical Science, the usual situation is to let the sample size tend to ∞, hence the terminology "large sample theory." An important use of statistical asymptotics is approximate calculations involving distributional properties of random variables, such as probabilities or moments. Also, asymptotics are often useful for statistical inferences. For instance, if one wishes to provide a confidence interval for a population mean from a sample when the population distribution is unknown, then the Central Limit Theorem and the consistency of the sample variance provide a simple approach which works well in many settings when the sample size is large. Finally, it has been common to examine "asymptotic optimality" of procedures. Thus, from a practical point of view, many "real world" applications of statistics involve an appeal to asymptotics.

3.1 Basic Asymptotics

This first section introduces some tools from the analytical (i.e. non-probabilistic) subject of asymptotics. There are many useful ideas and results from this subject which are generally learned in applied mathematics.

Suppose $\{ \ a_n : n \in I\!\!N \ \}$ is a sequence of vectors and $\{ \ b_n : n \in I\!\!N \ \}$ is a sequence of nonnegative real numbers. We write

$$\|a_n\| \lesssim b_n \quad \text{as} \quad n \to \infty$$

or, what is the same,

$$a_n \ = \ O(b_n) \quad \text{as} \quad n \to \infty \quad ,$$

if and only if there is some N and some $C \in (0, \infty)$ such that

$$\|a_n\| \;\leq\; C b_n \quad \text{for all} \quad n \geq N.$$

In words, $a_n = O(b_n)$ iff for all n sufficiently large, $\|a_n\|$ is bounded by a multiple of b_n. One says "a_n is big-oh of b_n," or "a_n is asymptotically dominated by b_n." In general, the notation $O(b_n)$ stands for any sequence $\{ \ a_n \ \}$ satisfying the above *which is not necessarily the same in each appearance*, even within the same expression or equation. This latter convention is the source of much confusion and error. An equivalent condition if $b_n > 0$ for all n is

$$\|a_n\| \lesssim b_n \quad \text{iff} \quad \exists C > 0 \text{ such that } \|a_n\| \;\leq\; C b_n \quad \text{for all} \quad n, \tag{3.1}$$

which avoids mentioning N, and one can avoid mentioning C as well by writing

$$\|a_n\| \lesssim b_n \quad \text{iff} \quad \limsup_n \frac{\|a_n\|}{b_n} \;<\; \infty \quad . \tag{3.2}$$

See Exercise 3.1.1. It is sometimes useful to write $b_n \gtrsim a_n$ to mean $a_n \lesssim b_n$.

We write

$$\|a_n\| \;\ll\; b_n \quad \text{as} \quad n \to \infty$$

or

$$a_n \;=\; o(b_n) \quad \text{as} \quad n \to \infty,$$

if and only if for all $\epsilon > 0$ there is some N such that

$$\|a_n\| \;\leq\; \epsilon b_n \quad \text{for all} \quad n \geq N.$$

An equivalent condition if $b_n > 0$ for all n is

$$\|a_n\| \;\ll\; b_n \quad \text{iff} \quad \lim_{n \to \infty} \frac{\|a_n\|}{b_n} \;=\; 0 \quad . \tag{3.3}$$

See Exercise 3.1.1.

If both a_n and b_n are nonnegative,

$$a_n \lesssim b_n \quad \text{and} \quad b_n \lesssim a_n \quad \text{as} \quad n \to \infty \quad ,$$

then we write

$$a_n \;\asymp\; b_n \quad \text{as } n \to \infty \ .$$

If $b_n > 0$ for all n then

$$a_n \;\asymp\; b_n \quad \text{as} \quad n \to \infty \quad \text{if and only if} \tag{3.4}$$

$$\exists c > 0 \text{ and } C > 0 \text{ such that } c b_n \;\leq\; a_n \;\leq\; C b_n \ , \forall n.$$

See Exercise 3.1.1.

Finally, if $\{a_n\}$ is a positive real sequence and $b_n > 0$ for all n, we write

$$a_n \sim b_n \quad \text{as } n \to \infty$$

iff

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 1 .$$

We will often drop the $n \to \infty$ in such statements, it being understood. Note that

$$a_n \sim b_n \quad \Rightarrow \quad a_n \asymp b_n ,$$

but not conversely. (To see this, note that $a_n \sim b_n$ implies both the $\liminf_n$ and $\limsup_n$ in (3.4) are 1.)

A warning: these notations are not entirely standard. The $O(\cdot)$ and $o(\cdot)$ are common and well standardized, but the notations $\lesssim$ and $\ll$ are somewhat rarer. Many use "$\approx$", "$\simeq$", or "$\cong$" to mean either what we defined as $\asymp$ or $\sim$, and sometimes "$\sim$" means what we defined as $\asymp$. One must clarify the meaning in each case. Our notations are, we believe, the most commonly used. Also, the meaning of $\asymp$ never seems to vary.

These notions are also useful when dealing with functions of a continuous variable. If $f(\underline{x})$ and $g(\underline{x}) \geq 0$ are real valued functions of a vector variable $\underline{x}$ defined for all $\underline{x}$ in some neighborhood of $\underline{x}_0$, then we write

$$|f(\underline{x})| \lesssim g(\underline{x}) \quad \text{as } \underline{x} \to \underline{x}_0$$

or

$$f(\underline{x}) = O(g(\underline{x})) \quad \text{as } \underline{x} \to \underline{x}_0 ,$$

if and only if there is some $\epsilon > 0$ and some $C \in (0, \infty)$ such that

$$|f(\underline{x})| \leq Cg(\underline{x}) \quad \text{for all} \quad \underline{x} \quad \text{satisfying} \quad 0 < \|\underline{x} - \underline{x}_0\| < \epsilon \quad .$$

Again, vector valued functions $\underline{f}$ can be treated by replacing the $|\cdot|$ with $\|\cdot\|$. One can extend these notions to one-sided or infinite limits, i.e. if x is a real variable, $x \to x_0 \pm 0$ or $x_0 = \pm\infty$ (Exercise 3.1.4). Here, we write $x \to x_0 - 0$ to mean x approaches x_0 from the left, and similarly $x \to x_0 + 0$ means x approaches x_0 from the right. Also, the other notions corresponding to $\ll$ (or $o(.)$), $\asymp$, and $\sim$ are defined analogously (Exercise 3.1.5).

Another extension which is frequently useful is to have uniformity in another variable. Suppose $f(\underline{x}, \underline{y})$ and $g(\underline{x}, \underline{y})$ are functions of two vector variables where $\underline{y} \in A$. Then

$$|f(\underline{x}, \underline{y})| \lesssim |g(\underline{x}, \underline{y})| \quad \text{as} \quad \underline{x} \to \underline{x}_0 \quad \text{uniformly} \quad \text{in} \quad \underline{y} \in A$$

iff there is some $\epsilon > 0$ and some $C \in (0, \infty)$ such that

$$\text{for all} \quad \underline{y} \in A, |f(\underline{x}, \underline{y})| \leq C|g(\underline{x}, \underline{y})|$$

$$\text{for all } \underline{x} \quad \text{satisfying} \quad 0 < \|\underline{x} - \underline{x}_0\| < \epsilon \quad .$$

Again, this can be extended to sequences (i.e. x is an $I\!N$ valued variable and $x_0 = \infty$), and so forth.

While these notions and notations are very useful, they are the source of many errors, especially the $O(\cdot)$ notation. Many students think that $a_n = O(b_n)$ means $|a_n| \asymp b_n$, *which is not true.* Note that $a_n = O(b_n)$ is a *one-sided* relation. For instance, $2^{-n} = O(n^{-2})$, but $n^{-2} \neq O(2^{-n})$. Also, just because $a_n = O(b_n)$ and $a_n = O(c_n)$ does not mean $O(b_n) = O(c_n)$! For instance, $2^{-n} = O(n^{-2})$ and $2^{-n} = O(n^{-1})$, but $n^{-1} \neq O(n^{-2})$! One must always be careful to remember what these notations mean. $O(b_n)$ stands for some sequence that is $\lesssim b_n$. The notation $\lesssim$ is preferable in that it is less likely to lead to error (e.g. one isn't likely to think that $a_n \lesssim b_n$ and $a_n \lesssim c_n$ implies $b_n \lesssim c_n$), but the $O(\cdot)$ notation is very useful in so many circumstances and is used so widely that it is worth learning about it and understanding it to avoid the pitfalls. To illustrate the usefulness, by the definition of the derivative we have that for f differentiable at x_0,

$$f(x_0 + h) \ = \ f(x_0) \ + \ hf'(x_0) + o(h) \quad , \quad \text{as} \quad h \to 0.$$

To use the $\ll$ notation, we would have to write something like

$$f(x_0 + h) \ = \ f(x_0) \ + \ hf'(x_0) + R(h) \quad ,$$

$$\text{where} \quad |R(h)| \ll h \quad \text{as} \quad h \to 0.$$

The usual practice is to collect all $O(\cdot)$ and $o(\cdot)$ terms on one side of an equation (usually the r.h.s.) and to simplify using results as in Proposition 3.1.1 below. See Example 3.1.1.

The practical utility of the asymptotics to be presented is not always clear. It will typically be the case that we have an expression such as

$$a_n \ = \ b_n \ + \ O(c_n) \ , \tag{3.5}$$

where a_n is a sequence we are interested in approximating but is very complicated, b_n is the approximating sequence which is easier to evaluate, and $O(c_n)$ is the "remainder" or "error" term. Generally, if this approximation is to be of any use, we need that $c_n = o(\|b_n\|)$, i.e. that the error term tends to be much smaller than our approximating sequence, or put more simply, the relative error of the approximation tends to 0. Even if this happens, we don't know that for a given (finite) value of n that b_n will be a good approximation to a_n, because our definitions all involve some statement about "for all n sufficiently large ...". If one looks through the proof of a theorem whose conclusion is of the form (3.5), one can find values for how large n has to be and the constant C. However, this is often not very productive useful one of the advantages of this whole asymptotical approach is that we can be rather "sloppy" about getting our constants. Thus, in general, if we find from the proof of equation (3.5) that for all $n \geq 1,000,000$

we have $\|a_n - b_n\| \leq 100,000\|c_n\|$, this may be rather discouraging. However, one could hope that in fact for much smaller values of n we have a much better bound. In applications, there is often a straightforward way to proceed: the desired sequence a_n can often be computed accurately for small values of n, and then we can calculate the norm of the error $\|a_n - b_n\|$, plot this vs. $\|c_n\|$, and look for some behavior in the plotted points which indicates a straight line for large enough n. Thus, it is often desirable to combine asymptotic results with computational results for small n to see where the asymptotics begins to "take hold" (in slang, when we have reached "asymptopia") with sufficient accuracy, and then we can use the asymptotic approximation after that point. See deBruijn (??) for further discussion on these practical issues. Example 3.1.1 below gives some indication of these ideas.

We next give some simple properties of these order relations.

Proposition 3.1.1 **(a)** *If $\|c_n\| \lesssim b_n \lesssim a_n$, then $\|c_n\| \lesssim a_n$. This is also written as $O(O(a_n)) = O(a_n)$.*

(b) *If $\|c_n\| \lesssim b_n \ll a_n$, then $\|c_n\| \ll a_n$, which is also written as $O(o(a_n)) = o(a_n)$. Similarly, either $o(O(a_n))$ or $o(o(a_n))$ is $o(a_n)$.*

(c) *$O(a_n)O(b_n) = O(a_n b_n)$.*

(d) *Either $O(a_n)o(b_n)$ or $o(a_n)o(b_n)$ is $o(a_n b_n)$.*

(e) *$O(a_n) + O(a_n) = O(a_n)$.*

(f) *$o(a_n) + o(a_n) = o(a_n)$.*

Partial Proof. Consider part (a). This looks "obvious" when stated in the form $\|c_n\| \lesssim b_n \lesssim a_n \Longrightarrow \|c_n\| \lesssim a_n$ because of the typographical similarity between "$\lesssim$" and "$\leq$", but looks can be deceiving, so one must be able to provide a proof. Now $\|c_n\| \lesssim b_n$ means there is an N and a C such that $\|c_n\| \leq Cb_n$ for all $n \geq N$, and $b_n \lesssim a_n$ means there is a N' and a C' such that $b_n \leq C'a_n$ for all $n \geq N'$. Thus, for all $n \geq \max\{N, N'\}$, $|c_n| \leq CC'a_n$, so we have $c_n \lesssim a_n$ as desired.

We consider the equality of the first and third members in part (d). Suppose $c_n = O(a_n)$, so there is an N and a C such that $\|c_n\| \leq Ca_n$ for all $n \geq N$. Suppose $d_n = o(b_n)$, i.e. given $\epsilon > 0$ there is an N' such that $|d_n| \leq \epsilon b_n$ for all $n \geq N'$. Now let $\epsilon' > 0$ be given. Take $\epsilon = \epsilon'/C$ in the previous sentence, so for $n \geq \max\{N, N'\}$, $|c_n d_n| \leq (Ca_n)(\epsilon b_n) = \epsilon'(a_n b_n)$, which shows that $c_n d_n = o(a_n b_n)$ as desired.

Next, consider the first and third members of (e). Suppose $b_n = O(a_n)$ i.e. there exist N and C such that for all $n \geq N$ we have $\|b_n\| \leq Ca_n$. Similarly, for another sequence $c_n = O(a_n)$, there exists C' and N' such that for all $n \geq N'$,

$\|c_n\| \leq C'a_n$. Then for all $n \geq \max\{N, N'\}$, $\|b_n + c_n\| \leq \|b_n\| + \|c_n\| \leq Ca_n + C'a_n = (C + C')a_n$, which shows $b_n + c_n = O(a_n)$.

Finally, consider (f). If $b_n = o(a_n)$, then given $\epsilon > 0$ there exist N such that for all $n \geq N$, $\|b_n\| < (\epsilon/2)a_n$. And for the other term, if $c_n = o(a_n)$ then there exists N' such that for all $n \geq N'$, $\|c_n\| < (\epsilon/2)a_n$. Then for all $n \geq \max\{N, N'\}$, $\|b_n + c_n\| \leq \|b_n\| + \|c_n\| < (\epsilon/2 + \epsilon/2)a_n = \epsilon a_n$, which shows the desired result.

The rest of the proof is left to the student (Exercise 3.1.6).

□

Before looking at an example, we should point out that if one blindly manipulates expressions and equalities involving the $O(\cdot)$ and $o(\cdot)$, one is bound to make errors. Equations involving these seem to defy many of the laws of algebra. For instance, when we write $O(a_n) + O(a_n) = O(a_n)$, we can't automatically subtract $O(a_n)$ from both sides and conclude $O(a_n) = 0$.

Example 3.1.1 Consider the approximation of the tails of the standard normal distribution, i.e.

$$\begin{aligned} 1 - \Phi(z) &= \int_z^\infty \phi(x)\,dx \qquad (3.6) \\ &= \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-x^2/2}\,dx \; . \end{aligned}$$

Of course, ϕ is the $N(0,1)$ density. Throughout this example, we will let $z \to \infty$ in all asymptotic formulae. Many asymptotic expansions for integrals are obtained through integration by parts, a useful trick to remember. If we apply this to (3.6) with

$$u = \frac{-1}{x}\;, \quad dv = -xe^{-x^2/2}\,dx$$

$$du = \frac{1}{x^2}\,dx\;, \quad v = e^{-x^2/2}$$

then

$$\int u\,dv = uv - \int v\,du$$

gives

$$\begin{aligned} 1 - \Phi(z) &= \frac{1}{\sqrt{2\pi}} \left\{ \left[\frac{-1}{x} e^{-x^2/2}\right]_{x=z}^{x=\infty} - \int_z^\infty \frac{1}{x^2}\, e^{-x^2/2}\,dx \right\} \\ &= \frac{\phi(z)}{z} - \int_z^\infty \frac{\phi(x)}{x^2}\,dx \quad . \end{aligned}$$

Now for $x \geq z > 0$, $1/x^2 \leq 1/z^2$ so

$$0 < \int_z^\infty \frac{\phi(x)}{x^2}\,dx < \frac{1}{z^2}\int_z^\infty \phi(x)\,dx = [1 - \Phi(z)]\,O\left(\frac{1}{z^2}\right) \quad .$$

Plugging this into the previous display gives

$$1 - \Phi(z) = \frac{\phi(z)}{z} + [1 - \Phi(z)]\, O\left(\frac{1}{z^2}\right)$$

and hence

$$[1 - \Phi(z)]\left[1 + O\left(\frac{1}{z^2}\right)\right] = \frac{\phi(z)}{z} \quad . \tag{3.7}$$

It is preferable that the order bounds end up on the r.h.s. with the approximant, and this is no problem since if $r(z) = o(1)$ as $z \to \infty$ (which just means $r(z) \to 0$), we have

$$\frac{1}{1 + r(z)} = 1 - \frac{r(z)}{1 + r(z)} = 1 - O(r(z)) = 1 + O(r(z)),$$

as the denominator $[1 + r(z)] \to 1$. Thus $[1 + O(1/z^2)]^{-1} = 1 + O(1/z^2)$, and so dividing both sides of (3.7) by $1 + O(1/z^2)$ gives

$$1 - \Phi(z) = \frac{\phi(z)}{z}\left[1 + O\left(\frac{1}{z^2}\right)\right] \quad . \tag{3.8}$$

Now we consider the practical utility of this result. Figure 1.1 shows a plot of the ratio of the computed error bound to the order estimate as a function of z. Here, this ratio is defined as

$$\frac{(1 - \Phi(z)) - \phi(z)/z}{\phi(z)/z^3} \quad .$$

The *S* statistical package was used for this calculation. We see that this function decreases down to near -1 at about $z = 7$ and then begins to oscillate erratically (actually, for $z > 8$ the erratic oscillations become quite extreme). These oscillations result from roundoff error in the computation of $\Phi(z)$. Indeed, for $z > 7$ an extension of this asymptotic approximation given in Exercise 3.1.13 would probably give more accurate results than the computer package. From this exercise it also follows that

$$\frac{(1 - \Phi(z)) - \phi(z)/z}{\phi(z)/z^3} \sim -1 \quad . \tag{3.9}$$

Until the numerical error causes the computed quantity to become unstable, this is more or less the behavior we see.

Figure 1.2 shows the relative error in the approximation of of $1 - \Phi(z)$ by $\phi(z)/z$. The relative error is defined to be

$$\text{Relative Error} = \frac{(1 - \Phi(z)) - \phi(z)/z}{(1 - \Phi(z))},$$

i.e., the error divided by the quantity being approximated.
From this plot, we conclude that the relative error for $z \geq 3.0$ is $\leq 10\%$, approximately. Since $1 - \Phi(z)$ is small, this is not necessarily so bad.

