Radically Elementary Probability and Statistics

By
Charles J. Geyer

Technical Report No. 657

School of Statistics
University of Minnesota

April 16, 2007

Radically Elementary Probability and Statistics

Charles J. Geyer

April 16, 2007

Contents

Preface

The book *Radically Elementary Probability Theory* (Nelson, 1987, Preface)

> is an attempt to lay new foundations for probability theory using a tiny bit of nonstandard analysis. The mathematical background is little more than that which is taught in high school, and it is my hope that it will make deep results from the modern theory of stochastic processes readily available to anyone who can add, multiply, and reason.

The "tiny bit of nonstandard analysis" is really "tiny" entailing little more than a rigorous notion of infinitesimals. Nelson (1987) introduces it in three brief chapters. We give a more belabored treatment (our Part I), but almost everything one needs to know about nonstandard analysis are the arithmetic rules for infinitesimal, appreciable, and unlimited numbers (a number is unlimited if its reciprocal is infinitesimal, and a number is appreciable if it is neither infinitesimal or unlimited) given in the tables in our Section 3.3, the principle of external induction — an axiom of the nonstandard analysis used in Nelson (1987) and this book (Axiom IV in Section 2.1) — and the principle of overspill (our Section 3.4).

With this tiny bit of nonstandard analysis in hand Nelson (1987) radically simplifies probability theory, adopting two principles (explained in our Section 5.1). All probability spaces have

(i) finite sample space and

(ii) no nonempty events of probability zero.

These have the consequence that expectations always exist and are given by finite sums, conditional expectations are unique and given by a simple formula, also involving a finite sum, our equation (5.5), and no measure theory is necessary.

One might think that this "radical" simplification is too radical — throwing the baby out with the bathwater — but Nelson (1987) and this book provide some evidence that this is not so. Even though our theory has no continuous random variables or even discrete random variables with infinite sample space, hence no normal, exponential, Poisson, and so forth random variables. We shall see that finite approximations satisfactorily take their place.

Consider a Binomial(n, p) random variable X such that neither p nor $1-p$ is infinitesimal and n is unlimited. Then (the Nelson-style analog of) the central limit theorem says that $(X-np)/\sqrt{p(1-p)/n}$ has a distribution that is "nearly normal" in the sense that the distribution function of this random variable differs from the distribution function of the normal distribution in conventional probability only by an infinitesimal amount at any point (our Theorem 8.2).

Consider a Binomial(n, p) random variable X such that p is infinitesimal but np is appreciable. Then X has a distribution that is "nearly Poisson" in the sense that the distribution function of this random variable differs from the distribution function of the Poisson(np) distribution in conventional probability only by an infinitesimal amount at any point (unfortunately this result is in a yet to be written chapter of this book, but is easily proved).

Consider a Geometric(p) random variable X such that p is infinitesimal and choose a (necessarily infinitesimal) number ϵ such that $Y = \epsilon X$ has appreciable expectation μ. Then Y has a distribution that is "nearly exponential" in the sense that the distribution function of this random variable differs from the distribution function of the Exponential$(1/\mu)$ distribution in conventional probability only by an infinitesimal amount at any point.

Thus although Nelson-style probability theory does not have continuous random variables, it does have discrete analogs that take their place. One is entitled to ask what is the point? Nelson makes one point in the text quoted above: the level of mathematical difficulty is much lower. In teaching probability at levels below Ph. D. we are used to starting with Kolmogorov's axioms and Venn diagrams and mutually exclusive events and — just when the students are thoroughly confused — dropping the whole subject. When expectation is introduced, it is entirely the nineteenth-century, pre-Kolmogorov notion. Many topics are discussed with allusions to Kolmogorov-style theory but without rigor because measure theoretic abstraction is deemed too difficult for the level of the course. Nelson-style theory avoids the necessity for all this handwaving. It provides a different kind of rigor that is understandable, perhaps (as Nelson claims) even by high school students.