Figure 1.1: Ratio of Computed Error Bound to Order Estimate

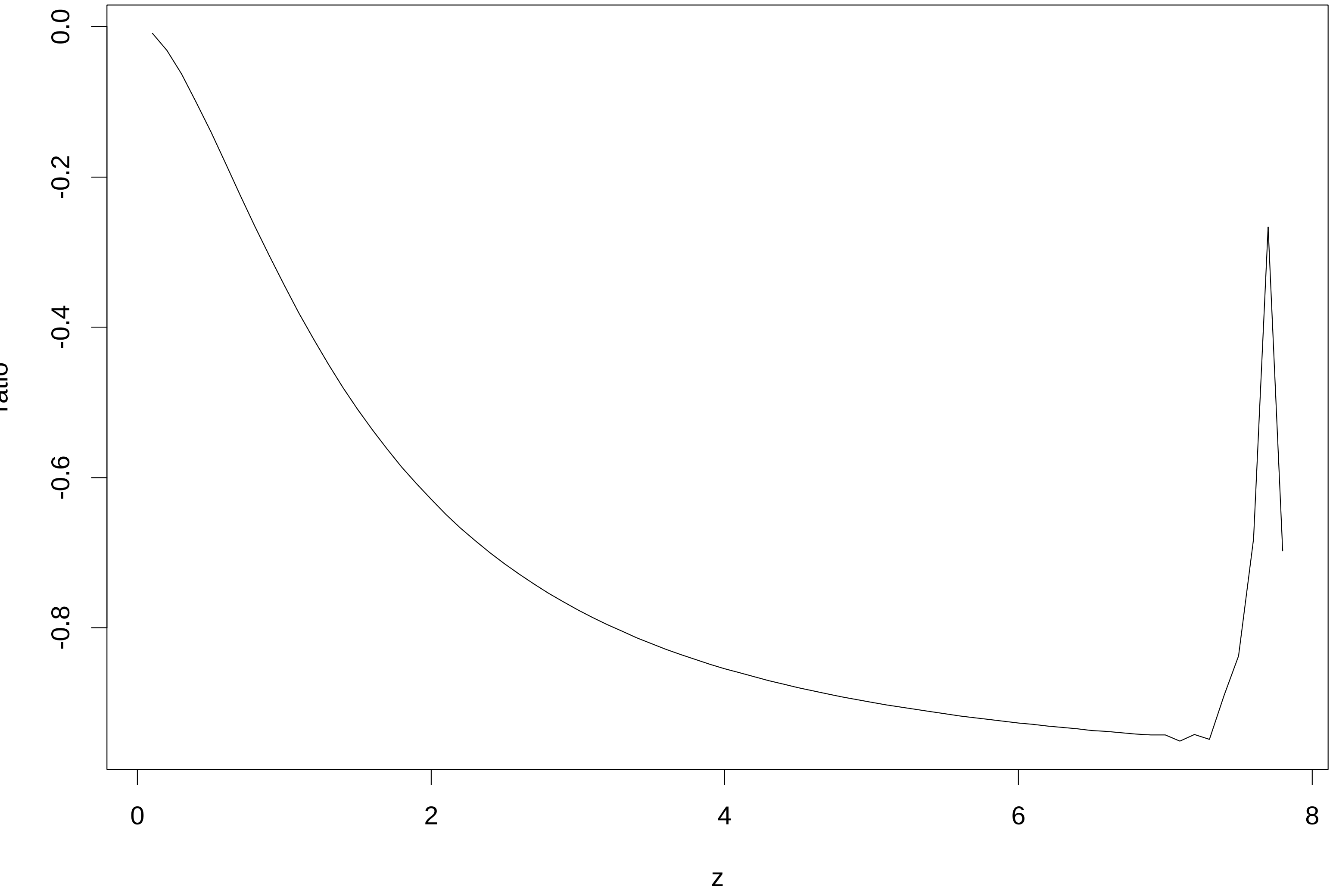
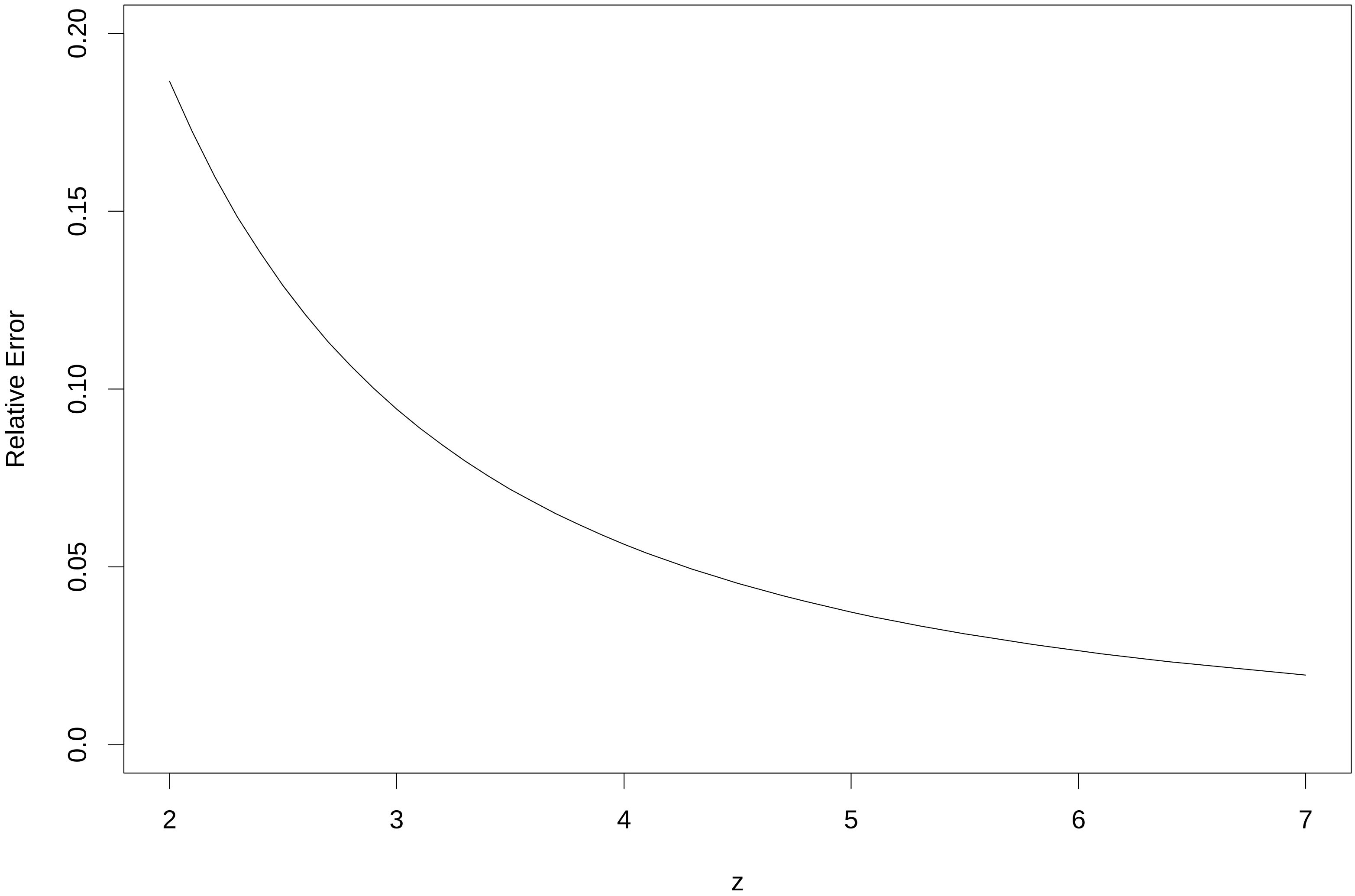


Figure 1.2: Plot of Relative Error in Normal Tail Approximation



relative error at z=7 is about 2%

□

In the next Proposition, we concentrate on the $O(\cdot)$, but analogous results hold for the $o(\cdot)$.

Proposition 3.1.2 **(a)** *If $a_n \lesssim b_n$ are nonnegative sequences, then $a_n^p \lesssim b_n^p$ for any $p > 0$.*

(b) *Let μ be a measure on Ω and suppose there is a μ-null set N such that $|f_n(x)| \lesssim |g_n(x)|$ as $n \to \infty$ uniformly in $x \in \Omega - N$. Then*

$$\left| \int_\Omega f_n(x)\, d\mu(x) \right| \lesssim \int_\Omega |g_n(x)|\, d\mu(x) \quad .$$

Partial Proof. Part (a) is left to the student (Exercise 3.1.8). Consider part (b). Of course, deleting the μ-null does not affect any of the integrals. We know that there exists an n_0 and a constant C such that for all $n \geq n_0$ and all $x \in \Omega - N$,

$$|f_n(x)| \leq C|g_n(x)| \quad \text{for all} \quad x \in \Omega - N \quad .$$

Recall that $|\int f_n\, d\mu| \leq \int |f_n|\, d\mu$, so integrating both sides of the last display gives

$$\left| \int f_n\, d\mu \right| \leq C \int |g_n(x)|\, d\mu$$

and this holds for all $n \geq n_0$. This is the desired result.

□

We illustrate the utility of these ideas with some applications in probability and statistics. The first argument is "combinatorial" in nature, whereas the second uses Taylor series, which is very common in asymptotics.

Theorem 3.1.3 *Suppose X, X_1, X_2, ..., are i.i.d. random variables with $E[X] = 0$. Let $S_n = \sum_{i=1}^n X_i$. If k is an integer ≥ 2 such that $E[|X|^k] < \infty$, then*

$$E[S_n^k] = O(n^{\lfloor k/2 \rfloor}) \quad ,$$

where $\lfloor k/2 \rfloor$ is the largest integer $\leq k/2$.

Proof. Expanding the product of a sum, we have

$$\begin{aligned} E[S_n^k] &= E\left[\prod_{j=1}^k \left(\sum_{i_j=1}^n X_{i_j} \right) \right] \\ &= \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_k=1}^n E\left[\prod_{j=1}^k X_{i_j} \right] \\ &= \sum_{\underline{i} \in \{1,2,\ldots,n\}^k} E\left[\prod_{j=1}^k X_{i_j} \right] \quad , \end{aligned}$$

where the summation in the last expression is over k-dimensional vectors $\underline{i}$ with positive integer components which are $\leq n$. Applying the triangle inequality, we have

$$|\, E[S_n^k] \,| \;\leq\; \sum_{\underline{i}\in\{1,2,\ldots,n\}^k} \;\Big|\; E\left[\prod_{j=1}^{k} X_{i_j}\right] \Big| \quad . \tag{3.10}$$

In the summand,

$$\Big|\; E\left[\prod_{j=1}^{k} X_{i_j}\right] \Big| \quad ,$$

the order of the factors in the product is irrelevant. Thus, for each summation multi-index $\underline{i}$, there is a corresponding multi-index $\underline{m}$ such that the components of $\underline{m}$ are obtained by permuting those of $\underline{i}$ so that they are in increasing order, i.e. there is a permutation π of $\{1, 2, \ldots, k\}$ such that

$$m_j \;=\; i_{\pi(j)} \;, \quad \text{for all} \quad j = 1, 2, \ldots, k \;\;,$$

and

$$m_1 \;\leq\; m_2 \;\leq\; \ldots \;\leq\; m_k \quad .$$

Then

$$\Big|\; E\left[\prod_{j=1}^{k} X_{i_j}\right] \Big| \;\;=\; \Big|\; E\left[\prod_{j=1}^{k} X_{m_j}\right] \Big| \quad .$$

Let $\mathbf{Sort}(\underline{x})$ be the vector with the same components as $\underline{x}$ but permuted in increasing order, so $\underline{m} = \mathbf{Sort}(\underline{i})$.

Now each such (unordered) multi-index $\underline{i}$ corresponds to a unique (ordered) $\underline{m}$, but each $\underline{m}$ may be obtained from several $\underline{i}$. The number of $\underline{i}$'s giving rise to the same $\underline{m}$ is at most $k!$, i.e. $\#\{\underline{i} : \mathbf{Sort}(\underline{i}) = \underline{m}\} \leq k!$. (It is exactly $k!$ if all components of $\underline{m}$ are distinct, but if there are ties in the components of $\underline{m}$, then the number of such $\underline{i}$'s whose ordered components result in the given $\underline{m}$ is $< k!$. For instance, if $\underline{m}$ has one tie between two components (so there are $k-1$ distinct values in the k components of $\underline{m}$, then there are $k!/2!$ unordered multi-indices $\underline{i}$ such that $\mathbf{Sort}(\underline{i})$ equals the given $\underline{m}$.)

Let M denote the collection of all possible values of $\underline{m}$, i.e.

$$M \;=\; \mathbf{Sort}\,\Big(\{1, 2, \ldots, n\}^k\Big) \quad .$$

If we break up the summation on the r.h.s. of (3.10) into the multi-indices $\underline{i}$ corresponding to a given $\underline{m}$, we obtain

$$\begin{aligned} &\sum_{\underline{i}\in\{1,2,\ldots,n\}^k} \;\Big|\; E\left[\prod_{j=1}^{k} X_{i_j}\right] \Big| \\ &\quad = \;\sum_{\underline{m}\in M}\; \sum_{\underline{i}:\mathbf{Sort}(\underline{i})=\underline{m}} \Big|\; E\left[\prod_{j=1}^{k} X_{m_j}\right] \Big| \end{aligned}$$

$$= \sum_{\underline{m} \in M} \#\{\underline{i} : \mathbf{Sort}(\underline{i}) = \underline{m}\} \left| \; E\left[\prod_{j=1}^{k} X_{m_j}\right] \right|$$

$$\leq k! \sum_{\underline{m} \in M} \left| \; E\left[\prod_{j=1}^{k} X_{m_j}\right] \right| \quad . \tag{3.11}$$

Now recall that the X_i's are i.i.d. with mean 0, so in the last summation, if any of the components of a multi-index $\underline{m}$ are unique (i.e. if $m_{j_0} \neq m_j$ for $j \neq j_0$), then the corresponding summand is 0 by using the expected value of a product of independent random variables is the product of the expectations. For example, suppose $k = 6$ and $\underline{m} = (2, 2, 4, 8, 8, 8)$, then

$$E\left[\prod_{j=1}^{6} X_{m_j}\right] = E[X_2 X_2 X_4 X_8 X_8 X_8]$$

$$= E[X_2^2] \times E[X_4] \times E[X_8^3] = E[X^2] \times 0 \times E[X^3] = 0 \quad .$$

In general, we may state that

$$\text{if} \quad m_j - 1 < m_j < m_j + 1 , \quad \text{then} \tag{3.12}$$

$$E\left[\prod_{j=1}^{k} X_{m_j}\right] = 0 \quad .$$

The same conclusion holds if $m_1 < m_2$, or $m_k - 1 < m_k$. Thus, in the last expression in (3.11), we can delete all summands in which one of the components of the ordered multi-index $\underline{m}$ is unique (i.e. not tied with another component). This is the driving force behind the proof, plus the fact that we can obtain a simple bound on the summands in (3.11) which are not 0 by this trick.

Now let us count the number of summands left over after applying the observation of the previous paragraph. If an ordered multi-index $\underline{m}$ of dimension k has no untied components, then there are at most $\lfloor k/2 \rfloor$ distinct values among the components of $\underline{m}$. This is because each component must be tied with at least one other component, and if k is odd, there is at least one three way tie. For instance, if $k = 7$, one possibility for the ties in $\underline{m}$ is

$$m_1 = m_2 < m_3 = m_4 < m_5 = m_6 = m_7 \quad , \tag{3.13}$$

so there are only $3 = \lfloor 7/2 \rfloor$ distinct values among the components of such an $\underline{m}$, namely m_1, m_3, and m_5. Let M_0 denote the subset of M consisting of ordered multi-indices of dimension k in which there are no untied components, i.e.

$$M_0 = \{ \underline{m} : m_1 \leq m_2 \leq ... \leq m_k - 1 \leq m_k ,$$

$$\text{and either } m_j = m_j + 1 \quad \text{or} \quad m_j = m_j - 1 \quad \text{for all} \quad j \} \quad .$$

We claim that

$$\#(M_0) \;\le\; C_0 n^{\lfloor k/2 \rfloor} \quad . \tag{3.14}$$

where C_0 is a finite constant which depends on k but not on n. (This is important because the C in the definition of $O(\cdot)$ cannot depend on n, of course.) To see (3.14), note that we may choose the first of the distinct values in n ways from the set $\{1,\ 2,\ ...,\ n\}$, then the second of the distinct values in $n-1$ ways, and so forth down to the last, say the R'th, of the distinct values in $(n-R+1)$ ways. Thus, if there are $R \le \lfloor k/2 \rfloor$ distinct values, the number of ways may be chosen is at most $\prod_{r=1}^{R}(n-r+1) \le n^R \le n^{\lfloor k/2 \rfloor}$. Thus, there are at most $n^{\lfloor k/2 \rfloor}$ ways of choosing the distinct values for the components of such an $\underline{m}$. Given such a selection for the distinct values, they can be rearranged into the ordered components of $\underline{m}$ in some number of ways which depends only on the number of components k, which gives rise to the constant C_0. Note that we may be "lazy" here and not count the number of such rearrangements.

Now after we delete summands from the last expression in (3.11) using the observation in (3.12), we may rewrite any expectation from any remaining summands in the last expression in (3.11) in the form

$$E\left[\prod_{j=1}^{k} X_{m_j}\right] \;=\; \prod_{r=1}^{R} E[X^{p_r}] \quad , \tag{3.15}$$

where $R \le \lfloor k/2 \rfloor$ by our remarks above, and the p_r are integers satisfying

$$2\ \le p_r\ ,\ \text{for all } r = 1, 2, ..., R\ ,\ \text{and} \tag{3.16}$$

$$\sum_{r=1}^{R} p_r \;=\; k \quad .$$

For example, if $k = 7$ and $\underline{m}$ is as in (3.13), then

$$E\left[\prod_{j=1}^{7} X_{m_j}\right] \;=\; E[X^2] \times E[X^2] \times E[X^3]\ ,$$

where $R = 3$, $p_1 = 2$ (because $m_1 = m_2$), $p_2 = 2$ (because $m_3 = m_4$), and $p_3 = 3$ (because $m_5 = m_6 = m_7$). Now since $p_r \le k$ by (3.16),

$$\Big|\ E[X^{p_r}]\Big| \quad\le\; \max_{2 \le j \le k} E[|X|^j] \;=\; C_1 \quad .$$

Note that our assumption that $E[|X|^k] < \infty$ guarantees that $C_1 < \infty$. Further,

$$\prod_{r=1}^{R} E[X^{p_r}] \;\le\; C_1^R \;\le\; (\,\max\{1, C_1\}\,)^R \tag{3.17}$$

$$\le\; (\ \max\{1, C_1\}\ )^{\lfloor k/2 \rfloor} \;=\; C_2 \quad .$$

Now we are ready to pull the details together and complete the proof. Let M_0 denote the subset of M consisting of ordered multi-indices of dimension k in which there is no component which is untied with some other component. Then collecting together (3.10), (3.11), (3.12), (3.14), and (3.17), we have

$$\begin{aligned}
|\, E[S_n^k] \,| &\leq k! \sum_{\underline{m} \in M} \Big|\, E\left[\prod_{j=1}^{k} X_{m_j} \right] \Big| \\
&= k! \sum_{\underline{m} \in M_0} \Big|\, E\left[\prod_{j=1}^{k} X_{m_j} \right] \Big| \\
&\leq k!\, \#(M_0) \max_{\underline{m} \in M_0} \Big|\, E\left[\prod_{j=1}^{k} X_{m_j} \right] \Big| \\
&\leq k!\, (C_0 n^{\lfloor k/2 \rfloor})\, C_2 \\
&= C n^{\lfloor k/2 \rfloor} \quad ,
\end{aligned}$$

where the constant C doesn't depend on n. This completes the proof.

□

Theorem 3.1.4 *Suppose X, X_1, X_2, ..., are i.i.d. random variables with $E[X^4] < \infty$. Let $E[X] = \mu$ and $Var[X] = \sigma^2$. Suppose $h : A \longrightarrow I\!R$ where A is an interval such that $X \in A$ a.s., and suppose that all derivatives of h up to order 4 exist and the fourth order derivative is bounded in A. Then denoting the sample average of the first n of the X_i's by $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$, we have*

$$E[h(\overline{X}_n)] = h(\mu) + \frac{1}{2n} \sigma^2 D^2 h(\mu) + O(\frac{1}{n^2}) \quad .$$

If the fourth order derivative of h^2 is also bounded in A, then

$$Var[h(\overline{X}_n)] = \frac{1}{n} \sigma^2 (Dh(\mu))^2 + O(\frac{1}{n^2}) \quad .$$

Proof. By Taylor series expansion of $h(x)$ about $x = \mu$ out to the terms of order 4,

$$h(\overline{X}_n) = h(\mu) + (\overline{X}_n - \mu) Dh(\mu) + \frac{1}{2} (\overline{X}_n - \mu)^2 D^2 h(\mu) \tag{3.18}$$

$$+ \frac{1}{3!} (\overline{X}_n - \mu)^3 D^3 h(\mu) + \frac{1}{4!} (\overline{X}_n - \mu)^4 D^4 h(Y_n)$$

where Y_n is between X_n and μ. We are using Lagrange's form of the remainder term here. Also, note that Y_n is a random variable (it depends on $\overline{X}_n$). Taking expectations, we see that the linear term in the Taylor expansion disappears since $E[\overline{X}_n] = \mu$. Also, for the quadratic term we have $E[(\overline{X}_n - \mu)^2] = \sigma^2/n$, and for the cubic term, $E[(\overline{X}_n - \mu)^3] = O(1/n^2)$ by Theorem 3.1.3 (see Exercise 3.1.9(b)). Finally, for the remainder term, we have

$$\begin{aligned} \Big|\ E\left[(\overline{X}_n - \mu)^4 D^4 h(Y_n)\right]\Big| \ &\leq\ E\left[|\overline{X}_n - \mu|^4 |D^4 h(Y_n)|\right] \\ \leq\ E\left[|\overline{X}_n - \mu|^4\right] \sup_{y\in A}|D^4 h(y)| \ &=\ E\left[(\overline{X}_n - \mu)^4\right] \sup_{y\in A}|D^4 h(y)| \\ &=\ O(1/n^2) \sup_{y\in A}|D^4 h(y)| \ =\ O(1/n^2) \quad , \end{aligned} \tag{3.19}$$

where we used $E[(\overline{X}_n - \mu)^4] = O(1/n^2)$ which follows from Theorem 3.1.3. Note that since A is a convex set, it contains both $\overline{X}_n$ and μ, and hence also Y_n. Thus, from our remarks and (3.19), when we take expectations of both sides of (3.18) we obtain

$$E[h(\overline{X}_n)] \ =\ h(\mu)\ +\ \frac{1}{2n}\sigma^2 D^2 h(\mu)\ +\ O(\frac{1}{n^2})\ +\ O(\frac{1}{n^2}) \quad ,$$

where the first $O(1/n^2)$ term comes from $E[(\overline{X}_n - \mu)^3] = O(1/n^2)$ and the second comes from (3.19). The result then follows from Proposition 3.1.1 (e).