We would not like to make the wrong impression. As Nelson (1987, p. 13) says, all mathematics involves abstractions. The notion of an unlimited number (which is larger than any number you could actually name) is just as much an abstraction as an infinite sequence. The question is which abstraction do we wish to use. If we choose to use actually infinite sequences, then measure theory is unavoidable. If we choose to use nonstandard analysis, then the law of large numbers (Nelson, 1987, Chapter 16) can be stated without reference to measure theory. The mathematical content is not *exactly* the same, but the same purposes are served. Both laws of large numbers (Kolmogorov and Nelson) serve equally well in applications, such as justifying statistical inference. Neither flavor of abstraction makes the proof of the theorem trivial. Charlie's law of conservation of mathematical difficulty says that, no matter what abstraction you use, at the end of the day you still have prove something is less than ϵ and that calculation is always essentially the same. Neither Nelson (1987) nor this book is easy reading. I admit that most American undergradu-

ates (never mind high school students) are not comfortable with mathematical arguments that run over several pages with many steps. Regardless of the level of abstraction, such arguments are unavoidable. Nevertheless Nelson (1987) is a remarkable *tour de force*. In 79 small format pages he goes from nothing to what he calls the de Moivre-Laplace-Lindeberg-Feller-Wiener-Lévy-Doob-Erdös-Kac-Donsker-Prohorov theorem, which

> is a version of the de Moivre-Laplace central limit theorem that contains Lindeberg's theorem on the sufficiency of his condition, Feller's theorem on its necessity, Wiener's theorem on the continuity of the trajectories for his process, the Lévy-Doob characterization of it as the only normalized martingale with continuous trajectories, and the invariance principle of Erdös and Kac as extended by Donsker and Prohorov.

Nelson ends by saying

> This is an arbitrary stopping point. More can be done. I hope that someone will write a truly elementary book on stochastic processes along these lines, complete with exercises and applications.

I cannot claim that this is that book. At least this book is "along these lines." We follow Nelson in using only "radically elementary" nonstandard analysis based on four axioms from Nelson (1987) (our Section 2.1) and only "radically elementary" probability theory based on principles (i) and (ii) above. We have filled in many details left untouched by Nelson, in particular the relationship of the central limit theorem mentioned above, which is the climax of Nelson (1987), to $x \mapsto \exp(-x^2/2)/\sqrt{2\pi}$. We also do weak convergence on arbitrary metric spaces, Prohorov metric, Lévy metric, the portmanteau theorem, Slutsky's theorem, the continuous mapping theorem, and the Glivenko-Cantelli theorem.

However, my interests are not in stochastic processes, except for spatial point processes and Markov chains, but in statistics. The original idea of this book was to rewrite statistics in "radically elementary" fashion, but I have only barely begun that task. The reason this incomplete draft is being turned into a technical report is so that the thesis of Bernardo Borba de Andrade, which deals with the "radically elementary" approach to Markov chains, can cite some of the results in the part of my planned book that now exists (this technical report), in particular, the relationship between characteristic functions and convergence in distribution (Chapter 11) that he uses to prove the central limit theorem for alpha-mixing stationary processes. I hope that some of the planned additions to this book will eventually be written.

Whether the "radically elementary" or "Nelson-style" approach is better or worse than the "conventional" or "Kolmogorov-style" approach is something we cannot yet say. I can only hope that a reader of Nelson (1987) and this book will concede that our approach has promise. It is a huge task to rewrite all of probability theory in the new style. Many people will not think it worth the

effort even if the new style is simpler, more elegant, and easier to understand. Moreover, it must be conceded that the new style is not always more elegant. In particular, I had a great deal of difficulty proving the Lévy continuity theorem (our Theorem 11.8) because characteristic functions are messier objects when only discrete random variables are allowed: compare the form (11.11) of the characteristic function of the discrete double exponential distribution to $t \mapsto 1/(1 + t^2/\alpha^2)$ of its continuous analog and the inelegant condition (11.16) that is required for the former to be well approximated by the latter. Other characteristic functions are worse. Generally one has no closed form expression, messy or otherwise.