The second statement follows by the following steps: (i) compute an asymptotic formula for $E[h(\overline{X}_n)^2]$ using the first part; (ii) plug in the asymptotic formula for $(E[h(\overline{X}_n)])^2$ from the first part; and (iii) simplify using the results of Proposition 3.1.1. The details are left as Exercise 3.1.15.

□

The foregoing result will be useful for evaluating the properties of $h(\overline{X}_n)$ as an estimator of $h(\mu)$.

Exercises for Section 3.1.

3.1.1 (a) Verify the claims about equations (3.1), (3.2), (3.3), and (3.4).

(b) Show that (3.2) and (3.3), are valid without the requirement $b_n \neq 0$ for all n provided we adopt the convention $0/0 = 0$.

3.1.2 Assume $b_n > 0$ for all n. Show that $a_n = O(b_n)$ if and only if there is a positive finite constant C such that $\|a_n\| \leq C\|b_n\|$ for all n.

3.1.3 Suppose

$$b_n = \begin{cases} 0 & \text{if } n \leq 10 \text{ or } n \text{ is even,} \\ n^{-1} & \text{otherwise.} \end{cases}$$

Determine if the following sequences are $O(b_n)$.

(i) $a_n = 2^{-n}$.

(ii) $a_n = (1 - (-1)^n)n^{-1}$.

3.1.4 Define $f(x) = O(g(x))$ as $x \to \infty$.

3.1.5 Define the analogues of $o(\cdot)$, $\asymp$, and $\sim$ for vector valued functions of a vector variable. State and prove analogues of equations (3.2), (3.3), and (3.4) for this setup as well.

3.1.6 Complete the proof of all parts of Proposition 3.1.1.

3.1.7 For each of the following statements, determine whether it is true or false and justify your answer.

(a) $O(a_n) - O(a_n) = 0$.

(b) If $a_n = b_n + o(b_n)$, and if $b_n > 0$, then $a_n \sim b_n$.

(c) If $a_n \asymp b_n$ and $b_n \asymp c_n$, then $a_n \asymp c_n$.

(d) The previous claim remains true if $\asymp$ is replaced by $\sim$ throughout.

(e) If $a_n \asymp b_n$ and $c_n \asymp d_n$, then $a_n + c_n \asymp b_n + d_n$.

(f) $x^{-1} = O(1)$ as $x \to \infty$.

(g) $1 = O(x)$ as $x \to \infty$.

(h) $x^{-1} = o(1)$ as $x \to \infty$.

(i) $x^{-1} = O(1)$ as $x \to 0$.

(j) $1 = O(x^{-1})$ as $x \to 0$.

(k) Suppose for each fixed $n = 1, 2, ...,$ we have $f_n(x) \asymp g_n(x)$ as $x \to x_0$. Then $\sum_{n=1}^{\infty} f_n(x) \asymp \sum_{n=1}^{\infty} g_n(x)$ as $x \to x_0$.

(l) Suppose for each fixed $n = 1, 2, \ldots, N$, we have $f_n(x) \asymp g_n(x)$ as $x \to x_0$. Then $\sum_{n=1}^{N} f_n(x) \asymp \sum_{n=1}^{N} g_n(x)$ as $x \to x_0$.

(m) If $a_n > 0$ and $b_n = o(a_n)$, then $\sqrt{a_n + b_n} = \sqrt{a_n}[1 + b_n/(2a_n) + o(b_n/a_n)]$.

3.1.8 (a) Prove Proposition 3.1.2 (a).

(b) State and prove the analogues of both parts of Proposition 3.1.2 for $\ll$ (or $o(\cdot)$).

3.1.9 (a) Verify that

$$\left(\sum_{i=1}^{n} x_i\right)^k = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_k=1}^{n}\prod_{j=1}^{k} x_{i_j} \quad .$$

(b) Let $\overline{X}_n = n^{-1}S_n$ denote the mean of a sample of size n. Under the same assumptions as in Theorem 3.1.3, show that $E[\overline{X}_n^k] = O(n^{\lfloor k/2\rfloor - k})$. Show in particular that $E[\overline{X}_n^3] = O(n^{-1})$ and $E[\overline{X}_n^4] = O(n^{-1})$.

(c) Show in Theorem 3.1.3 that if we drop the hypothesis that $E[X] = 0$, then the best result we can obtain is $E[S_n^k] = O(n^k)$.

3.1.10 Suppose $f(x) \geq 0$ satisfies $f(x) = O(e^{-ax})$ as $x \to \infty$ for some $a > 0$. Show that

$$\int_z^\infty f(x)\,dx \lesssim e^{-az} \text{ as } z \to \infty \quad .$$

3.1.11 The results of Example 3.1.1 can be used to derive an asymptotic expression for extreme quantiles of the $N(0,1)$ distribution. For $0 < \alpha < 1$ define the upper α quantile

$$z_\alpha = \Phi^{-1}(1-\alpha) \quad .$$

Derive that

$$z_\alpha \sim \sqrt{-2\log\alpha} \text{ as } \alpha \to 0 \quad .$$

Hints: Formally solving for z_α from (3.8) gives

$$z_\alpha^2 = -2\log\alpha - 2\{\log z_\alpha - \log(\sqrt{2\pi}) + \log[1 + O(1/z_\alpha^2)]\} \quad .$$

It is easy to verify that $z_\alpha \to \infty$ as $\alpha \to 0$, so $\log z_\alpha = o(z_\alpha^2)$ as $\alpha \to 0$.

3.1.12 Give an asymptotic approximation similar to the one in Example 3.1.1 but for the right tail of the Gamma distribution, i.e. for

$$\int_z^\infty \frac{x^{\alpha-1}}{\Gamma(\alpha)} e^{-x}\,dx \text{ as } z \to \infty \quad .$$

3.1.13 (a) Show that for any integer $m \geq 0$ and any real $z > 0$,

$$1 - \Phi(z) = \frac{\phi(z)}{z}\sum_{k=0}^{m}\frac{(-1)^k\,(2k)!}{2^k(k!)\,z^{2k}} + (-1)^{(m+1)}\frac{(2m+2)!}{2^{m+1}(m+1)!}R_m(z) \quad ,$$

where

$$R_m(z) \;=\; \int_z^\infty \frac{\phi(x)}{x^{2m+2}}\, dx$$

(b) Show that

$$R_m(z) \lesssim \frac{\phi(z)}{z^{2m+2}}\,, \quad \text{as} \quad z \to \infty\,.$$

(c) Verify equation (3.9).

(d) Show that

$$R_m(z) \;\sim\; \frac{\phi(z)}{z^{2m+3}}\,, \quad \text{as} \quad z \to \infty\,.$$

3.1.14 Show that the series

$$\sum_{k=0}^\infty \frac{(-1)^k\,(2k)!}{z^{2k}\,2(k!)}$$

diverges for any $z > 0$.

3.1.15 Verify the asymptotic formula for $\mathrm{Var}[h(\overline{X}_n)]$ given in Theorem 3.1.4.

3.1.16 In the proof of Theorem 3.1.4, we carried out the Taylor series expansion to terms of order 4. Explain why carrying it out to terms of order 3 and only assuming that the third derivative of h is bounded will not work.

3.2 Probabilistic Asymptotics.

3.2.1 Convergence Modes for Sequences of Random Variables.

Recall that for a sequence of nonrandom vectors $\{\underline{x}_n\}$ we have $\underline{x}_n \to \underline{x}$ as $n \to \infty$ if and only if $\|\underline{x}_n - \underline{x}\| \to 0$, so that convergence of vectors is reduced to convergence of real numbers. One can check that $\underline{x}_n \to \underline{x}$ iff each component of $\underline{x}_n$ converges to the corresponding component of $\underline{x}$. For sequences of vector valued functions $\{\underline{f}_n\}$ there are several possible modes of convergence, e.g. pointwise convergence ($\underline{f}_n(x) \to \underline{f}(x)$ for each x in the domain) and uniform convergence ($\sup_x \|\underline{f}_n(x) - \underline{f}(x)\| \to 0$ as $n \to \infty$ where the supremum is over all x in the domain; note that uniform convergence implies pointwise convergence, but not *vice versa*. Random vectors are functions (whose domain is some underlying probability space), but none of these modes of convergence is useful for probability theory.

Definition 3.2.1 *Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random k-vectors defined on a probability space $(\Omega, \mathcal{F}, P)$.*

(a) *We say $\underline{X}_n$ converges to $\underline{X}$* almost surely *(abbreviated a.s.) or* with probability one *(abbreviated w.p.1), and write $\underline{X}_n \xrightarrow{a.s.} \underline{X}$ or* $\lim_n \underline{X}_n = \underline{X}$, *$P$-a.s., iff $P\{\,\omega \in \Omega : \underline{X}_n(\omega) \to \underline{X}(\omega)\,\} = 1$.*

(b) *For $p > 0$ we say $\underline{X}_n$ converges to $\underline{X}$ in* p^{th} mean, p^{th} moment, *or* in L_p, *and write $\underline{X}_n \xrightarrow{L_p} \underline{X}$, iff $E[\|\underline{X}_n - \underline{X}\|^p] \to 0$. Convergence in L_2 is sometimes referred to as* quadratic mean *convergence and written $\underline{X}_n \xrightarrow{q.m.} \underline{X}$.*

(c) *$\underline{X}_n$ converges to $\underline{X}$* in probability, *written $\underline{X}_n \xrightarrow{P} \underline{X}$, if and only if $\forall \epsilon > 0$, $P[\|\underline{X}_n - \underline{X}\| > \epsilon] \to 0$.*

Of course, we are already familiar with a.s. convergence and do not need a definition, but we find it useful to put all three definitions together. Of the three modes of convergence, a.s. convergence stands apart. Note that establishing convergence in probability and convergence in L_p only requires knowing the bivariate distributions Law$[\underline{X}_n, \underline{X}]$ so as to be able to compute $P[\|\underline{X}_n - \underline{X}\| > \epsilon]$ or $E[\|\underline{X}_n - \underline{X}\|^p]$. One typically needs to use the entire (infinite dimensional) distribution Law$[\underline{X}, \underline{X}_1, \underline{X}_2, \ldots]$ to establish a good a.s. convergence result.

The next result appears in Billingsley, Theorem 20.5, p. 274.

Proposition 3.2.1 *(a) $\underline{X}_n \xrightarrow{a.s.} \underline{X}$ implies $\underline{X}_n \xrightarrow{P} \underline{X}$.*

(b) $\underline{X}_n \xrightarrow{L_p} \underline{X}$ implies $\underline{X}_n \xrightarrow{P} \underline{X}$.

Proof. (a) Assume $\underline{X}_n \overset{a.s.}{\to} \underline{X}$. We want to show that for all $\epsilon > 0$ and all $\delta > 0$ there exists m such that for all $n \geq m$, $P[\|\underline{X}_n(\omega) - \underline{X}(\omega)\| \geq \epsilon] \leq \delta$. Fix $\epsilon > 0$. Let B be the null set where convergence fails in a.s. convergence. Then for all $\omega \notin B$, there exists $N = N(\omega)$ such that for all $n \geq N$, $\|\underline{X}_n(\omega) - \underline{X}(\omega)\| < \epsilon$. Now given $\delta > 0$ we can find m such that $P[N > m] < \delta$. Then for all $n > m$ we have $P[\|\underline{X}_n(\omega) - \underline{X}(\omega)\| \geq \epsilon] \leq P(B \cup [N > m]) \leq \delta$, as desired.

(b) By Markov's inequality, if $\underline{X}_n \overset{L_p}{\to} \underline{X}$, then

$$P[\,\|\underline{X}_n - \underline{X}\| > \epsilon\,] \;\leq\; \frac{E[\|\underline{X}_n - \underline{X}\|^p]}{\epsilon^p} \;\to\; 0 \quad ,$$

so $\underline{X}_n \overset{P}{\to} \underline{X}$.

□

The last result shows that convergence in probability is the weakest of the three modes of convergence introduced above.

3.2.2 Further Results on Almost Sure Convergence and Convergence in Probability.

Our first result is a classical one.

Proposition 3.2.2 (Borel-Cantelli Lemma) *Suppose B_n is a sequence of events on a probablity space and*

$$\sum_{n=1}^{\infty} P(B_n) \;<\; \infty.$$

Then

$$P\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} B_n\right) \;=\; 0.$$

Proof. Let

$$\begin{aligned} C_n &= \bigcup_{m=n}^{\infty} B_n \\ C_\infty &= \bigcap_{n=1}^{\infty} C_n. \end{aligned}$$

Note that $C_n \supseteq C_{n+1}$. We claim

$$P(C_\infty) \;=\; \lim_{n\to\infty} P(C_n) \quad . \tag{3.20}$$

To see this, define

$$D_n \;=\; C_n \setminus C_{n+1}.$$

Then

$$C_n \;=\; C_\infty \cup \bigcup_{k=n}^{\infty} D_n,$$

and the union is disjoint, so

$$P(C_n) \;=\; P(C_\infty) + \sum_{k=n}^{\infty} P(D_n)$$

and as $n \to \infty$, the summation on the r.h.s. tends to 0 since $\sum_{k=1}^{\infty} P(D_n)$ is finite.

Now

$$P(C_n) \;\leq\; \sum_{m=n}^{\infty} P(B_n),$$

by subadditivity. Since $\sum_{n=1}^{\infty} P(B_n) \;<\; \infty$, we have $\sum_{m=n}^{\infty} P(B_m) \to 0$ as $n \to \infty$, which proves the result.

□

The next result shows how this may be used to establish a.s. convergence.

Proposition 3.2.3 *Let a_n be a sequence of nonnegative numbers with $a_n \to 0$. Suppose X_n is a sequence of r.v.'s such that*

$$\sum_{n=1}^{\infty} P[|X_n| > a_n] \;<\; \infty.$$

Then $X_n \overset{a.s.}{\to} 0$.

Proof. Define the events

$$B_n \;=\; [|X_n| > a_n].$$

Clearly we can apply the Borel-Cantelli Lemma to the sequence B_n. It is merely a matter of unravelling the meaning of $P\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} B_n\right)$. Now

$$\begin{aligned} C_n &= \bigcup_{m=n}^{\infty} B_n \\ &= \{\omega : \omega \in B_m \text{ for some } m \geq n\} \\ &= \{\omega : \exists m = m(\omega) \geq n \text{ such that } |X_m(\omega)| > a_m\}. \end{aligned}$$

In the last line, we have written $m = m(\omega)$ to show that the m where $|X_m| > a_m$ depends on ω. It is a r.v. Now we have

$$\begin{aligned} C_\infty &= \bigcap_{n=1}^{\infty} C_n \\ &= \{\omega : \forall n \, \omega \in C_n\} \\ &= \{\omega : \forall n \, \exists m \geq n \text{ such that } |X_m(\omega)| > a_m\}. \end{aligned}$$

Now we know $P(C_\infty) = 0$, so let's see what happens on C_∞^c, the set of probability 1:

$$C_\infty^c \;=\; \{\omega : \exists n = n(\omega)\; \forall m \geq n \quad |X_m(\omega)| \leq a_m\}\,.$$

Thus, for $\omega \in C_\infty^c$, from some point on (namely, from $n(\omega)$ onward), the sequence a_m dominates the nonnegative sequence of real numbers $|X_m(\omega)|$. But this clearly implies $|X_m(\omega)| \to 0$ since $a_m \to 0$.

□

Remarks 3.2.1 If a_n is a sequence, then a subsequence a_{n_k} is obtained by taking the sequence of integers $n_1 < n_2 < n_3 < \ldots$. Note how we requiring the n_k's to be strictly increasing. If x_n is any sequence of real numbers and $x_n \to x$, then for any subsequence x_{n_k} we have $x_{n_k} \to x$ as $k \to \infty$. One can show further that if x_n is any sequence of real numbers, then $x_n \to x$ if and only if given any subsequence x_{n_k} there is a sub-subsequence $x_{n_{k_j}}$ such that $x_{n_{k_j}} \to x$. To see this, note that $x_n \to x$ implies $x_{n_k} \to x$ and hence $x_{n_{k_j}} \to x$ for any sub-subsequence $x_{n_{k_j}}$. Conversely, or contrapositively, if x_n does not converge to x, then there is some subsequence x_{n_k} which stays some minimal distance away from x (say, x_{n_k} never gets closer than ϵ to x), and so the same must be true of any subsequence of x_{n_k}, so no such sub-subsequence $x_{n_{k_j}}$ converges to x. These discussions, while a bit confusing and seemingly not very deep, do indicate that it is sometimes worthwhile to consider subsequences and sub-subsequences. Actually, in probability, there is a long tradition of "subsequence" arguments. We shall just touch on the subject.

□

Proposition 3.2.4 *Suppose* $\underline{X}_n \xrightarrow{P} \underline{X}$. *Then there is a subsequence* $\underline{X}_{n_k} \overset{a.s.}{\to} \underline{X}$.

Proof. Consider any sequences $0 < a_n \to 0$ and $b_n > 0$ with $\sum_n b_n < \infty$. Since $P[\|\underline{X}_n - \underline{X}\| > a_1] \longrightarrow 0$, we can find n_1 such that $P[\|\underline{X}_{n_1} - \underline{X}\| > a_1] < b_1$. Suppose we have constructed $n_1 < n_2 < \ldots n_k$ such that $P[\|\underline{X}_{n_j} - \underline{X}\| > a_j] < b_j$, $\forall j \leq k$. Since $P[\|\underline{X}_n - \underline{X}\| > a_{k+1}] \longrightarrow 0$, we can find $n_{k+1} > n_k$ such that $P[\|\underline{X}_{n_{k+1}} - \underline{X}\| > a_{k+1}] < b_{k+1}$. Thus, we have shown by induction that we can obtain a subsequence $\underline{X}_{n_k}$ such that

$$P[\|\underline{X}_{n_k} - \underline{X}\| > a_k] \;<\; b_k.$$

Now the result follows by the previous proposition.

□

The next result is especially useful.

Proposition 3.2.5 $\underline{X}_n \xrightarrow{P} \underline{X}$ *if and only if for every subsequence* $\underline{X}_{n_k}$ *there is a sub-subsequence* $\underline{X}_{n_{k_j}} \xrightarrow{a.s.} \underline{X}$.

Proof. If $\underline{X}_n \xrightarrow{P} \underline{X}$, then so does any subsequence: $\underline{X}_{n_k} \xrightarrow{P} \underline{X}$. Thus, by the previous proposition we can extract a sub-subsequence $\underline{X}_{n_{k_j}} \xrightarrow{a.s.} \underline{X}$.

Conversely, if $\underline{X}_n$ does not converge in probabilty to $\underline{X}$, then there is some $\epsilon > 0$ such that $P[\|\underline{X}_n - \underline{X}\| > \epsilon]$ does not tend to 0. But then there must be a subsequence $\underline{X}_{n_k}$ and a $\delta > 0$ such that $P[\|\underline{X}_{n_k} - \underline{X}\| > \epsilon] > \delta$, for all k. If $\underline{X}_{n_{k_j}}$ is any subsequence of $\underline{X}_{n_k}$, then it cannot converge a.s. to $\underline{X}$ since then it would converge in probability to $\underline{X}$ by Proposition 3.2.1 (a), but that would violate the fact $P[\|\underline{X}_{n_{k_j}} - \underline{X}\| > \epsilon] > \delta$ for all j.

□

From this last proposition, many results about a.s. convergence can be carried over to convergence in probability. The following is an example of this.

Proposition 3.2.6 *Suppose* $X_n \xrightarrow{P} X$. *Then the monotone and dominated convergence theorems apply.*

Proof. We will consider the dominated convergence theorem, so suppose $|X_n| \leq Y$ where Y is integrable. Put $a_n = E[X_n]$, and consider any subsequence a_{n_k} and the corresponding X_{n_k}. Then by the previous result, we have a subsequence $X_{n_{k_j}} \xrightarrow{a.s.} X$. Now the dominated convergence theorem applies to $X_{n_{k_j}}$ and we conclude $a_{n_{k_j}} \to E[X]$. The result now follows from Remarks 3.2.1

□

3.2.3 Continuous Mapping Principles.

In this subsection we consider the first two of the "continuous principles" we will need. Suppose $\underline{x}_n$ is a sequence of vectors and $\underline{x}_n \to \underline{x}$. If h is a function which is continuous at $\underline{x}$, then it is easy to show $h(\underline{x}_n) \to h(\underline{x})$. For stochastic convergence of random vectors, it is more complicated. The key definition we will need is the following.

Definition 3.2.2 *Let* $\underline{X}$ *be a random d-vector and* $h : I\!R^d \longrightarrow I\!R^k$ *a Borel measurable function. We say* h is a.s. continuous at $\underline{X}$ *if*

$$P\{\omega : h \text{ is continuous at } \underline{X}(\omega)\} \;=\; 1.$$

Proposition 3.2.7 (Continuous Mapping Principle for $\xrightarrow{a.s.}$ and $\xrightarrow{P}$) *Suppose* h *is a.s. continuous at* $\underline{X}$.

(a) If $\underline{X}_n \xrightarrow{a.s.} \underline{X}$, *then* $h(\underline{X}_n) \xrightarrow{a.s.} h(\underline{X})$.