The new style also takes a lot of getting used to. It is hard to abandon the unique normal distribution of conventional theory. In our new theory we call "normal" any distribution whose distribution function differs from the distribution function of the normal distribution in conventional probability only by an infinitesimal amount at any point. This huge class of distributions is very like the conventional normal distribution in some respects and very unlike it in others. In particular, the moments need be nothing like those of a conventional normal distribution. Even if we add the condition that our "Nelson-style" normal distributions be (Nelson-style) L^2 so that first and second moments agree (to within an infinitesimal amount) with those of the conventional normal distribution (our Theorem 10.7), higher moments need not agree.

Now is this a good thing or a bad thing? Originally, the normal distribution was not thought of as a real distribution but only as a limit arising in the de Moivre-Laplace theorem. Gradually, through the work of Gauss, Quetelet, and others the normal distribution was reified so we now think of it as a real distribution. But much of late twentieth century statistics, especially nonparametrics and robustness, had the objective of knocking the normal distribution off its pedestal. Is is part of the problem or part of the solution that $E(X^4) = 3\sigma^4$ for the normal distribution? Confidence intervals for the population variance based on the F distribution give incorrect results, even asymptotically incorrect, unless the population fourth central moment is exactly three times the population second central moment squared. We have no reason to believe that will be true in real applications. A well-known introductory textbook (Moore and McCabe, 1989, pp. 568–569) said

> The F test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice.

reproducing figures from Pearson and Please (1975) as evidence. So is the uniqueness of the normal distribution in conventional theory a benefit or a trap for the unwary? Here it seems to a trap. The theory of "exact" confidence intervals based on the F distribution (assuming "exact" normality) is elegant but worthless in application precisely because "exact" normality is too much to ask. In Nelson-style theory the unique continuous normal distribution goes back to being only a limit that is never reached and doesn't really exist, so we are not tempted to take it too seriously.

Many other cases could be considered where conventional theory "takes infinity too seriously." Marginalization paradoxes arising from the use of improper priors in Bayesian inference (Dawid, et al., 1973) are an example. Nelson-style improper priors do not exist because any positive measure on a finite sample space can be renormalized so that it is a probability measure. There are Nelson-style analogs to improper priors. For example, the uniform distribution on a grid of points that has infinitesimal spacing and extends to unlimited numbers in both positive and negative directions behaves in some respects just like Lebesgue measure as an improper prior. But since the Nelson-style "improper" prior is in fact proper (a probability distribution), no paradoxes can arise. Here the supposed simplicity of conventional theory is not simple at all. It leads only to confusion and incoherence. The extreme technical difficulty of deciding when an improper prior leads to sensible inference (Eaton, 1992) is notorious. In Nelson-style theory, the issue simply does not arise.

The same point has been made about using finitely additive probability theory by Sudderth (1980). Nelson-style "radically elementary" probability theory is in this respect and some other respects — especially in Nelson's definition of "almost surely" (see our Section 6.1) — much like finitely additive probability theory because, of course, if the sample space is finite, then countable additivity is vacuous. In other respects, Nelson-style theory is not much like finitely additive theory. As we have stressed, in Nelson-style theory the level of abstraction is much lower than conventional Kolmogorov-style theory. The level of abstraction of finitely additive theory tends to be much higher than conventional theory.