(b) If $\underline{X}_n \xrightarrow{P} \underline{X}$, *then* $h(\underline{X}_n) \xrightarrow{P} h(\underline{X})$.

Proof. For part (a), let $N_1 = \{\omega : \underline{X}_n(\omega)$ does not converge to $\underline{X}(\omega)\}$ and $N_2 = \{\omega : h$ is not continuous at $\underline{X}(\omega)\}$. Both the N_i's are null sets, and hence also is $N = N_1 \cup N_2$. If $\omega \notin N$, then by our remark about sequences, $\underline{X}_n(\omega) \to \underline{X}(\omega)$, so the result follows.

Turning to part (b), suppose $\underline{X}_n \xrightarrow{P} \underline{X}$ and consider any subsequence $\underline{X}_{n_k}$. By Proposition 3.2.5 there is a sub-subsequence $\underline{X}_{n_{k_j}} \xrightarrow{a.s.} \underline{X}$. By part (a), we have $h(\underline{X}_{n_{k_j}}) \xrightarrow{a.s.} h(\underline{X})$. Thus, for the sequence $h(\underline{X}_n)$, we have shown that any subsequence has a sub-subsequence converging a.s. to $h(\underline{X})$. It follows from Proposition 3.2.5 that $h(\underline{X}_n) \xrightarrow{P} h(\underline{X})$.

□

We next turn to the issue of characterizing rates of convergence in probability.

3.2.4 Stochastic Orders of Convergence.

Definition 3.2.3 *Let $\{\underline{X}_n : n = 1, 2, \ldots\}$ be a sequence of random d-vectors and $\{b_n\}$ a sequence of nonnegative reals.*

(a) $\underline{X}_n = O_P(b_n)$ as $n \to \infty$ iff for all $\delta > 0$ there exist $C < \infty$ and $N < \infty$ such that for all $n \geq N$,

$$P\,[\,\|\underline{X}_n\| \leq C b_n\,] \;\geq\; 1 - \delta.$$

(b) $\underline{X}_n = o_P(b_n)$ iff for all $\epsilon > 0$ and $\delta > 0$ there exist $N < \infty$ such that for all $n \geq N$,

$$P\,[\,\|\underline{X}_n\| \leq \epsilon b_n\,] \;\geq\; 1 - \delta.$$

□

Remarks 3.2.2 (a) Assuming $b_n > 0$ for all n, $\underline{X}_n = o_P(b_n)$ is equivalent to $\underline{X}_n / b_n \xrightarrow{P} 0$ as $n \to \infty$.

(b) If $\underline{X}_n = a_n$ a.s. where $\{a_n\}$ is a sequence of constants (i.e. all r.v.'s are degenerate), then $\underline{X}_n = O_P(b_n)$ (or $= o_P(b_n)$, respectively) if and only if $a_n = O(b_n)$ ($= o(b_n)$, respectively).

□

Proposition 3.2.8 *Suppose $p > 0$ and $E[\|\underline{X}_n\|^p] = O(b_n^p)$ ($o(b_n^p)$, respectively). Then $\underline{X}_n = O_P(b_n)$ ($= o_P(b_n)$, respectively).*

Proof. By Markov's inequality, if $C_0 > 0$,

$$P\,[\,\|\underline{X}_n\| \,>\, C_0 a_n\,] \;=\; P\,[\,\|\underline{X}_n\|^p \,>\, (C_0 a_n)^p\,] \;\leq\; \frac{E[\|\underline{X}_n\|^p]}{(C_0 a_n)^p} \quad .$$

Now take N and C_1 such that for all $n \geq N$,

$$E[\|\underline{X}_n\|^p] \;\leq\; C_1 a_n^p \quad .$$

Also, choose C_0 above so that

$$\frac{C_1}{C_0^p} \;<\; \delta \quad ,$$

where $\delta > 0$ is given. Then, with this choice of C_0, we have for all $n \geq N$ that

$$P\,[\,\|\underline{X}_n\| \;\leq\; C_0 a_n\,] \;\geq\; 1 \,-\, \delta \quad .$$

The proof of the claim about $o_P(\cdot)$ is left as an exercise.

□

Proposition 3.2.9 **(a)** *$O_P(O(a_n)) = O_P(a_n)$.*

(b) *Any of $O_P(o(a_n))$, $o_P(O(a_n))$, or $o_P(o(a_n))$ is $o_P(a_n)$.*

(c) *$O_P(a_n)O_P(b_n) = O_P(a_n b_n)$.*

(d) *Either of $O_P(a_n)o_P(b_n)$ or $o_P(a_n)O_P(b_n)$ is $o_P(a_n b_n)$.*

(e) *$O_P(a_n) + O_P(a_n) = O_P(a_n)$.*

(f) *$o_P(a_n) + o_P(a_n) = o_P(a_n)$.*

Partial Proof. Consider the equality of the second and fourth members of (b). Suppose $b_n = O(a_n)$ and $\underline{X}_n = o_P(b_n)$. Then there exist C_1 and N_1 such that

$$b_n \;\leq\; C_1 a_n \;,\; \text{for all } n \geq N$$

and given $\epsilon > 0$, $\delta > 0$ there is an N_2 such that

$$P\left[\,\|\underline{X}_n\| \;\leq\; \epsilon C_1^{-1} b_n\right] \;\geq\; 1 \,-\, \delta \text{ for all } n \geq N_2 \quad .$$

Then for $n \geq N = \max\{N_1, N_2\}$, since $C_1^{-1} b_n \leq a_n$ for $n \geq N$,

$$P\,[\,\|\underline{X}_n\| \;\leq\; \epsilon a_n\,] \;\geq\; P\left[\,\|\underline{X}_n\| \;\leq\; \epsilon C_1^{-1} b_n\right] \;\geq\; 1 \,-\, \delta \quad .$$

This shows $\underline{X}_n = o_P(a_n)$.

Consider (c). Suppose $X_n = O_P(a_n)$ and $Y_n = O_P(b_n)$. Then given $\delta > 0$ there exist N_a, C_a, N_b, and C_b such that

$$P\,[\,|X_n| \;\leq\; C_a a_n\,] \;\geq\; 1 \,-\, \delta/2 \quad \text{for all } n \geq N_a \quad ,$$

$$P\,[\,|Y_n| \,\le\, C_b b_n\,] \,\ge\, 1 \,-\, \delta/2 \quad \text{for all } n \ge N_b \quad .$$

Now taking $N = \max\{N_a, N_b\}$ and $C = C_a C_b$, we have for all $n \ge N$,

$$P\,[\,|X_n Y_n| \,\le\, C a_n b_n\,]$$

$$\ge\; P\,[\,|X_n| \,\le\, C_a a_n \,\&\, |Y_n| \,\le\, C_b b_n\,]$$

$$=\; 1 \,-\, P\,[\,|X_n| \,>\, C_a a_n \text{ or } |Y_n| \,>\, C_b b_n\,]$$

$$\ge\; 1 \,-\, (\,P\,[\,|X_n| \,>\, C_a a_n\,] \,+\, P\,[\,|Y_n| \,>\, C_b b_n\,]\,)$$

$$\ge\; 1 \,-\, (\delta/2 \,+\, \delta/2) \;=\; 1 \,-\, \delta \quad .$$

□

The next result is related to parts (a) and (b) of the previous proposition.

Proposition 3.2.10 *Let $\underline{X}_n$ be a sequence of random d-vectors, a_n a sequence of nonnegative constants, and f a function defined on $I\!R^d$.*

(a) If $f(\underline{x}) = O(\|\underline{x} - \underline{x}_0\|)$ as $\underline{x} \to \underline{x}_0$ and $\underline{X}_n = \underline{x}_0 + O_P(a_n)$ where $a_n \to 0$, then $f(\underline{X}_n) = O_P(a_n)$.

(b) If $f(\underline{x}) = o(\|\underline{x} - \underline{x}_0\|)$ as $\underline{x} \to \underline{x}_0$ and $\underline{X}_n = \underline{x}_0 + O_P(a_n)$ where $a_n \to 0$, then $f(\underline{X}_n) = o_P(a_n)$.

(c) If $f(\underline{x}) = o(\|\underline{x} - \underline{x}_0\|)$ as $\underline{x} \to \underline{x}_0$ and $\underline{X}_n = \underline{x}_0 + o_P(a_n)$ where $a_n = O(1)$, then $f(\underline{X}_n) = o_P(a_n)$.

Proof. We prove part (b) only. Given $\eta > 0$ there is an M such that $P[\|\underline{X}_n - \underline{x}_0\| < Ma_n] > 1 - \eta$ for all n sufficiently large. Let $\epsilon > 0$ be given. Since $f(\underline{x}) = o(\|\underline{x} - \underline{x}_0\|)$ there is a $\delta > 0$ such that $\|\underline{x} - \underline{x}_0\| < \delta$ implies $\|f(\underline{x})\| < \epsilon M^{-1}\|\underline{x} - \underline{x}_0\|$. Taking n sufficiently large that $Ma_n \le \delta$ (which is possible since $a_n \to 0$), we have with probability $> 1 - \eta$ that $\|f(\underline{X}_n)\| < \epsilon M^{-1}\|\underline{X}_n - \underline{x}_0\| < \epsilon M^{-1} M a_n = \epsilon a_n$, which completes the proof of the claim.

□

3.2.5 The Laws of Large Numbers.

Now we state some of the classical limit theorems of probability.

Theorem 3.2.11 *Let $\underline{X}_1$, $\underline{X}_2$, ... be i.i.d. random d-vectors and $\overline{\underline{X}}_n = (1/n)\sum_{i=1}^n \underline{X}_i$.*

(a) (Khintchine's Weak Law of Large Numbers) If $E[\underline{X}_i]$ exists and equals $\underline{\mu}$, then $\overline{\underline{X}}_n \xrightarrow{P} \underline{\mu}$.

(b) (Kolmogorov's Strong Law of Large Numbers). $E[\underline{X}_i]$ exists and equals $\underline{\mu}$ if and only if $\overline{\underline{X}}_n \xrightarrow{a.s.} \underline{\mu}$.

□

Proofs of these results may be found in standard advanced texts in probability theory. For example, see Billingsley, pp. 80-81. More general results appear in Gnedenko and Kolmogorov (???). The weak law is easy to prove if one assumes the X_i's have finite second moment, for then (assuming $d = 1$ for convenience) if $\sigma^2 = \mathrm{Var}[X_i]$ we have $\mathrm{Var}[\underline{\overline{X}}_n] = \sigma^2/n \to 0$ which implies $\underline{\overline{X}}_n \overset{q.m.}{\to} \mu$, and this implies $\underline{\overline{X}}_n \overset{P}{\to} \mu$ by Proposition 3.2.1 (b). In fact, by Proposition 3.2.8, $\mathrm{Var}[X_i] < \infty$ implies $\underline{\overline{X}}_n = \mu + O_P(n^{-1/2})$, which gives a rate of convergence in the weak law. For the rate of convergence in the strong law, the Law of the Iterated Logarithm (Billingsley, Theorem 9.5, p. 151) states that if $\mathrm{Var}[X_i] < \infty$, then

$$\underline{\overline{X}}_n = \mu + O(n^{-1/2}(\log\log n)^{1/2}) , \quad a.s.$$

Here, $Y_n = O(a_n)$ almost surely means that there is a null set $\mathcal{N} \subset \Omega$ such that $Y_n(\omega) = O(a_n)$ for all $\omega \notin \mathcal{N}$. Note that the constants C and N in the definition of the $O(\cdot)$ are random (depend on ω). In the Law of the Iterated Logarithm, the factor $(\log\log n)^{1/2} \to \infty$, but at a very slow rate, so the "almost sure" refinement for the order of the error bound tends to 0 slightly slower than the "in probability" error bound $O_P(n^{-1/2})$.

Exercises for Section 3.2.

3.2.1 Show that $\underline{X}_n \xrightarrow{P} \underline{X}$ if and only if for all $\epsilon > 0$ there is an N such that for all $n \geq N$, $P[\|\underline{X}_n - \underline{X}\| > \epsilon] < \epsilon$.

3.2.2 Show that a sequence of random k-vectors $\{\underline{X}_n\}$ converges in probability to $\underline{X}_0$ if and only if each component $X_{ni} \xrightarrow{P} X_{0i}$, $1 \leq i \leq k$.

3.2.3 Verify Remarks 3.2.2 (a) and (b).

3.2.4 Prove Proposition 3.2.8 for $o(\cdot)$ and $o_P(\cdot)$.

3.2.5 Complete the proof of Proposition 3.2.9.

3.2.6 Complete the proof of Proposition 3.2.10.

3.2.7 Suppose X_1, X_2, $\ldots$ are i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be continuously differentiable in a neighborhood of μ. Show that $h(\overline{X}_n) = h(\mu) + O_P(n^{-1/2})$, where $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$ is the sample mean of n observations.

3.2.8 Suppose X_1, X_2, $\ldots$ are i.i.d. random variables with mean 0 and variance $\sigma^2 < \infty$. Let

$$Y_n \;=\; \sum_{i=1}^{n} i^p X_i$$

where p is a given real number. Show that

$$Y_n \;=\; \begin{cases} O_P(n^{p+1/2}) & \text{if } p > -1/2, \\ O_P((\log n)^{1/2}) & \text{if } p = -1/2, \\ O_P(1) & \text{if } p < -1/2. \end{cases}$$

3.2.9 Proposition 3.2.2 is sometimes referred to as the "first Borel-Cantelli Lemma." We consider here the "second Borel-Cantelli Lemma." Suppose the B_n are mutually independent (i.e., $P(\bigcap_k B_{n_k}) = \prod_k P(B_{n_k})$ for any finite subsequence B_{n_1}, $\ldots$, B_{n_m}). Then $\sum_n P(B_n) = \infty$ implies

$$P\left(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} B_m\right) \;=\; 1.$$

Hints: Consider the complementary probability. We have for any $n_2 > n_2$,

$$P\left(\bigcap_{m=n_1}^{n_2} B_m^c \;=\; \prod_{m=n_1}^{n_2}\right)[1 - P(B_m)]\,.$$

Now use that $1 - x < e^{-x}$.

3.2.10 Here we consider counterexamples to the converses of Proposition 3.2.1.

(a) Let U be a r.v. with uniform distribution on $[0,1]$. Define

$$X_n \;=\; n^{2/p} I_{[0,1/n]}(U).$$

Show that $X_n \xrightarrow{P} 0$ and $X_n \xrightarrow{a.s.} 0$, but we don't have $X_n \xrightarrow{L_p} 0$.

(b) We define a sequence of r.v.'s X_n which takes values 0 or 1. Define the "dyadic blocks" $N_k = \{2^{k-1},\ 2^{k-1}+1,\ \ldots\ , 2^k - 1\}$. Note that for each k, N_k is the set of positive integers whose binary expansion has k binary digits. ?? Let X_n be 1 at a position chosen uniformly over the set of values in N_k and independently among the blocks, and otherwise X_n is 0. Show that $X_n \xrightarrow{P} 0$ but that $X_n = 1$ infinitely often so we can't have $X_n \xrightarrow{a.s.} 0$.

3.2.11 Let X_1, X_2, ... be i.i.d. r.v.'s with *Expo*(1) distribution, and put

$$Y_n \;=\; \min\{X_1, X_2, \ldots, X_n\}.$$

(a) Show that $Y_n = O_P(1/n)$.

(b) Show that $Y_n = O((\log n)/n)$ a.s. (Hint: look at $\sum_n P[Y_n > k(\log n)/n]$ for suitable k.)

Remark: It is typical that if one has a $O_P(n^{-k})$ sequence for some $k > 0$, one can show it is $O(h(n)n^{-k})$ a.s. where h is some function which goes to ∞ very slowly with n, such as a power of $\log n$.

3.3 Convergence in Distribution.

The next mode of convergence we consider is the weakest and also the most useful.

Definition 3.3.1 *Let P, P_1, P_2, ... be Borel probability measures on $I\!R^d$. Then the sequence P_n* converges weakly *to P (written $P_n \Rightarrow P$) iff for every bounded and continuous function $h : I\!R^d \longrightarrow I\!R$,*

$$\int_{R^d} h(\underline{x})\, dP_n(\underline{x}) \;\to\; \int_{R^d} h(\underline{x})\, dP(\underline{x}) \quad .$$

Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random d-vectors, not necessarily defined on the same probability space. Then the sequence $\{\underline{X}_n\}$ converges in distribution to $\underline{X}$ (written $\underline{X}_n \xrightarrow{D} \underline{X}$) iff $Law[\underline{X}_n] \Rightarrow Law[\underline{X}]$, i.e. $E[h(\underline{X}_n)] \to E[h(\underline{X})]$ for arbitrary bounded and continuous $h : I\!R^d \longrightarrow I\!R$.

□

We depart from the usual definition in terms of convergence of the cumulative distribution functions at points of continuity (as on p. 335 and p. 390 in Billingsley) because for many purposes, this definition is much simpler and easier to use.

Theorem 3.3.1 *Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random d-vectors and F, F_1, F_2, ... the corresponding c.d.f.'s. The following are equivalent:*

(i) *$\underline{X}_n \xrightarrow{D} \underline{X}$;*

(ii) *$F_n(\underline{x}) \to F(\underline{x})$ at every point $\underline{x} \in I\!R^d$ at which F is continuous;*

(iii) *There exist random vectors $\underline{Y}$, $\underline{Y}_1$, $\underline{Y}_2$, ... such that $Law[\underline{Y}] = Law[\underline{X}]$, $Law[\underline{Y}_n] = Law[\underline{X}_n]$, $\forall n$, and $\underline{Y}_n \xrightarrow{P} \underline{Y}$.*

Before giving the proof of the theorem, we give a couple of other results that are useful in their own right.

Lemma 3.3.2 *Let $-\infty \leq a < b \leq \infty$ and suppose $G : (a, b) \longrightarrow I\!R$ is monotone nondecreasing. The following hold:*

(i) *Any discontinuity of G is a jump discontinuity, i.e. a jump discontinuity, i.e., both limits*

$$\begin{aligned} G(x-0) &= \lim_{y \uparrow x} G(y) \\ G(x+0) &= \lim_{y \downarrow x} G(y), \end{aligned}$$

exist, and there is a discontinuity if and only if the "jump"

$$J(x) \;=\; G(x+0) - G(x-0),$$

is positive (it is always nonnegative).

(ii) $D_G = \{x : G$ *is discontinuous at* $x\}$ *is countable.*

(iii) *Any nonempty interval* $(c,d) \subset (a,b)$ *contains infinitely many points where* G *is continuous.*

Sketch of Proof. Part (i) follows from monotonicity and the fact that a nonempty set of reals bounded below (above) has a finite infimum (supremum). The main result used in the proof of (ii) is that a countable union of countable sets is countable. It suffices to look at a subinterval $(a_1, b_1) \subset (a,b)$ where G is bounded (we may have $G(x)$ approaching $-\infty$ as $x \downarrow a$ or $+\infty$ as $x \uparrow b$), as we can get countably many such subintervals where G is bounded. If G is bounded on (a_1, b_1), then the number of jumps where $J(x) > 1/n$ for any n must be finite (as $G(b_1 - 0) - G(a_1 + 0)$ is bounded and greater than the sum of all the jumps in (a_1, b_1)), so we can represent the number of positive jumps in (a_1, b_1) as a countable union of finite sets. Part (iii) is immediate from (ii) since any nonempty interval has uncountably many points.

□

Recall the definition of the lower quantile function F^- in Definition 1.1.5.

Theorem 3.3.3 *Let* F, F_1, F_2, $\ldots$ *be cumulative distribution functions on* $I\!R$ *and assume* $F_n(x) \to F(x)$ *as* $n \to \infty$ *at every point* x *in the set*

$$C \;=\; \{x \in I\!R : F \text{ is continuous at } x\}.$$

Then $F_n^-(u) \to F^-(u)$ *as* $n \to \infty$ *at every point* u *in*

$$B \;=\; \{u \in (0,1) : F^- \text{ is continuous at } u\}.$$

Furthermore, $F_n^+(u) \to F^+(u)$ *at every point in* B.

Proof. Let $u \in B$ and $x = F^-(u)$. Put $C = \{x \in I\!R : F$ is continuous at $x\}$. The basic idea of the proof is to show that for $\epsilon > 0$ we can find x_1, x_2 in C with

$$x - \epsilon \;<\; x_1 \;\;<\; x \;<\;\; x_2 \;<\; x + \epsilon \tag{3.21}$$

$$F(x_1) \;\;<\; u \;<\;\; F(x_2) \quad . \tag{3.22}$$

Then, since $F_n(x_i) \to F(x_i)$, there exists m such that for all $n \geq m$, we have

$$F_n(x_1) \;<\; u \;<\; F_n(x_2) \quad . \tag{3.23}$$

But these inequalities imply (by definition of F_n^- and right continuity of F) that

$$x_1 \; < \; F_n^-(u) \; < \; x_2 \quad , \tag{3.24}$$

which in conjunction with (3.21) yields $|F_n^-(u) - F^-(u)| < \epsilon$, for all $n > m$.