For a long time I have wondered why probability theory is the way it is. Why is based on Kolmogorov (1933)? As mentioned above, finitely additive theory has been studied, although nowhere near as intensively as the Kolmogorov-style countably additive theory. Kolmogorov himself (along with others) initiated other approaches, such as the so-called algorithmic complexity approach (Li and Vitanyi, 1997). There is a remarkable paper by Solovay (1970) which shows that is it possible to make set-theoretic assumptions (incompatible with the axiom of choice) so that every subset of the real numbers is Lebesgue measurable. So-called Loeb measure (Loeb, 1975) uses Robinson-style nonstandard analysis (see our Chapter 1 for brief explanation of that) to construct probability measures. The emphasis of probability over expectation in conventional theory can be reversed, as in Whittle (2005). Finally, as everyone is well aware, probability theory had nearly 300 years of history before 1933 in which many important theorems were proved without measure theory. So why do we do probability theory the way we do?

This is not just a question of the history and sociology of mathematics. In statistics, especially, theorems often only have heuristic value. The central limit theorem does not assure that normal approximation is good in any particular application, only that it will be good for some sufficiently large sample size, perhaps billions and billions of times larger than in any application. If it were really the case that sample sizes of billions and billions were required, then no one would care about the theorem, despite its beauty. Statisticians who worry

about the validity of normal approximation sometimes do simulations to check, and they find that it sometimes works well and sometimes not, depending on the details of the application. So the theorem does, at least sometimes, provide a useful guide to practice, even if it provides no guarantee. When we change the abstractions we use, we often change the atmospherics, hence we often change the heuristics, even though the mathematical assertions of the theorems are very similar. Nelson's central limit theorem is very similar to the conventional central limit theorem in what it says, but because the limiting distribution is not unique (it is unique up to near equivalence, of course), it feels different. The normal distribution, as we said above, doesn't really exist in Nelson-style theory. Hence we take some of its properties, such as its fourth moment, much less seriously.

I started working on "radically elementary" probability and statistics merely out of curiosity. I had to struggle to unlearn some ingrained habits from conventional theory. But I have been richly rewarded. Conventional theory looks different to me now. I have a broader perspective, a wider context. To me, that is even more important than simplicity and elegance.

Part I

Nonstandard Analysis

Chapter 1

Introduction

This part of the book introduces so-called "nonstandard analysis," which is the branch of mathematics that makes rigorous use of infinitesimals. The early history of infinitesimals starts with the invention of calculus by Newton and Leibniz in the late seventeenth century. From then until the late nineteenth century all calculus and analysis (the mathematician's term for advanced calculus, as in the title of Cauchy's *Cours d'Analyse de l'École Royale Polytechnique*) used infinitesimals, although questions about rigor persisted throughout this period. In the late nineteenth century, the delta-epsilon definitions of limits, continuity, derivatives, and integrals still in use today provided the first rigorous foundations for calculus. From then until the middle of the twentieth century infinitesimals were banished for lack of rigor.

The reasons for banishment of infinitesimals disappeared with the publication of Robinson (1996, first edition 1966). Interestingly, Robinson's work grew out of the process of formalization of mathematics that made infinitesimals seem nonrigorous in the first place. What goes around comes around.

Model theory is the branch of mathematical logic that deals with models for axiomatic systems. Given a set of axioms, a *model* for those axioms is a mathematical object that satisfies the axioms. If the axioms in question are those for a vector space, then a model is just a vector space. This may seem trivial, but it is profound. For one thing, the existence of a model means the axioms are consistent, meaning they imply no contradiction (statements that follow from the axioms and contradict each other), because no model can satisfy a contradiction. For another thing, it leads to the question of what sort of models can satisfy the axioms. Again, if the axioms in question are those for a vector space, then the question is what vector spaces exist and what are their properties? So this question encompasses the whole subject of linear algebra.

If the axioms in question are those for the real numbers, then the question is what real number systems can exist? We are taught in conventional mathematics about *the* real number system. But model theoretic study of the axioms for the real numbers shows that uniqueness of the real numbers cannot be proved, and that "nonstandard" models exist. Robinson exploited this fact to bring

infinitesimals back into rigorous mathematics. He constructed, in conventional mathematics, nonstandard models of the real numbers having elements that could be interpreted as infinitesimals. For a recent, fairly elementary, treatment of such model-theoretic constructions, see Goldblatt (1998).