To establish the existence of x_1 and x_2 in C and satisfying (3.21) and (3.22), we consider 4 cases. The first is if $x \in C$. Then, because $u \in D$, we have $F(y) < F(x) < F(z)$ for all $y < x$ and $z > x$. This follows since $u \in D$ and $x \in C$, we have $u = F(x)$ (refer to Figure 1.1.3), and there is not a jump at u, i.e., an interval in which F is constant.

The second case we consider is when $x \notin C$ and $F(x-0) < u < F(x)$. Then $F(y) < u < F(x) \leq F(z)$ for all $y < x$ and $z > x$, so we can easily find $x_1 \in (x-\epsilon, x) \cap C$ and $x_2 \in (x, x+\epsilon) \cap C$ satisfying (3.21) and (3.22). The other two cases are when $x \notin C$ and $u = F(x)$ or $u = F(x-0)$. Each of these is a "mixture" of the previous cases, so one can adapt the arguments accordingly.

The claim about F^+ is immediate since B is the set where $F^+ = F^-$.

□

Partial Proof of Theorem 3.3.1. We will treat in detail only the case $d = 1$. For (i) ⇒ (ii), assume $X_n \overset{D}{\to} X$ and we wish to show $F_n(x) \to F(x)$ if F is continuous at x. Now $F_n(y) = E[h(X_n)]$ where $h(x) = I_{(-\infty,y]}(x)$ is the indicator of the interval $(-\infty, y]$. This h is bounded but not continuous. We will approximate it by continuous functions given by

$$h_\epsilon(x) \; = \; \begin{cases} 1 & \text{if } x \leq y - \epsilon, \\ 1 - \epsilon^{-1}[x - (y-\epsilon)] & \text{if } y - \epsilon < x \leq y, \\ 0 & \text{if } x > y \end{cases}$$

$$g_\epsilon(x) \; = \; \begin{cases} 1 & \text{if } x \leq y, \\ 1 - \epsilon^{-1}(x - y) & \text{if } y < x \leq y + \epsilon, \\ 0 & \text{if } x > y + \epsilon. \end{cases}$$

Here, $\epsilon > 0$. Plots of these functions are given in Figure 3.3. Then $h_\epsilon(x) \leq I_{(-\infty,y]}(x) \leq g_\epsilon(x)$, so

$$\begin{array}{ccccc} E[h_\epsilon(X_n)] & \leq & F_n(y) & \leq & E[g_\epsilon(X_n)] \\ \downarrow {\scriptstyle n\to\infty} & & & & {\scriptstyle n\to\infty} \downarrow \\ E[h_\epsilon(X)] & \leq & F(y) & \leq & E[g_\epsilon(X)] \end{array}$$

Thus, for all $\epsilon > 0$,

$$E[h_\epsilon(X)] \; \leq \; \liminf F_n(y) \; \leq \; \limsup F_n(y) \; \leq \; E[g_\epsilon(X)].$$

Now $\lim_{\epsilon\to 0} E[g_\epsilon(X)] = F(y)$ by dominated convergence ($|g_\epsilon(x)| \leq 1$ for all x and all $\epsilon > 0$) and $\lim_{\epsilon\to 0} E[h_\epsilon(X)] = E[I_{(-\infty,y)}(X)] = F(y-0)$. If y is a continuity point of F, then $F(y) = F(y-0)$ so $\liminf F_n(y) = \limsup F_n(y) = F(y)$.

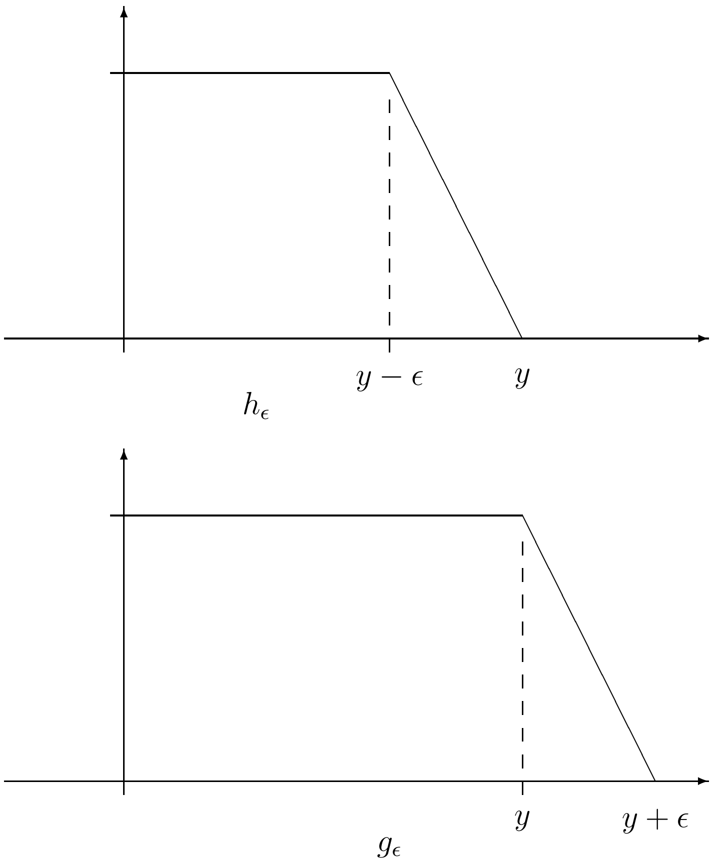


Figure 3.1: Graphs of h_ϵ and g_ϵ.

Turning now to (ii) ⇒ (iii). Let U be uniformly distributed on $[0,1]$ and put

$$\begin{aligned} Y_n &= F_n^-(U) \\ Y &= F^-(U), \end{aligned}$$

where F^- is given in Definition 1.1.5. It follows from Proposition 1.2.4 that the c.d.f.'s of Y and Y_n are F and F_n, respectively, so their distributions are the same as those of X and X_n, respectively (see Theorem 1.1.6 (b)). Note that by Lemma 3.3.2, the set of values u where F^- is discontinuous is countable and so has measure 0 under Law$[U]$. It follows from Theorem 3.3.3 that $F_n^-(U) \to F^-(U)$ a.s., and hence $Y_n \xrightarrow{P} Y$ (using Proposition 3.2.1 (a)).

For (iii) implies (i), assume (iii) holds and let h be a bounded continuous function from $I\!R \longrightarrow I\!R$. Since X_n and Y_n have the same distributions, $E[h(X_n)]$ $= E[h(Y_n)]$. Since h is continuous, we have $h(Y_n) \xrightarrow{P} h(Y)$ by (iii) and Proposition 3.2.7 (b). Since h is bounded, we have $E[h(Y_n)] \to E[h(Y)]$ by the dominated convergence theorem for convergence in probability (Proposition 3.2.6), and of course $E[h(Y)] = E[h(X)]$, so we can conclude that $E[h(X_n)] \to E[h(X)]$ for any bounded continuous h, and hence $X_n \overset{D}{\to} X$.

If $d > 1$ the proof of (i) $\Rightarrow$ (ii) can be done with functions which are products of the h_ϵ's and g_ϵ's in each component. The proof of (iii) from either (ii) or (i) is difficult. The proof that (iii) implies (i) already given works in any dimension d.

□

One can find the equivalence of (i) and (ii) in Billingsley, Theorem 25.8, page 344, for the one dimensional case, and Theorem 29.1, page 390, for the higher dimensional setting. The equivalence of (i) and (iii) is a famous result due to Skorohod (1956). See pp. 399-403 of Billingsley. Much research in modern probability has been motivated by this result.

Now we give various results on convergence in distribution.

Proposition 3.3.4 *(a) If $\underline{X}_n \overset{P}{\to} \underline{X}$, then $\underline{X}_n \overset{D}{\to} \underline{X}$.*

(b) If $\underline{X}_n \overset{D}{\to} \underline{x}$, a constant (nonrandom) vector, then $\underline{X}_n = \underline{x} + o_P(1)$, and so if all $\underline{X}_n$'s are defined on the same probability space, then $\underline{X}_n \overset{P}{\to} \underline{x}$.

Proof. Part (a) is immediate from the equivalence of (i) and (iii) in Theorem 3.3.1. Part (b) is left as an exercise.

□

Remarks 3.3.1 Part (a) shows that $\overset{D}{\to}$ is the weakest of the modes of convergence we have studied. Schematically,

$$\left.\begin{array}{lll} \underline{X}_n & \overset{a.s.}{\to} & \underline{X} \implies \\ \underline{X}_n & \overset{L_p}{\to} & \underline{X} \implies \end{array}\right\} \underline{X}_n \overset{P}{\to} \underline{X} \implies \underline{X}_n \overset{D}{\to} \underline{X}.$$

Of course, $\underline{X}_n \overset{P}{\to} \underline{X}$ means all random vectors are defined on the same probability space.

□

Proposition 3.3.5 (Continuous mapping principle) *Suppose $\underline{X}_n \overset{D}{\to} \underline{X}$ and ψ is a.s. continuous at $\underline{X}$. Then $\psi(\underline{X}_n) \overset{D}{\to} \psi(\underline{X})$.*

Proof. Using condition (iii) of Theorem 3.3.1, we can obtain $\underline{Y}$, $\underline{Y}_n$ with the same distributions and $\underline{Y}_n \overset{P}{\to} \underline{Y}$. By the continuous mapping principle for $\overset{P}{\to}$, Proposition 3.2.7 (b), we have that $\psi(\underline{Y}_n) \overset{P}{\to} \psi(\underline{Y})$, and $\psi(\underline{Y})$ and $\psi(\underline{Y}_n)$ have the same distributions as $\psi(\underline{X})$ and $\psi(\underline{X}_n)$, so Theorem 3.3.1 applies again to show $\psi(\underline{X}_n) \overset{D}{\to} \psi(\underline{X})$. (See Theorem 25.7, p. 343, of Billingsley.)

□

Theorem 3.3.6 (Lévy-Cramér Continuity Theorem) *Suppose $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, … have characteristic functions ϕ, ϕ_1, ϕ_2, …, respectively. Then $\underline{X}_n \overset{D}{\to} \underline{X}$ if and only if $\phi_n(\underline{u}) \to \phi(\underline{u})$ for all $\underline{u}$.*

Partial Proof. ($\Rightarrow$) Since $\phi_n(\underline{u}) = E[\cos(\underline{u}'\underline{X}_n)] + iE[\sin(\underline{u}'\underline{X}_n)]$ and $\cos(\underline{u}'x)$ and $\sin(\underline{u}'x)$ are bounded, continuous functions of x for each fixed u, it follows that $\phi_n(\underline{u}) \to \phi(\underline{u})$ for each fixed $\underline{u}$. Note how easy this proof is with our definition of convergence in distribution.

The proof of the converse is quite difficult. See Theorem 26.3, p. 359, of Billingsley.

□

Theorem 3.3.7 (Cramér-Wold device) *Random d-vectors $\underline{X}_n \overset{D}{\to} \underline{X}$ if and only if $\underline{u}'\underline{X}_n \overset{D}{\to} \underline{u}'\underline{X}$ for all $\underline{u} \in I\!R^d$.*

Proof. ($\Rightarrow$) Assuming $\underline{X}_n \overset{D}{\to} \underline{X}$, the mapping $h(\underline{x}) = \underline{u}'\underline{x}$ is continuous in $\underline{x}$ for each fixed $\underline{u}$, so $\underline{u}'\underline{X}_n \overset{D}{\to} \underline{u}'\underline{X}$ by the continuous mapping principle.

($\Leftarrow$) Assuming $\underline{u}'\underline{X}_n \overset{D}{\to} \underline{u}'\underline{X}$ for all $\underline{u}$ then by the Lévy-Cramér Continuity Theorem applied to the random variables $\underline{u}'\underline{X}_n$ with their characteristic functions evaluated at 1,

$$E[\exp(i\underline{u}'\underline{X}_n)] \ \to \ E[\exp(i\underline{u}'\underline{X})]$$

which implies that the characteristic functions for $\underline{X}_n$ converge to that of $\underline{X}$ for all $\underline{u}$, and hence that $\underline{X}_n \overset{D}{\to} \underline{X}$ by the other direction of Theorem 3.3.6.

□

Lemma 3.3.8 (Tightness) *If $\underline{X}_n \overset{D}{\to} \underline{X}$ then $\forall\epsilon > 0$ $\exists M$ such that $\forall n$,*

$$P\ [\ \|\underline{X}_n\| \ > \ M\ ] \ < \ \epsilon \quad , \tag{3.25}$$

i.e., $\underline{X}_n = O_P(1)$.

Proof. Fix $\epsilon > 0$. There is an M_1 such that $P[\|\underline{X}\| > M_1] < \epsilon/2$. If $M_2 \geq M_1$ is a continuity point of F, the c.d.f. of $\|\underline{X}\|$, (note: any nonempty interval of reals contains continuity points of F) then $P[\|\underline{X}_n\| > M_2] \to 1 - F(M_2)$ (since $x \mapsto \|x\|$ is a continuous function, it follows by the continuous mapping principle that $\|\underline{X}_n\| \overset{D}{\to} \|\underline{X}\|$). Hence, there is an N s.t. $P[\|\underline{X}_n\| > M_2] < \epsilon/2$ for all $n \geq N$. Now for each $n < N$ there exists K_n such that $P[\|\underline{X}_n\| > K_n] < \epsilon/2$, and if we put $M = \max\{K_1, K_2, \ldots, K_{n-1}, M_2\}$ then we have the desired result.

□

Theorem 3.3.9 (Convergence Equivalence) *Suppose* $(\underline{X}_1, \underline{Y}_1)$, $(\underline{X}_2, \underline{Y}_2)$, ... *are random 2d-vectors such that for each* $\epsilon > 0$,

$$\lim_{n \to \infty} P\,[\,\|\underline{X}_n - \underline{Y}_n\| \,>\, \epsilon\,] \;=\; 0 \quad , \tag{3.26}$$

(i.e., $\underline{X}_n = \underline{Y}_n + o_P(1)$*). If* $\underline{X}_n \overset{D}{\to} \underline{X}$*, then also* $\underline{Y}_n \overset{D}{\to} \underline{X}$*.*

Proof. We will assume the X's and Y's are 1-dimensional. The extension to higher dimensions is left as Exercise 3.3.2. Let F, F_n, and G_n denote the c.d.f.'s of X, X_n, and Y_n, respectively. Suppose ξ is a point of continuity of F and let $\epsilon > 0$ be such that $\xi \pm \epsilon$ are also continuity points of F. Then

$$\exists N \text{ such that } \forall\, n \geq N, \quad P[|X_n - Y_n| > \epsilon] \;<\; \epsilon \quad ,$$

so that $G_n(\xi) = P[Y_n \leq \xi] \leq P[X_n \leq \xi + \epsilon] + P[|X_n - Y_n| > \epsilon] \leq F_n(\xi + \epsilon) + \epsilon$, and similarly, arguing with the event $[Y_n > \xi] \subset [X_n > \xi - \epsilon] \cup [|X_n - Y_n| > \epsilon]$ we obtain $G_n(\xi) \geq F_n(\xi - \epsilon) - \epsilon$. Letting $n \to \infty$ we obtain

$$F(\xi - \epsilon) \,-\, \epsilon \;\leq\; \liminf_n G_n(\xi) \;\leq\; \limsup_n G_n(\xi) \;\leq\; F(\xi + \epsilon) \,+\, \epsilon \quad .$$

Letting $\epsilon \downarrow 0$ through a sequence of points for which $\xi \pm \epsilon$ are continuity points of F shows that $\lim_n G_n(\xi) = F(\xi)$. This proves $Y_n \overset{D}{\to} X$ by Theorem 3.3.1.

□

Theorem 3.3.10 (Slutsky's Theorem.) *Suppose* (X_1, Y_1), (X_2, Y_2), ... *are random 2-vectors such that* $X_n \overset{D}{\to} X$ *and* $Y_n \overset{D}{\to} c$ *where* c *is a constant. Then*

$$X_n + Y_n \;\overset{D}{\to}\; X + c \tag{3.27}$$

$$X_n Y_n \;\overset{D}{\to}\; cX \tag{3.28}$$

$$X_n / Y_n \;\overset{D}{\to}\; X/c \quad \textit{provided } c \neq 0 \tag{3.29}$$

Proof. We will prove (3.28) only, leaving the others as exercises. We will show that $X_n Y_n$ is convergence equivalent to cX_n. By the Tightness Lemma 3.3.8, given $\eta > 0$ there exists $M = M(\eta)$ such that for all n, $P[|X_n| > M] < \eta$. Now if $\epsilon > 0$ is given then

$$\begin{aligned} & P\,[\,|X_n Y_n - X_n c| \,\geq\, \epsilon\,] \\ & \quad \leq \quad P\,[\,|X_n(Y_n - c)| \geq \epsilon \text{ and } |X| \leq M\,] \;+\; P\,[\,|X| > M\,] \\ & \quad \leq \quad P\,[\,|Y_n - c| \geq \epsilon/M\,] \;+\; \eta \quad . \end{aligned}$$

Now given $\delta > 0$ choose $\eta \leq \delta/2$ in the above. For the corresponding $M = M(\eta)$ find N such that for all $n \geq N$, $P[|Y_n - c| \geq \epsilon/M] \leq \delta/2$. This follows by the argument in the proof of Proposition 3.3.4. Now for all $n \geq N$ we have $P[|X_nY_n - c| \geq \epsilon] \leq \delta$ which implies $P[|X_nY_n - cX_n| \geq \epsilon] \to 0$, i.e. that X_nY_n and cX_n are convergence equivalent. By the continuous mapping principle, $cX_n \xrightarrow{D} cX$, so by Theorem 3.3.9, $X_nY_n \xrightarrow{D} cX$.

□

Perhaps the most important theorem regarding convergence in distribution is the next one.

Theorem 3.3.11 (Lindeberg-Lévy Central Limit Theorem.) *Let X_1, X_2, ... be i.i.d. random variables with mean μ and finite variance σ^2. Put $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Then*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{D} N(0, \sigma^2) \quad .$$

The proof may be found in Billingsley, Section 27. We have abused notation slightly and written a distribution rather than a random variable on the r.h.s. of $\xrightarrow{D}$, but the meaning should be clear. Note where the mean and variance enter in as well as the factor $\sqrt{n}$. We will have numerous opportunities to apply this result.

Exercises for Section 3.3.

3.3.1 True or False (proof or counterexample):

(a) For random d-vectors, $\underline{X}_n \xrightarrow{D} \underline{X}$ if and only if each component $X_{ni} \xrightarrow{D} X_i$, $1 \leq i \leq d$.

(b) $X_n \xrightarrow{D} X$ implies $E[X_n] \rightarrow E[X]$.

(c) If $X_n \xrightarrow{D} Z$ where Z is $N(0,1)$, then $X_n^2 \xrightarrow{D} U$ where U has a χ_1^2 distribution.

(d) Part (3.27) of Slutsky's Theorem holds if the X_i and Y_i are random vectors.

3.3.2 Complete the proof of Theorem 3.3.9 to cover random vectors.

3.3.3 Using the notations and assumptions of the Central Limit Theorem show that $\sqrt{n}\overline{X}_n$ does not converge in distribution to anything if $\mu \neq 0$.

3.3.4 Prove Proposition 3.3.4 (b) by showing that $\underline{X}_n \xrightarrow{D} \underline{x}$ implies for all $\epsilon > 0$, $P[\|\underline{X}_n - \underline{x}\| > \epsilon] \rightarrow 0$, irrespective of whether or not the $\underline{X}_n$ are defined on the same probability space.

3.3.5 Prove parts (3.27) and (3.29) of Slutsky's theorem.

3.3.6 Let $\underline{X}_1$, $\underline{X}_2$, ... be i.i.d. random d-vectors satisfying $E\|\underline{X}_i\|^2 < \infty$. Put $\underline{\mu} = E[\underline{X}_i]$ and $V = \mathrm{Cov}[\underline{X}_i]$. Let $\bar{\underline{X}}_n$ be the average of n of the $\underline{X}_i$'s.

(a) Extend the Central Limit Theorem as stated in the text to i.i.d. random vectors $\underline{X}_i$: show that

$$\sqrt{n}\left[\bar{\underline{X}}_n - \underline{\mu}\right] \xrightarrow{D} N(\underline{0}, V) \quad .$$

(b) Suppose V is nonsingular. Show that

$$n\,(\bar{\underline{X}}_n - \underline{\mu})'V^{-1}(\bar{\underline{X}}_n - \underline{\mu}) \xrightarrow{D} \chi_d^2 \quad .$$

3.3.7 (a) Suppose $\underline{X}_n$ is a sequence of random vectors and $\underline{X}_n \xrightarrow{D} \underline{X}$. Let A_n be a sequence of random matrices such that $A_n \xrightarrow{P} A$ where A is a fixed matrix which is nonsingular. Show that $A_n^{-1}\underline{X}_n \xrightarrow{D} A^{-1}\underline{X}$ where A_n^{-1} is defined arbitrarily when it does not exist.

(b) Part (a) is a multidimensional generalization of (3.29). Give a corresponding generalization of (3.28).

3.4 Further Convergence in Distribution Results.

Here we give some other very useful results on convergence in distribution.

3.4.1 Asymptotic Linearization: The Delta Method.

We first give a very useful limit theorem for statistical applications. The result does not seem to have a formal name, but it widely known as the "δ–method."

Before stating the result, we review the notion of a derivative for a vector valued function of a vector variable. In general, for a function $\underline{f}$ defined on a domain in $\mathbb{R}^K$ taking values in $\mathbb{R}^J$, the derivative $D\underline{f}(\underline{x}_0)$ at $\underline{x}_0$ is a $J \times K$ matrix which satisfies

$$\underline{f}(\underline{x}) \;=\; \underline{f}(\underline{x}_0) \,+\, D\underline{f}(\underline{x}_0)(\underline{x} - \underline{x}_0) \,+\, o(\|\underline{x} - \underline{x}_0\|) \quad ,$$

as $\|\underline{x} - \underline{x}_0\| \to 0$. We assume $\underline{x}_0$ is an interior point of the domain of $\underline{f}$. If the derivative at $\underline{x}_0$ exists (i.e. if there is a $J \times K$ matrix which satisfies this last equation) then the (j, k) entry of $D\underline{f}(\underline{x}_0)$ is given by $\partial f_k/\partial x_j$ evaluated at $\underline{x}_0$. (Mnemonic Note: one can determine the dimensions of the matrix as $J \times K$ rather than $K \times J$, since if they were the latter the matrix multiplication $D\underline{f}(\underline{x}_0)(\underline{x} - \underline{x}_0)$ would not make sense. Once these dimensions have been determined, then it is easy to see that the (j, k) entry must be $\partial f_k/\partial x_j$ and not $\partial f_j/\partial x_k$.) This derivative matrix is sometimes called the Jacobian or the Jacobian matrix. In statistics, the "Jacobian" usually means the determinant of this matrix (when $J = K$ so the matrix is square), so we will always call it the Jacobian matrix. Note that if $\underline{f}$ is real valued ($J = 1$), then the Jacobian matrix is a $K \times 1$ matrix or a K-dimensional row vector, which is just the transpose of the usual gradient vector.

Theorem 3.4.1 *Suppose X_n is a sequence of random d-vectors and μ is a constant d-vector such that $\sqrt{n}(X_n - \mu) \xrightarrow{D} N(0, V)$ where V is a covariance matrix. Let h be a $\mathbb{R}^k$ valued function defined in a neighborhood of μ and suppose h is differentiable at μ. Then*

$$\sqrt{n}[h(X_n) - h(\mu)] \;\xrightarrow{D}\; N\left(0, Dh(\mu)VDh(\mu)'\right) \quad .$$

Before embarking on the proof, we make a remark. Note that this is very different from the continuous mapping principle, which would say that $h(\sqrt{n}[X_n - \mu]) \xrightarrow{D} h(Z)$ where Z is $N(0, V)$, if h is only continuous. One would not expect a normal distribution for $h(Z)$ in this case, unless h was very special (e.g. linear).

Proof. We first explain the ideas behind the proof, assuming that the X_n's and h are 1-dimensional. Now convergence in distribution of $\sqrt{n}(X_n - \mu)$ implies $X_n = \mu + O_P(n^{-1/2})$, and in particular tends to be quite close to μ for large

n. Thus, in a small neighborhood of μ, $h(x)$ looks like its first order Taylor expansion,

$$h(x) \;=\; h(\mu) \,+\, (x-\mu)\, Dh(\mu) \,+\, \ldots \quad .$$

With a little algebraic rearrangement, this gives

$$\sqrt{n}[h(X_n) - h(\mu)] \;=\; Dh(\mu)\, \sqrt{n}[X_n - \mu] \,+\, \ldots$$

and we see that the term on the r.h.s. $\xrightarrow{D}$ $N\left(0, V[Dh(\mu)]^2\right)$ (here, V is just a positive scalar). The only thing that remains is to make the result rigorous by taking care of the remainder denoted "..." above.

In fact, we will only complete the proof for real valued h ($k = 1$). The extension to vector valued h is left as an exercise. According to the definition of differentiablity of h at μ, there is a vector g (which is $\bigtriangledown h(\mu)$) such that

$$h(x) \;=\; h(\mu) \,+\, g'(x-\mu) \,+\, R(x) \quad ,$$

where the remainder is

$$R(x) \;=\; o(\|x-\mu\|) \quad \text{as } x \to \mu \quad .$$

Now we want to plug in the random vector X_n. By tightness, $\sqrt{n}[X_n - \mu]$ converging in distribution implies that $\sqrt{n}[X_n - \mu] = O_P(1)$ and hence that $X_n - \mu = O_P(n^{-1/2})$. It then follows from Proposition 3.2.10(b) that

$$R(X_n) \;=\; o_P(n^{-1/2}) \quad . \tag{3.30}$$

Then we have

$$\sqrt{n}[h(X_n) - h(\mu)] \;=\; g'\left\{\sqrt{n}[X_n - \mu]\right\} \,+\, \sqrt{n} o_P(n^{-1/2}) \quad . \tag{3.31}$$

Now

$$\sqrt{n} o_P(n^{-1/2}) \;=\; o_P(1)$$

by Proposition 3.2.9 (d) and the fact that $\sqrt{n} = O_P(n^{1/2})$. It follows that $\sqrt{n}[h(X_n) - h(\mu)]$ and $g'\{\sqrt{n}[X_n - \mu]\}$ are convergence equivalent. However, by the continuous mapping principle,

$$g'\{\sqrt{n}[X_n - \mu]\} \;\xrightarrow{D}\; N(0, g'Vg)$$

and so this is also the limit in distribution for $\sqrt{n}[h(X_n) - h(\mu)]$ by Theorem 3.3.9. This completes the proof.

□

Example 3.4.1 Let X_1, X_2, ... be i.i.d. $N(\mu, 1)$ where μ is unknown. Consider the estimator $\delta_n = \Phi(x_0 - \overline{X}_n)$ of $P_\mu[X_i > x_0]$ where x_0 is a given number. Now $n^{1/2}[\overline{X}_n - \mu]$ is exactly $N(0,1)$ and so also $\xrightarrow{D}$ $N(0,1)$. It follows from the δ-method that

$$\sqrt{n}\;[\delta_n \;-\; \Phi(x_0 - \mu)] \;\xrightarrow{D}\; N\left(0,\, \phi^2(x_0 - \mu)\right) \quad .$$

3.4.2 The Lindeberg Central Limit Theorem.

In this section, we give the most general version of the central limit theorem for independent random variables. Two applications of this theorem are considered: the asymptotic distribution of regression estimates assuming i.i.d. (but not necessarily normal) errors with mean 0 and finite variance, and the asymptotic distribution of the sample median.

The structure of the theorem is a little complicated. A *triangular array of row-wise independent r.v.'s* is a doubly indexed collection of r.v.'s $\langle Y_{ij} : 1 \leq j \leq n_i, \quad i \geq 1\rangle$ with n_i entries in each row. We suppose $n_i \to \infty$ as $i \to \infty$, and that for each $i \geq 1$, the i'th row $\langle Y_{ij} : 1 \leq j \leq n_i\rangle$ consists of independent r.v.'s. Thus, there may be dependence between the rows, but not within the rows. We will assume that the Y_{ij}'s have finite second moments denoted

$$E[Y_{ij}] \;=\; \mu_{ij} \qquad \mathrm{Var}[Y_{ij}] \;=\; \sigma_{ij}^2.$$

Let

$$S_i \;=\; \sum_{j=1}^{n_i} Y_{ij}$$

denote the sum of the r.v.'s in the i'th row, and

$$\mathrm{Var}[S_i] \;=\; s_i^2 \;=\; \sum_{j=1}^{n_i} \sigma_{ij}^2.$$

The *Lindeberg Condition* is that

$$\forall \epsilon > 0, \quad \lim_{i \to \infty} \sum_{j=1}^{n_i} \frac{1}{s_i^2} \int_{[|Y_{ij}-\mu_{ij}|>\epsilon s_i]} (Y_{ij} - \mu_{ij})^2 \, dP \;=\; 0. \tag{3.32}$$

Theorem 3.4.2 *[Lindeberg C.L.T.] Under the Lindeberg condition,*

$$\frac{S_n - E[S_n]}{s_n} \overset{D}{\to} N(0,1).$$

□

The proof may be found on pp. 369-371 of Billingsley. The Lindeberg condition is one of those technical mathematical conditions that even mathematicians dislike, but it is the best condition that has been found to obtain asymptotic normality of a sum of independent random variables.

Theorem 3.4.3 *Let X_1, X_2 , ..., X_n be i.i.d. random variables with c.d.f. F. Let $0 < \alpha < 1$ and denote the alpha'th quantile of the distribution by*

$$q_\alpha \;=\; F^{-}(\alpha),$$

where F^- is defined in Definition 1.1.5. Assume F is differentiable at q_α and $F'(q_\alpha) = f(q_\alpha) > 0$. Let Q_n be any sample α quantile, by which we mean

$$X_{(\lfloor \alpha n \rfloor)} \;\leq\; Q_n \;\leq\; X_{(\lceil \alpha n \rceil)}.$$

Then

$$\sqrt{n}\,(Q_n - q_\alpha) \;\xrightarrow{D}\; N\left(0, \alpha(1-\alpha)/f(q_\alpha)^2\right)$$

We used the notation f for the derivative of F as if F has a Lebesgue density, but that is not necessary. Of course, usually it will be the case.

Proof. The application of Theorem 3.4.2 is a little tricky. Let

$$Z_n \;=\; \frac{Q_n - q_\alpha}{(\sqrt{\alpha(1-\alpha)/n})/f(q_\alpha)}.$$

We wish to show that for any $z \in I\!R$,

$$P\,[Z_n \leq z] \;\rightarrow\; \Phi(z)$$

as $n \to \infty$. From our assumption on Q_n,

$$P\left[X_{(\lceil \alpha n \rceil)} \leq q_\alpha + z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha)\right] \;\leq\; P\,[Z_n \leq z] \;\leq$$

$$P\left[X_{(\lfloor \alpha n \rfloor)} \leq q_\alpha + z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha)\right],$$

where $X_{(i)}$ denotes the i'th order statistic. So, it suffices to show that the lower and upper bounds have the same limit, $\Phi(z)$. We will treat only the upper bound, it being clear from the argument that the lower bound follows as well.

To approximate the probability, let

$$Y_{nj} \;=\; \begin{cases} 1 & \text{if } X_j \leq q_\alpha + z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha); \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then for each n, the Y_{nj}'s are independent Bernoulli random variables with success probability

$$\begin{aligned} p_n \;&=\; F\left(q_\alpha + z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha)\right) \\ &=\; F(q_\alpha) \,+\, F'(q_\alpha)z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha) \,+\, o\left(\frac{1}{\sqrt{n}}\right) \\ &=\; \alpha \,+\, z\sqrt{\alpha(1-\alpha)/n} \,+\, o\left(n^{-1/2}\right) \end{aligned}$$

Then

$$S_n \;=\; \sum_{j=1}^{n} Y_{nj} \;\sim\; B(n, p_n),$$

has variance

$$s_n^2 \;=\; np_n(1-p_n) \;\sim\; n\alpha(1-\alpha).$$

Checking the Lindeberg condition in this case is relatively easy. Note that $|Y_{nj} - p_n| < 1$, so the set $\{|Y_{nj} - p_n| > \epsilon s_n\}$ is empty if $\epsilon s_n > 1$, which is the same as $n > 1/[\epsilon^2 p_n(1-p_n)]$, and this can be guaranteed for all n large enough that both $p_n(1-p_n) > \alpha(1-\alpha)/2$ and $n > 2/[\epsilon^2\alpha(1-\alpha)]$. Thus, all the integrals in the Lindeberg condition are eventually 0, so it holds trivially. Then we have

$$\frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} \;\xrightarrow{D}\; N(0,1),$$

or by a Slutsky argument,

$$\frac{S_n - np_n}{\sqrt{n\alpha(1-\alpha)}} \;\xrightarrow{D}\; N(0,1).$$

Now the event

$$\left[X_{(\lfloor \alpha n \rfloor)} \leq q_\alpha + z\sqrt{\alpha(1-\alpha)/n}/f(q_\alpha)\right]$$

is the same as the event $[S_n \geq \lfloor \alpha n \rfloor]$. To approximate the probability of this event by our CLT result about the S_n's we would consider

$$P\left[\frac{S_n - np_n}{\sqrt{n\alpha(1-\alpha)}} \;\geq\; \frac{\lfloor \alpha n \rfloor - np_n}{\sqrt{n\alpha(1-\alpha)}}\right].$$

There is a slight problem here as the limit on the r.h.s. of the inequality in the event is changing with n. However,

$$\lfloor \alpha n \rfloor \;=\; \alpha n \,+\, O(1)$$

and using our approximation of p_n above we obtain

$$\begin{aligned}
\frac{\lfloor \alpha n \rfloor - np_n}{\sqrt{n\alpha(1-\alpha)}} \;&=\; \frac{n(\alpha - p_n) \,+\, O(1)}{\sqrt{n\alpha(1-\alpha)}} \\
&=\; \frac{-n[z\sqrt{\alpha(1-\alpha)/n} \,+\, o(1/\sqrt{n})] \,+\, O(1)}{\sqrt{n\alpha(1-\alpha)}} \\
&=\; -z \,+\, o(1).
\end{aligned}$$

Thus,

$$\frac{S_n - np_n}{\sqrt{n\alpha(1-\alpha)}} \;-\; \left\{\frac{\lfloor \alpha n \rfloor - np_n}{\sqrt{n\alpha(1-\alpha)}} \,+\, z\right\}$$

is convergence equivalent with $(S_n - np_n)/\sqrt{n\alpha(1-\alpha)}$. Hence, the desired result follows from

$$P\left[\frac{S_n - np_n}{\sqrt{n\alpha(1-\alpha)}} \;\geq\; -z\right] \;\rightarrow\; \Phi(z).$$

□

Proposition 3.4.4 *Suppose $\underline{Y}_1$, $\underline{Y}_2$, ..., are i.i.d. random k-vectors with $E[\|\underline{Y}_i\|^2] < \infty$, $E[\underline{Y}_i] = \underline{\mu}$, $Cov[\underline{Y}_i] = V$. Suppose for $n = 1, 2, \ldots$, that $\underline{c}_{n1}, \ldots, \underline{c}_{nn}$ is a triangular array of nonrandom vectors such that*

$$\frac{\max_{1\leq i\leq n}\|\underline{c}_{ni}\|^2}{\sum_{i=1}^{n}\|\underline{c}_{ni}\|^2} \;\rightarrow\; 0, \tag{3.33}$$

as $n \rightarrow \infty$. Assume

$$\sigma_n^2 \;=\; \sum_{i=1}^{n} \underline{c}_{ni}^t V \underline{c}_{ni}$$

is positive for all n sufficiently large. Put

$$Z_n \;=\; \sigma_n^{-1}\sum_{i=1}^{n} \underline{c}_{ni}^t(\underline{Y}_i - \underline{\mu}).$$

Then $Z_n \xrightarrow{D} N(0,1)$.

□

The proof is left as an exercise (Exercise 3.4.6). In fact, one can show that the condition (3.33) is necessary and sufficient for the triangular array of r.v.'s $\{X_{ni} = c_{ni}Y_i : 1 \leq i \leq n, n = 1, 2, \ldots\}$ to satisfy the Lindeberg condition (Exercise 3.4.7). In this setting (with a triangular array of constants that multiply i.i.d. r.v.'s) we see that the Lindeberg condition is intuitively equivalent to the notion that no single one of the r.v.'s dominates in the sum – that they are all "small" compared to the sum. This result has useful applications in statistics.

3.4.3 Convergence to a Poisson.

The normal distribution is not the only distribution that occurs as the limiting distribution for a triangular array of random variables. We next consider a generalization of the Poisson approximation to the Binomial.

Proposition 3.4.5 *Suppose $\langle Y_{ij} : 1 \leq j \leq n_i, \quad i \geq 1\rangle$ is a triangular array of row-wise independent Bernoulli r.v.'s with $p_{ij} = P[Y_{ij} = 1]$. Suppose*

$$\sum_{j=1}^{n_i} p_{ij} \;\;\rightarrow\;\; \lambda > 0, \tag{3.34}$$

$$\max_{1\leq j\leq n_i} p_{ij} \;\;\rightarrow\;\; 0. \tag{3.35}$$

Then

$$S_i \;=\; \sum_{j=1}^{n_i} Y_{ij} \;\xrightarrow{D}\; Poisson(\lambda).$$

Proof. The ch.f. for Y_{ij} is

$$\phi_{ij}(u) \;=\; 1 + p_{ij}(e^{iu} - 1),$$

and the ch.f. for S_i is

$$\phi_i(u) \;=\; \prod_{j=1}^{n_i} \left[1 + p_{ij}(e^{iu} - 1)\right].$$

Now

$$\log(1 + x) \;=\; x \;+\; R(x),$$

where

$$R(x) \;=\; O(|x|^2), \tag{3.36}$$

as $x \to 0$ in the complex plane. Thus

$$\begin{aligned}
\log \phi_i(u) \;&=\; \sum_{j=1}^{n_i} \log\left[1 + p_{ij}(e^{iu} - 1)\right] \\
&=\; (e^{iu} - 1)\sum_{j=1}^{n_i} p_{ij} \;+\; \sum_{j=1}^{n_i} R(p_{ij}(e^{iu} - 1)).
\end{aligned}$$

Using (3.36) there exists K and δ so that $|x| < \delta$ implies $|R(x)| < K|x|^2$. By (3.35), there is an i_0 such that for all $i \geq i_0$, $p_{ij} < \delta/2$ which implies $p_{ij}|e^{iu}-1| < \delta$. For all $i \geq i_0$ we have

$$\begin{aligned}
\left|\sum_{j=1}^{n_i} R(p_{ij}(e^{iu} - 1))\right| \;&\leq\; K\sum_{j=1}^{n_i} p_{ij}^2 |e^{iu} - 1|^2 \\
&\leq\; 4K\sum_{j=1}^{n_i} p_{ij}^2 \\
&\leq\; 4K\sum_{j=1}^{n_i}\left(\max_{1\leq k\leq n_i} p_{ik}\right) p_{ij} \\
&=\; \left(\max_{1\leq k\leq n_i} p_{ik}\right)\left(4K\sum_{j=1}^{n_i} 4K\sum_{j=1}^{n_i} p_{ij}\right).
\end{aligned}$$

By (3.34) the second factor in the last expression is $O(1)$ as $i \to \infty$ and by (3.35) the first factor is $o(1)$, so the whole expression is $o(1)$, i.e.

$$\begin{aligned}
\log \phi_i(u) \;&=\; (e^{iu} - 1)\sum_{j=1}^{n_i} p_{ij} \;+\; o(1) \\
&\to\; (e^{iu} - 1)\lambda.
\end{aligned}$$

One recognizes the last expression as the log of the characteristic function of the $Poisson(\lambda)$ distribution.

□

3.4.4 Extreme Value Theory.

Now we consider a very different situation: the maximum of a sequence of i.i.d. r.v.'s as opposed to a mean or sum. For this subsection let Y_1, Y_2, …, be i.i.d. r.v.'s with c.d.f. F and consider the largest order statistic

$$X_n \;=\; \max_{1\le i\le n} Y_i.$$

The c.d.f. of X_n is of course

$$F_{X_n}(x) \;=\; (F(x))^n.$$

In order to make use of the limit relationship

$$\left[1 + \frac{a}{n} + o\left(\frac{1}{n}\right)\right]^n \;\to\; e^a, \tag{3.37}$$

we will work with

$$\bar{F}_{X_n}(x) \;=\; 1 - F_{X_n}(x) \;=\; \left(1 - \bar{F}(x)\right)^n,$$

where of course $\bar{F}(x) = 1 - F(x)$. Our interest is in whether or not there exist nonrandom sequences a_n and b_n such that $W_n = (X_n - a_n)/b_n \overset{D}{\to} W$ where W has a nondegenerate distribution. We see

$$F_{W_n}(w) \;=\; F_{X_n}(a_n + b_n w).$$

If we can obtain a result utilizing (3.37), it must be that

$$\bar{F}(a_n + b_n w) \;=\; -\frac{\gamma(w)}{n}(1 + o(1)), \tag{3.38}$$

for some function $\gamma(w)$ and then

$$\bar{F}_{X_n}(a_n + b_n w) \to \exp[-\gamma(w)],$$

so that $1 - \exp[-\gamma(w)]$ will give the limiting distribution.

Definition 3.4.1 *A function $h : (0,\infty) \longrightarrow (0,\infty)$ is called* slowly varying *if and only if for any $a > 0$,*

$$\lim_{x\to\infty} h(ax)/h(x) \;=\; 1.$$

□

The most common nonconstant slowly varying function is $\log x$, and one can obtain others from it such as $(\log x)^p$, $\log\log x$, etc.

Theorem 3.4.6 (Extreme Value Distributions) *Let $Y_1, Y_2, \ldots,$ be i.i.d. r.v.'s with c.d.f. F and $X_n = \max_{1\le i\le n} Y_i$. Let $y_0 = \sup\{y : F(y) < 1\}$. Put $\bar{F} = 1-F$.*

(a) *Suppose $y_0 = \infty$ and $\bar{F}(y) = y^{-p}h(y)$ for some positive constant p and some slowly varying function h. Let b_n be any sequence such that*

$$\bar{F}(b_n) \;=\; \frac{1}{n} + o\left(\frac{1}{n}\right). \tag{3.39}$$

Then $X_n/b_n \xrightarrow{D} W$ where W has c.d.f.

$$G_{1,p}(w) \;=\; \begin{cases} 1-\exp(-w^{-p}) & \textit{if } w \ge 0, \\ \\ 0 & \textit{if } w < 0. \end{cases} \tag{3.40}$$

(b) *Suppose $y_0 < \infty$ and $\bar{F}(y) = (y_0-y)^p h(1/(y_0-y))$ for some positive constant p and some slowly varying function h. Let b_n be any sequence satisfying (3.39). Then $(X_n - y_0)/b_n \xrightarrow{D} W$ where W has c.d.f.*

$$G_{2,p}(w) \;=\; \begin{cases} 1 & \textit{if } w \ge 0, \\ \\ \exp(-(-w)^p) & \textit{if } w < 0. \end{cases} \tag{3.41}$$

(c) *Suppose there exists a function $g(y)$ such that for all t,*

$$\frac{\bar{F}(y+tg(y))}{\bar{F}(y)} \;\to\; e^{-t}, \quad \textit{as } y \uparrow y_0. \tag{3.42}$$

Then $(X_n - a_n)/b_n \xrightarrow{D} W$ where a_n and b_n satisfy

$$\begin{aligned} \bar{F}(a_n) \;&=\; \frac{1}{n} + o\left(\frac{1}{n}\right), \\ b_n \;&\sim\; g(a_n), \end{aligned}$$

and W has c.d.f.

$$G_3(w) \;=\; \exp[-e^{-w}]. \tag{3.43}$$

Proof. For part (a), note that $b_n \to \infty$ and

$$\begin{aligned} \bar{F}(b_n w) \;&=\; (b_n w)^{-p} h(b_n w) \\ &=\; (b_n w)^{-p} h(w)(1+o(1)) \\ &\qquad \text{since } h \text{ is slowly varying} \\ &=\; w^{-p}\bar{F}(b_n)(1+o(1)) \\ &=\; \frac{w^{-p}}{n}(1+o(1)), \end{aligned}$$

where the last line follows from (3.39). The result now follows.

Part (b) is very similar to part (a) and is left as Exercise 3.4.8.

For part (c), consider first

$$\begin{aligned}\bar{F}(a_n + g(a_n)w) &= \bar{F}(a_n)e^{-w}(1 + o(1)) \\ &\quad \text{by (3.42)} \\ &= \frac{e^{-w}}{n}(1 + o(1)),\end{aligned}$$

where the last line follows from (3.43). This shows $\tilde{W}_n = (X_n - a_n)/g(a_n) \xrightarrow{D} W$ where W has the given distribution. Now (3.43) says $b_n/g(a_n) \to 1$, so by Slutsky's theorem $(b_n/g(a_n))\tilde{W}_n = (X_n - a_n)/b_n \xrightarrow{D} W$.

□

The distribution functions introduced in parts (a), (b), and (c) of the theorem are called *Extreme Value* or *Gumbel distributions of types I, II, and III*, respectively. It can be shown that these are the only possible limiting nondegenerate distributions for W_n of the form $(X_n - a_n)/b_n$, and that the distribution F must satisfy one of the three conditions in the theorem. See reference ???.

Example 3.4.2 *[The Weibull Distribution]* Consider a distribution function F which satisfies

$$F(y) \; = \; y^q h(1/y), \quad y \geq 0, \tag{3.44}$$

where h is a slowly varying function. For instance, if F is q times differentiable from the right at 0 and the first $q-1$ derivatives vanish and whose q'th derivative from the right is nonzero. (The derivative from the right is defined in the same way as the ordinary derivative except one takes limits from the right, i.e. the derivative from the right at y is $F'(y+0) = \lim_{h\downarrow 0}(F(x+h) - F(x))/h$. It is necessary to use derivatives from the right since $F(y) = 0$ for $y < 0$ and the usual derivative (from both sides) won't exist unless $F'(0+0) = 0$.) Then

$$F(y) \; = \; \frac{F^{(q)}(0+0)}{q!}y^q \; + \; R(y), \quad y \geq 0,$$

where $F^{(q)}(0+0)$ is the q'th derivative from the right at $y = 0$ and $R(y) = o(y^q)$. We define

$$h(x) \; = \; F^{(q)}(0+0)/q! \; + \; R(1/x)x^q, \quad x > 0,$$

and then for any $a > 0$,

$$\frac{h(ax)}{h(x)} \; = \; \frac{F^{(q)}(0+0)/q! \; + \; a^q R(1/(ax))x^q}{F^{(q)}(0+0)/q! \; + \; R(1/x)x^q}.$$

Now both $R(1/(ax))x^q$ and $R(1/x)x^q$ tend to 0 as $x \to \infty$. Since $F^{(q)}(0+0)/q!$ is nonzero, it follows that the last expression tends to 1 as $x \to \infty$.

A special case that satisfies (3.44) is the $Gamma(\alpha, \beta)$ distribution. Of course

$$F(y) \;=\; \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^y x^{(\alpha-1)} e^{-x/\beta}\, dx.$$

Since $e^{-x/\beta} \leq 1$ for all $x \geq 0$ we have

$$F(y) \;\leq\; \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^y x^{(\alpha-1)}\, dx \;=\; \frac{1}{\Gamma(\alpha+1)\beta^\alpha} y^\alpha.$$

(Recall that $\alpha\Gamma(\alpha) \;=\; \Gamma(\alpha+1)$.) On the other hand, for any $x_0 > 0$ we have $e^{-x/\beta} \geq e^{-x_0/\beta}$ for all $x \in [0, x_0]$ and hence

$$F(y) \;\geq\; \frac{1}{\Gamma(\alpha+1)\beta^\alpha} y^\alpha e^{-x_0/\beta}, \quad \forall y \in [0, x_0].$$

We see then that

$$\begin{aligned} F(y) \;&=\; \frac{1}{\Gamma(\alpha+1)\beta^\alpha} y^\alpha \, [1 + o(1)] \\ &=\; \frac{1}{\Gamma(\alpha+1)\beta^\alpha} y^\alpha \;+\; o(y^\alpha), \end{aligned}$$

which implies that F satisfies (3.44) by the same argument as in the previous paragraph. We could in fact have adapted the result of the previous paragraph to this case, but it is easier to simply argue directly.

Now let Y_1, Y_2, ... be i.i.d. with $Y_i > 0$ a.s. and assume their distribution function F satisfies (3.44). We consider

$$X_n \;=\; \min_{1 \leq i \leq n} Y_i.$$

Of course, we may put this in the framework of Theorem 3.4.6 (b) by taking $\tilde{Y}_i$ $= -Y_i$ and applying the results to $\tilde{F}(y) = 1 - F(y-0)$. These will give $\tilde{a}_n =$ $\sup\{y : \tilde{F}(y) < 1\} = 0$ and $\tilde{b}_n$ such that $[(-X_n) - \tilde{a}_n]/\tilde{b}_n \xrightarrow{D} W$, which turns into $X_n/b_n \xrightarrow{D} -W$ where and $b_n = \tilde{b}_n$. One easily sees that b_n solves

$$F(b_n) \;=\; 1/n \;+\; o(1/n).$$

The limiting distribution is

$$G_{-2,p}(y) \;=\; 1 \;-\; \exp[-w^q].$$

In particular, if $h(1/y) = c + o(1)$ as in our examples, then we may take

$$b_n \;=\; (nc)^{-1/q}.$$

It is simpler write the convergence in distribution result as

$$n^{1/q}X_n \xrightarrow{D} W,$$

where W has distribution function

$$F_W(w) \;=\; 1 \;-\; \exp[-(w/c^{1/q})], \quad w > 0.$$

The distribution function $1-\exp[-(x/\beta)^\alpha]$, $x > 0$, is known as the *Weibull distribution* function with *shape parameter* $\alpha > 0$ and *scale parameter* $\beta > 0$, denotes $Weibull(\alpha, \beta)$. This is found to be applicable in many practical situations, e.g. in breaking strength of materials or lifetimes of components where the weakest or shortest lived component determines the overall result.

□

Example 3.4.3 Suppose the Y_i's have a $N(0,1)$ distribution. We know from Example 3.1.1 that

$$\bar{F}(y) \;=\; \frac{\phi(y)}{y}(1+o(1)).$$

This clearly won't satisfy (b) of the theorem as $y_0 = \infty$, and neither will it satisfy (a) since the $\bar{F}(y)$ goes to 0 faster than any power of y, and the slowly varying function goes slower than any power. Thus, we must try to fit it into the framework of (c). Now

$$\frac{\bar{F}(y+tg(y))}{\bar{F}(y)} \;\sim\; \exp\left[-ytg(y) - (tg(y))^2/2\right]\frac{y}{y+tg(y)}.$$

The first term in the exponent looks something like $-t$. Let us try

$$g(y) \;=\; 1/y.$$

Then

$$\frac{\bar{F}(y+tg(y))}{\bar{F}(y)} \;\sim\; \exp[-t-t^2/(2y^2)]\frac{y}{y+tg(y)} \;\sim\; e^{-t},$$

as $y \to \infty$. Now let us find a_n so that

$$\bar{F}(a_n) \;\sim\; \frac{\phi(a_n)}{a_n} \;=\; \exp\left[-a_n^2/2 \;-\; \log a_n \;-\; \log\sqrt{2\pi}\right] \;=\; \frac{1}{n}(1+o(1)).$$

Taking logs and noting that $\log(1+o(1)) = o(1)$ we observe that the last equation is equivalent to

$$a_n^2/2 \;+\; \log a_n \;+\; \log\sqrt{2\pi} \;=\; \log n + o(1). \tag{3.45}$$

Let's try ignoring all but the first (and dominant) term on the l.h.s. which gives

$$a_{n1} \;=\; \sqrt{2\log n}.$$

Clearly $\log a_{n1} + \log\sqrt{2\pi}$ is not $o(1)$. In order to derive a "correction" to a_{n1}, let's use Newton's method which says that if a_1 is an approximate solution to $\psi(a) = 0$, then we can possibly improve on a_1 with $a_2 = a_1 - \psi(a_1)/\psi'(a_1)$. Here,

$$\begin{aligned}\psi(a) &= a^2/2 + \log a + \log\sqrt{2\pi} - \log n,\\ \psi'(a) &= a + 1/a.\end{aligned}$$

If one tries $a_{n2} = a_{n1} - \psi(a_{n1})/\psi'(a_{n1})$, one will see that the second term in the $\psi'(a_{n1})$ can be dropped. So, let us try

$$a_n \;=\; \sqrt{2\log n} - \frac{\log\sqrt{2\log n} + \log\sqrt{2\pi}}{\sqrt{2\log n}} \;=\; \sqrt{2\log n} - \frac{\log\sqrt{4\pi\log n}}{\sqrt{2\log n}}.$$

Note that this is $\sqrt{2\log n}+o(1)$. Also, $\log a_n = \log[\sqrt{2\log n}+o(1)] = \log[\sqrt{2\log n}(1+o(1))] = \log\sqrt{2\log n} + o(1)$. Thus

$$\begin{aligned}&a_n^2/2 + \log(\sqrt{2\pi}a_n)\\ &\quad= \log n - \left[\log\sqrt{2\log n} + \log\sqrt{2\pi}\right] + o(1) + \log\left(\sqrt{2\log n} + o(1)\right) + \log\sqrt{2\pi}\\ &\quad= \log n + o(1).\end{aligned}$$

We may take

$$1/a_n \;\sim\; 1/\sqrt{2\log n} \;=\; b_n.$$

Thus, for this case we have that if X_n is the largest order statistic in n i.i.d. standard normal random variables, then

$$\sqrt{2\log n}\left[X_n - \sqrt{2\log n} + \log\sqrt{4\pi\log n}/\sqrt{2\log n}\right] \;\xrightarrow{D}\; W,$$

where W has the G_3 distribution. Note in particular that we obtain

$$X_n \;=\; \sqrt{2\log n} - \log\sqrt{2\log n}/\sqrt{2\log n} + O_P\left(1/\sqrt{2\log n}\right). \tag{3.46}$$

One can show that the minimal order statistic is asymptotically independent of the maximal order statistic, and hence that if R_n is the range (difference between the maximal and minimal order statistics) in an i.i.d. sample of size n from $N(\mu, \sigma^2)$, then

$$\sqrt{2\log n}(R_n - 2a_n\sigma)/\sigma \;\xrightarrow{D}\; W_1 + W_2, \tag{3.47}$$

where W_1 and W_2 are i.i.d. from G_3. See Exercise 3.4.9. This shows that $R_n/(2a_n)$ $\xrightarrow{P}$ σ and in fact we can get a rate of convergence as

$$\frac{R_n}{2a_n} - \sigma \;=\; O_P(1/\log n). \tag{3.48}$$

This is much slower than the sample standard deviation which converges at the rate $O_P(1/\sqrt{n})$.

□

Exercises for Section 3.4.

3.4.1 Suppose $\sqrt{n}[X_n - \mu] \xrightarrow{D} N(0, \sigma^2)$. Let $h(x) = x^2$. What is the limiting distribution of $h(\sqrt{n}[X_n - \mu])$? What is the limiting distribution of just $h(X_n)$, with no centering and scaling? What is the limiting distribution of $\sqrt{n}[h(X_n) - h(\mu)]$?

3.4.2 Let X_1, X_2, …, X_n be i.i.d. with the exponential density with mean $\mu > 0$, i.e. the density is

$$f(x) \;=\; \mu^{-1} \exp[-x/\mu] \quad , \quad x > 0 \quad .$$

(a) Find a limiting normal distribution for $\overline{X}_n$, meaning for $\sqrt{n}[\overline{X}_n - \mu]$.

(b) Suppose we wish to estimate the probability $P[X > x_0] = \exp[-x_0/\mu]$, where x_0 is a given value. We use the estimate $\delta_n = \exp[-x_0/\overline{X}_n]$. Find a limiting normal distribution for δ_n, after suitable centering and scaling, of course.

3.4.3 Let X_1, X_2, …, X_n be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X}_n$ be the sample mean and

$$S_n^2 \;=\; \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

the sample variance.

(a) Show that $\sqrt{n}[S_n^2 - \sigma^2]$ has a limiting normal distribution and find the asymptotic variance. (Hint: nS_n^2/σ^2 has exactly a χ^2_{n-1} distribution, which is the same as the distribution of the sum of $n - 1$ i.i.d. χ^2_1 random variables.)

(b) Let $S_n = \sqrt{S_n^2}$ be the sample standard deviation. Find the limiting normal distribution for $\sqrt{n}[S_n - \sigma]$.

3.4.4 Complete the proof of Theorem 3.4.1 for vector valued h. (Hint: Cramér–Wold Device.)

3.4.5 Let X_1, X_2, …, X_n be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X}_n$ be the sample mean and S_n^2 the sample variance as defined in Exercise **??**. Find the asymptotic normal distribution (after appropriate centering and rescaling) for

$$\delta_n \;=\; \Phi([x_0 - \overline{X}]/S_n) \quad .$$

Why might this quantity be of interest?

3.4.6 Give the proof of Proposition 3.4.4.

3.4.7 Suppose Y_1, Y_2, … are i.i.d. r.v.'s with finite second moments and $\sigma^2 = \mathrm{Var}[Y_i] > 0$. Let $X_{ni} = c_{ni} Y_i$ where c_{ni} is a triangular array of constants as in Proposition 3.4.4. Show that a necessary condition for the X_{ni} to satisfy the Lindeberg condition is that

$$\frac{\max_{1 \le i \le n} c_{ni}^2}{\sum_{i=1}^{n} c_{ni}^2} \;\rightarrow\; 0, \quad \text{as } n \rightarrow \infty.$$

3.4.8 Prove part (b) of Theorem 3.4.6.

3.4.9 Let Y_1, Y_2, …, be i.i.d. r.v.'s such that both F_Y and F_{-Y} satisfy one of the three conditions of Theorem 3.4.6 (not necessarily the same one). Let $X_{n1} = \min\{Y_i : 1 \leq i \leq n\}$ and $X_{n2} = \max\{Y_i : 1 \leq i \leq n\}$. Let a_{ni} and b_{ni} be as in Theorem 3.4.6 such that $W_{ni} = (X_{ni} - a_{ni})/b_{ni} \xrightarrow{D} W_i$, $i = 1, 2$, where the W_i have one of the three extreme value distributions. Let $R_n = X_{n2} - X_{n1}$ be the sample range of Y_1, …, Y_n.

(a) Show that $(W_{n1}, W_{n2}) \xrightarrow{D} (W_1, W_2)$ where W_1 and W_2 are independent. (Hint: Consider the limit of $P[x_{n1} < X_{n1},\ X_{n2} \leq x_{n2}]$ for appropriate x_{ni}.)

(b) Assume that $b_{n1} = cb_{n2}$ for some $c > 0$. Show that $(R_n - (a_{n2} - a_{n1}))/b_{n1} \xrightarrow{D} W_1 + cW_2$ where (W_1, W_2) are as in part (a).

(c) Suppose $b_{n1} = o(b_{n2})$. Show that $(R_n - a_{n2})/b_{n2} \xrightarrow{D} W_2$.

3.5 Accuracy of Asymptotic Distributions.

??

The results of the previous two sections are widely used in statistics to derive asymptotic distribution results. As discussed in the first section of this chapter, any asymptotic result is an approximation, and one should have some idea of how accurate the approximation is before using it in practice. In this section, we study some examples of asymptotic approximations and assess their accuracy with finite sample calculations. In the first example, the finite sample calculations are possible without resorting to Monte Carlo simulation, but the second example is easily dispensed with by simulation, although it too can be done without simulation. However, Monte Carlo is quick and easy.

Example 3.5.1 We return to the setup of Example 3.4.1. Such a problem may arise in practice as follows: a water chemist is interested in estimating the percentage of households in a certain area whose tap water lead concentrations exceed a standard recently set by the Environmental Protection Agency. He takes tap water samples from n houses more or less randomly selected in the study area. (Of course, the selection process is of great concern to the statistician. Although many scientists seem to think that one obtains a random sample rather easily by simply not being too thoughtful about how to select houses, in this case, that in fact is not true. Subtle and unconscious forces can operate to bias the sample, but after the data is collected the statistician can do little more than gently warn the scientist to seek assistance in sample selection in the future. The statistician will then analyze the data as if it were a random sample and hope that it is not too biased.) Some data snooping indicated that the log-normal distribution seemed to fit reasonably well, so after taking logarithms of the data, a normal distribution is fit by estimating mean and variance. In order to keep things simple for now, we assume that the variance is known. A more practical result would be obtained by using Exercise 3.4.5. One point to be made here – this experiment will be carried out over and over for various locales. The water chemist wants a statistical procedure that can be routinely applied in field work. Therefore, it is important that it be simple, e.g. implementable on a hand calculator possibly in conjunction with a book of tables.

Thus, we showed in Example 3.4.1 that $\delta_n = \Phi(x_0 - \overline{X}_n)$ satisfies $\sqrt{n}[\delta_n - \Phi(x_0 - \mu)] \xrightarrow{D} N(0, \phi^2(x_0 - \mu))$. Now in this case, we can obtain a closed form solution for the c.d.f. of δ_n, or to make comparison with our asymptotic distribution result more easy, for $Z_n := \sqrt{n}[\delta_n - \Phi(x_0 - \mu)]/\phi(x_0 - \mu)$, which has an asymptotic $N(0,1)$ distribution. We have

$$F_n(z) \;=\; P[Z_n \leq z] \;=\; 1 \;-\; \Phi\left(\sqrt{n} \left\{ \xi \;-\; \Phi^{-1} \left(\Phi(\xi) + n^{-1/2}\phi(\xi) z \right) \right\} \right) \quad , \tag{3.49}$$

where

$$\xi \;=\; x_0 - \mu \quad .$$

The derivation is left as Exercise 3.5.1. An S-Plus function was written to evaluate $F_n(z)$ for a given value of ξ and n. The listing of the function is

```
> pr1
function(z, x0, n)
{
#computes the c.d.f. at z of Phi(x0-barXn) under mu=0; accepts vector z
t1 <- sqrt(n)
t2 <- (dnorm(x0) * z)/t1
t2 <- pnorm(x0) + t2
t2 <- qnorm(t2)
t2 <- t1 * (x0 - t2)
t2 <- pnorm( - t2)
return(t2)
}
```

Some comments on this function. The argument `z` may be a whole vector, which is especially convenient for generating plots. The argument `x0` is what was denoted ξ above. The function was already input into the system (see *The New S Language* manual for a discussion of this), and we typed the function name "`pr1`" at the S prompt ("> ") to produce this listing. The first line ("`function(z, x0, n)`") just tells us that this S object is a function, and lists the arguments. Then the lines between the braces "{ …}" give the actual statements in the function. The line beginning with the pound sign "#" is a comment. Note that we have broken up the evaluation of the function into a number of steps. For such a complicated function, this is a good idea as it makes it easier to find possible errors. The S functions `dnorm`, `pnorm`, and `qnorm` correspond to ϕ, Φ, and Φ^{-1}, respectively, in our notation, i.e. the standard normal density, distribution, and quantile functions.

The c.d.f. F_n was evaluated 101 equally spaced values of z between ± 2.5 inclusive, i.e. -2.50, -2.45, -2.40, …, 2.45, 2.50, and for $\xi = 0$, .5, 1.0, 1.5, 2.0, 2.5. A problem can arise here since $0 < \Phi(x_0 - \overline{X}_n) < 1$ which implies

$$\frac{-\sqrt{n}\Phi(\xi)}{\phi(\xi)} \; < \; Z_n \; < \; \frac{\sqrt{n}[1-\Phi(\xi)]}{\phi(\xi)} \quad .$$

If z goes outside this range, then the $\Phi^{-1}(\cdot)$ in (3.49) will fail to be defined (since it's argument must be inside $(0,1)$). In fact, we had problems with this when ξ $= 2.0$ and $\xi = 2.5$, so we eliminated them from further consideration. (In a more careful study, we would have written the S code for the function F_n to recognize this and give the right answer, either 0 or 1 depending on what side of the limits in the last display z was.)

In Figure 3.5.1 in the upper left plot is shown the $N(0,1)$ c.d.f. and the 4 c.d.f.'s for the different ξ values. (We denoted ξ as `x0` in this figure, since the two agree when $\mu = 0$.)

The limiting $N(0,1)$ practically coincides with the $\xi = 0$ case, and the approximation becomes worse as ξ increases to 1.5. In general, one might think the approximation looks quite good, but this is somewhat misleading as the c.d.f.'s have a roughly common shape and the "centering" is pretty close (note how well they match in the neighborhoos of $z = 0$). We now consider some different ways of looking at this approximation error. The first alternative we consider is the probability-probability or P-P plot. Since $F_n(z)$ is supposed to be close to $\Phi(z)$, if we plot points $(\Phi(z), F_n(z))$ for various values of z, they should fall nearly on the 45 degree line within the unit square $[0,1] \times [0,1]$. This plot is shown in the upper right of Figure 3.5.1. In the lower left of Figure 3.5.1 we show a blowup of the upper right corner of the upper right plot in Figure 3.5.1, and have also overlaid the 45 degree line. We see for example that when $\xi = 1.5$, if $\Phi(z) = .95$ then $F_n(z)$ is approximately .99. Now this is important since we will be interested in probabilities in this range when we construct approximate confidence intervals in the next chapter (in particular, $\Phi(z) = .95$ will be important for constructing 90% confidence intervals). Looking at $1 - \Phi(z)$ and $1 - F_n(z)$, we see that the approximation doesn't look so great – the approximating value $1 - \Phi(z)$ is about 5 times the true value $1 - F_n(z)$.

Now in the construction of confidence intervals, we will want to find a value $z_{.05}$ such that $F_n(z_{.05}) = .95$, i.e. $z_{.05} = F_n^{-1}(1 - .05)$. Using the approximating normal distribution, we would use $\Phi^{-1}(1 - .05)$ as the approximation. Now it is clearly of interest to see how well the quantiles of Φ approximate the quantiles of F_n. To assess this, we plot $(\Phi^{-1}(p), F_n^{-1}(p))$ for various values of p in $(0,1)$. This gives a Quantile-Quantile or QQ-plot. This is given in the lower right of Figure 3.5.1. Note for instance that for $\xi = 1.5$, $\Phi^{-1}(.975) = 1.96$ is not so close to $F^{-1}(.975)$ which is about 1.49. Based on this, we see that a 95% confidence interval constructed from the asymptotic normal distribution might tend to be too large, by roughly 1/3. There is an extra complication here – we will have to estimate the standard deviation $\phi(x_0 - \mu)$ of the limiting normal distribution, which will introduce further approximation errors. These matters are taken up in the a subsequent chapter. As one can see by comparing the three types of plots (c.d.f. plot, PP-plot, and QQ-plot), the QQ-plot tends most to magnify differences between the distributions, especially in the tails for typical distributions. See Exercise 3.5.2 for some justification of this. For this reason, the QQ-plot is preferred by statisticians when comparing distributions. There is an added advantage when one wishes to compare with a normal distribution of unknown mean and variance – simply plotting against N(0,1) quantiles will produce a straight line if the given distribution is $N(\mu, \sigma^2)$, no matter what the values are for μ and σ^2. See Exercise 3.5.3.

What we have learned from this example is that the asymptotic approximation is quite good when $n = 30$ if x_0 is close to μ (i.e. ξ is close to 0). As x_0 gets bigger, the accuracy goes down, although it stays good in a neighborhood of $z = 0$ (this is true roughly speaking since the Taylor series expansion is exact at

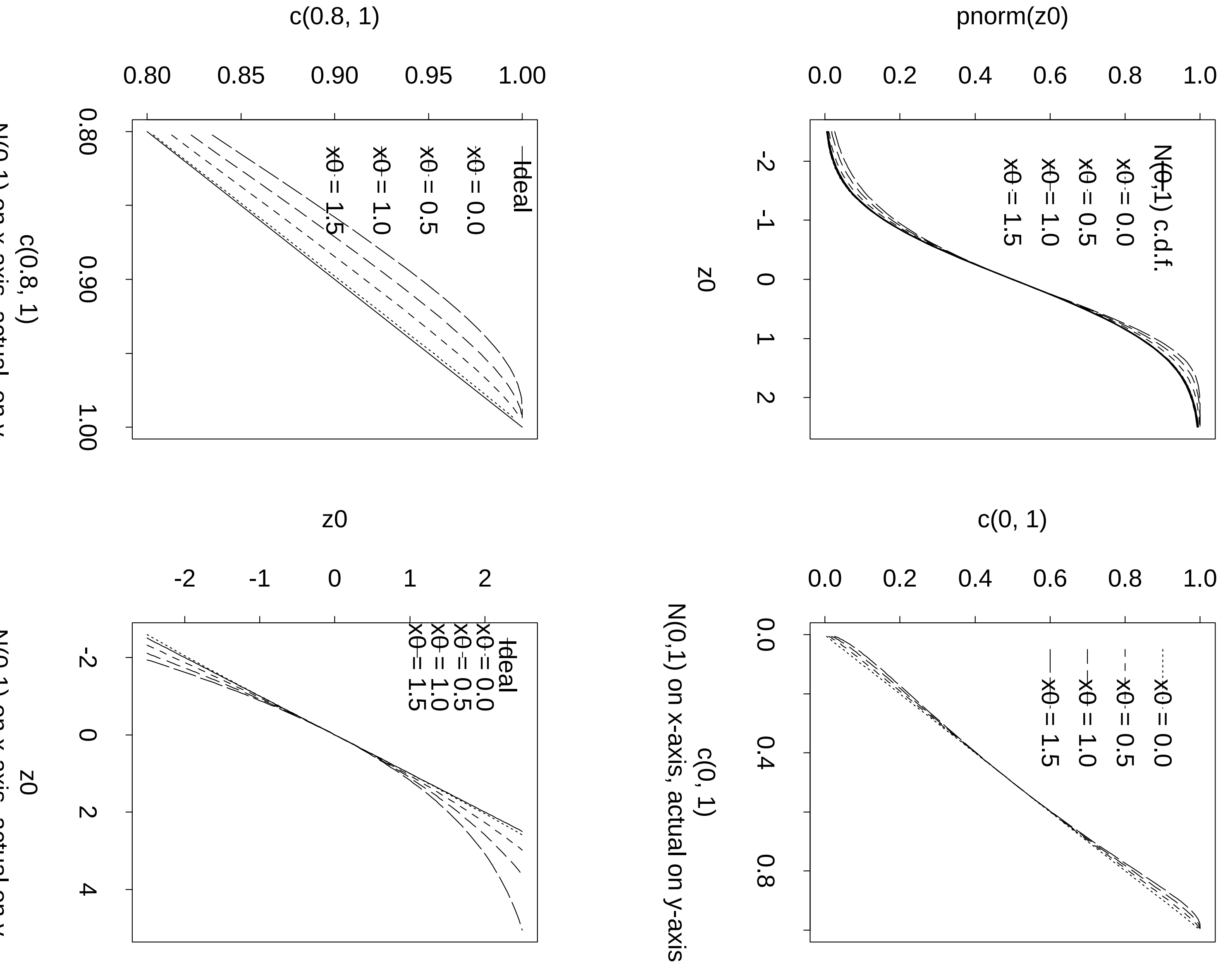


Figure 3.2: Plots to Assess Accuracy in Example 3.5.1

$z = 0$). In the context of our practical example involving the logarithms of lead concentrations in drinking water, the value of x_0 is actually much larger than μ (as the E.P.A. limit is somewhat larger than the lead concentrations typically found in drinking water, at least we hope), so this suggests that the asymptotic approximation may not work so well, certainly when $n = 30$. Further, at least for x_0 between 0 and 1.5, the tails of the actual distribution are lighter than the normal on the right (note that $F_n(z) > \Phi(z)$ for $z > 0$, i.e. the right tail tends to 1 more quickly than the normal approximation) and heavier on the left (the left tails of F_n tend to 0 more slowly than the normal). One could of course repeat this for a larger value of n, and would presumably find that the normal approximation worked better. Given some values for how large an approximation error is acceptable, one can find how large n must be to achieve that.

□

Example 3.5.2 It follows from Theorem 3.4.3 that if X_1, ..., X_n are i.i.d. from a distribution which has a Lebesgue density $f(x)$ which is continuous and positive at the median m then the sample median M_n has the following asymptotic distribution:

$$\sqrt{n}[M_n - m] \overset{D}{\to} N(0, 1/[4f^2(m)]) \quad .$$

We will investigate the accuracy of this asymptotic result for a couple of different distributions. Monte Carlo simulation will be used for this purpose. We will generate a large number N of samples of size n from given distributions, and for each sample compute M_n and

$$Z_n \;=\; 2f(m)\sqrt{n}[M_n - m]$$

and compare this with the standard normal as in the previous example. We will use $N = 1,000$ and $n = 100$. It is important to distinguish between the underlying sample size n and the number of Monte Carlo trials N.

We first try this with the $N(0, 1)$ distribution. Note that $m = 0$ and $\phi(m) = 1/\sqrt{2\pi} = 0.3989423$. The following Splus code generates the 1000 sample medians from samples of 100 N(0,1) random variables, centers and normalizes, and then sorts (into ascending order):

```
> Zn_apply(matrix(rnorm(100*1000),nrow=100),2,median)
> dnorm(0)
[1] 0.3989423
> Zn_2*0.3989423*sqrt(100)*Zn
> Zn_sort(Zn)
```

The first line performs a lot of calculations with a few keystrokes. The command "`(rnorm(100*1000)`" generates 100 × 1000 N(0,1) random variates (it is common to speak of the output of a random number generator as a " random variate").

These are returned in a big vector (think of it as a column vector) of length 100 × 1000. The `matrix` function takes a vector and constructs a matrix. In this case, the first 100 entries of the vector returned by "`(rnorm(100*1000)`" go into the first column, the second 100 entries into the second column, and so forth, until the vector is exhausted, giving 1000 columns. The `apply` function is extremely useful. If `x` is a matrix and `func` is a real valued function of a vector, then "`apply(x,2,func)`" returns a vector whose length is the number of columns of `x`, and whose i^{th} entry is the result of applying `func` to the i^{th} column of `x`. To apply `func` to the rows of `x`, use "`apply(x,1,func)`." The first line above could have been replaced with the following for the same result:

```
> Zn_NULL
> for(i in 1:1000) Zn_c(Zn,median(rnorm(100)))
```

However, this last piece of code is guaranteed to take much, much longer than the `apply` statement. **One should avoid the use of `for` loops in Splus whenever possible.**

Next, we turn on the graphics device and plot the empirical c.d.f. and overlay the N(0,1) c.d.f. Recall that the empirical c.d.f. of data X_1, ..., X_n is given by

$$\hat{F}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \;=\; \frac{1}{n}\#\{i : X_i \leq x\} \quad .$$

The Splus code:

```
> X11() #on-screen graphics device in X-windows
> plot(Zn,(1:1000)/1000)
> lines(Zn,pnorm(Zn))
> #Almost perfect fit!
```

Note that it is critical that we used the sorted `Zn` in the `plot` command. The plot may be found in the upper left of Figure 3.5.2. We also want to make the corresponding QQ-plot. There is a minor difficulty here as the empirical distribution is discontinuous at each of the data points. To overcome this, we associate the i^{th} order statistic of the data with the probability $(i - .5)/N$. In particular, this way we avoid 0 and 1. The corresponding code is

```
> plot(qnorm((1:1000)/1000 -.5/1000),Zn)
> #find a good range for drawing in the 45 degree line
> range(Zn)
[1] -3.954932  2.769698
> qnorm(.5/1000)
[1] -3.290527
> lines(c(-3.290527,2.769698),c(-3.290527,2.769698))
```

The plot may be found in upper right of Figure 3.5.2. Again, in this case the asymptotic approximation works extremely well.

Turning now to another distribution, we consider $n = 100$ again but this time from a gamma distribution with $\alpha = 1.5$, which is also a χ^2_3 distribution. The code for generating the data, centering, and normalizing:

```
> Zn_apply(matrix(rgamma(100*1000,1.5),nrow=100),2,median)
> qgamma(.5,1.5)
[1] 1.182987
> dgamma(qgamma(.5,1.5),1.5)
[1] 0.3759935
> Zn_2*0.3759935*sqrt(100)*(Zn-1.182987)
```

The rest is the same as for the $N(0,1)$ case. The plot of the empirical c.d.f. and the QQ-plot are given in the lower part of Figures 3.5.2. Again, the approximation is excellent.

We don't mean to leave the reader with the impression that the asymptotic distribution of the median always works extremely well. One can construct distributions where it doesn't work well, for instance if $f(m)$ is rather small as for a bimodal distribution where the median is between the two modes. See Exercise 3.5.4.

□

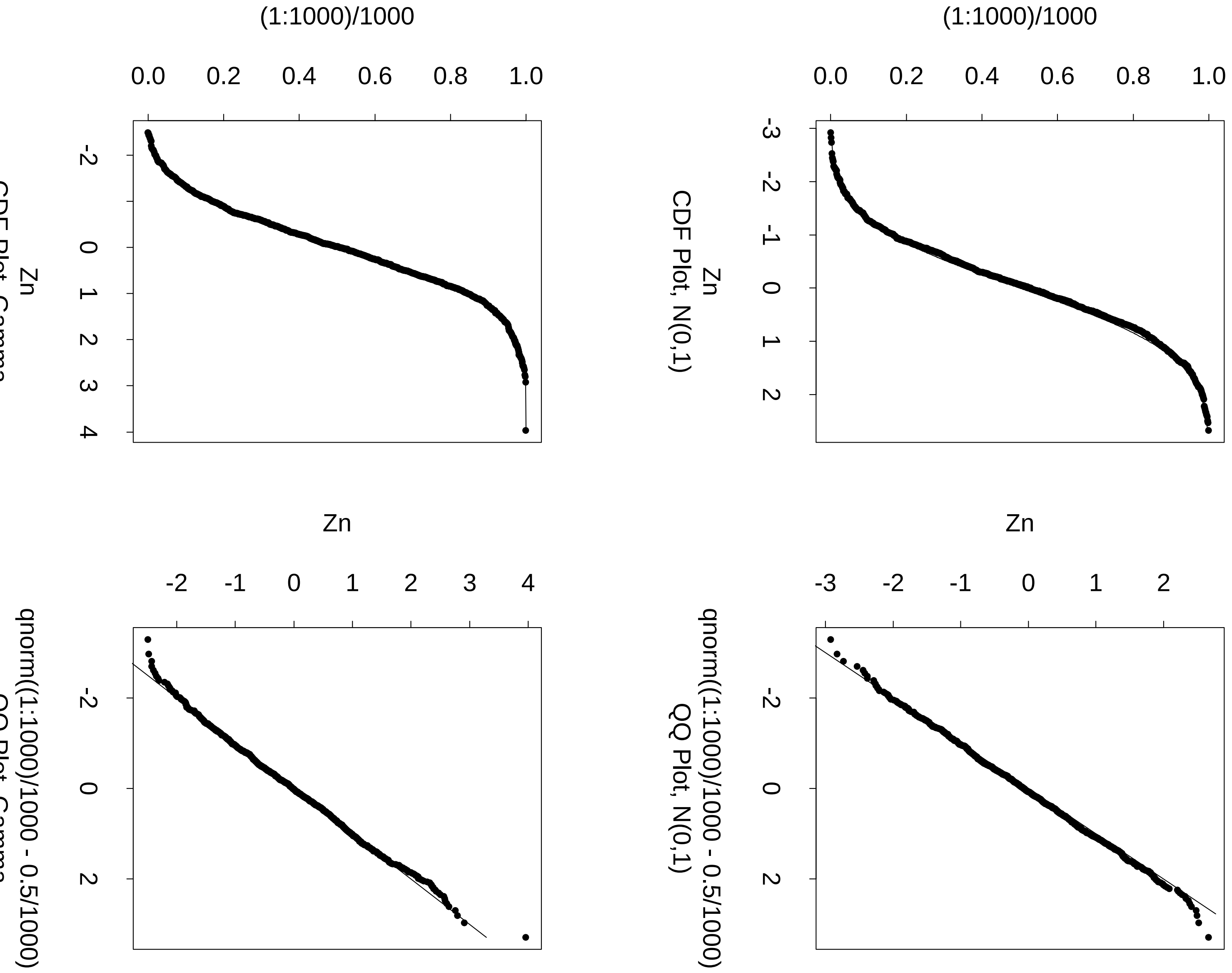


Figure 3.3: Distributions of the Sample Median

Exercises for Section 3.5.

Elementary and Review Exercises.

3.5.1 Verify equation (3.49).

3.5.2 Here we investigate why the QQ-plot tends to magnify differences between c.d.f.'s in the tails. Suppose F is a given c.d.f. with density f which is everywhere positive and we consider c.d.f.'s G satisfying

$$\sup_{-\infty<x<\infty} |G(x) - F(x)| \ \leq \ \epsilon$$

where ϵ is a small positive number. Given u satisfying $\epsilon < u < 1 - \epsilon$, put $x = F^{-1}(u)$.

(a) Given G in our class above, show that

$$F^{-1}(u - \epsilon) \ \leq \ G^{-1}(u) \ \leq \ F^{-1}(u + \epsilon) \quad .$$

(b) Argue that, approximately,

$$F^{-1}(u \pm \epsilon) \ \doteq \ x \pm \frac{\epsilon}{f(x)} \quad .$$

(c) Assuming $f(x) \to 0$ as $|x| \to \infty$, give an heuristic justification for the statement that the QQ-plot will magnify differences in the tails.

3.5.3 (a) Show that the QQ-plot of a $N(\mu, \sigma^2)$ vs. a $N(0, 1)$ will be a straight line, and determine the slope and intercept.

(b) Generalize the result of part (a) to an arbitrary location–scale family, i.e. a QQ-plot of $F_{a,b}$ vs. $F_{0,1}$ where $F_{a,b}(y) = F_{0,1}([y-a]/b)$. Here, $a \in I\!R$ is arbitrary and $b > 0$.

(c) Discuss the application of (b) to exponential distributions when the mean is unknown. How can one "read off" the mean from the QQ-plot?

Advanced Exercises.

3.5.4 (a) Let Y be a Bernoulli random variable with success probability $p = 1/2$, Z be a $N(0, 1)$ random variable independent of Y, and $\mu > 0$ be given. Put $X = Z + (2Y - 1)\mu$. Find the density for the distribution of X.

(b) Repeat the study of Example 3.5.2 when the observations are i.i.d. with the same distribution as X in part (a). Use $\mu = 2$.

3.5.5 Show that if one has available a function to compute the distribution of $B(n, p)$ for arbitrary (n, p) and to compute the c.d.f. of an individual observation X_i, then it is unnecessary to use Monte Carlo simulation in Example 3.5.2, i.e. one can compute $P[2f(m)\sqrt{n}(M_n - m) \leq z]$ similarly to Example 3.5.1.

3.5.6 Perform a study of the accuracy of the asymptotic approximation in Exercise 3.4.2.

3.5.7 Perform a study of the accuracy of the asymptotic approximation in Exercise 3.4.5.