Nelson (1977) introduced axiomatic nonstandard analysis in which the behavior of nonstandard concepts is derived from axioms rather than by explicit construction of models. In some ways this is simpler than the approach of Robinson (1996), since it requires no knowledge of model theory. Many other axiomatic theories of nonstandard analysis have since been developed. Gordon, Kusraev and Kutateladze (2002) and Kanovei and Reeken (2004) give encyclopedic coverage of them. These various versions of nonstandard analysis can be considered different formalizations of the same subject. Although axiomatic nonstandard analysis following Nelson (1977) requires no model theory, it does require a lot of set theory. In this book we are going to use a very simple version of nonstandard analysis found in Nelson (1987), which requires very little set theory.

In this version of nonstandard analysis, we consider the real number system we use to be the same as the real number system of conventional mathematics in the sense that every conventional mathematical theorem about it remains the same. Those theorems are silent about infinitesimals. They neither prove infinitesimals exist nor prove they do not exist. Hence we may assume that infinitesimals do exist and are real numbers just like any other real numbers.

We might worry that this will cause trouble, but we shall be very careful. Our mathematics with infinitesimals will be based on axioms (taken from Nelson, 1987), which are theorems of the version of nonstandard analysis in Nelson (1977); see p. 80 in Nelson (1987). A theorem in Nelson (1977), attributed to Powell, asserts that the theory of that paper and conventional mathematics (as formalized in Zermelo-Frankel set theory with the axiom of choice, usually abbreviated ZFC) are equiconsistent (meaning both theories are consistent or both are inconsistent). Since the celebrated inconsistency theorem of Gödel says that the consistency of ZFC cannot be proved within ZFC, this is all that can be said. Our nonstandard analysis is consistent if conventional mathematics is itself consistent.

Chapter 2

The Natural Numbers

The objects of this version of nonstandard analysis are the objects of conventional mathematics: the integers, the real numbers, and so forth. No new objects are added. Nothing in conventional mathematics is changed. Every theorem of conventional mathematics remains true.

In particular, the natural numbers, the set $\mathbb{N} = \{0, 1, 2, \ldots\}$, remains the same as it is in conventional mathematics.

2.1 Axioms

Our only addition to conventional mathematics is an undefined property *standard* and four axioms that govern the use of this property.

(I) 0 is standard.

(II) If $n \in \mathbb{N}$ is standard, then so is $n + 1$.

(III) There exists an $n \in \mathbb{N}$ that is not standard.

(IV) If A is any property such that $A(0)$ holds and $A(n) \to A(n + 1)$ holds for all standard $n \in \mathbb{N}$, then $A(n)$ holds for all standard $n \in \mathbb{N}$.

We define *nonstandard* to mean not standard. Hence we usually say axiom (III) asserts the existence of a nonstandard $n \in \mathbb{N}$. We also define *limited* as a synonym of standard and *unlimited* as a synonym of nonstandard. (In the next chapter these terms will become nonsynonymous when limited and unlimited can be applied to real numbers and standard and nonstandard cannot.) We could replace "standard" everywhere it appears in the axioms by "limited," but we do not do so, partly to follow Nelson (1987) and partly to distinguish between the undefined term "standard" that has no meaning other than that given by its appearance in these axioms and the term "limited" which is defined in terms of "standard."

The complicated axiom (IV) is called *external induction*. We also, of course, inherit conventional mathematical induction from conventional mathematics.

To explain the difference, we need the following terminology. A property from conventional mathematics (defined without any reference, direct or indirect, to the property "standard") is called *internal*. A property that is not internal is *external*. Our only examples of external properties so far are *standard*, *nonstandard*, *limited*, and *unlimited*, but later on we shall meet *infinitesimal*, *nearly continuous*, and many more. Following Nelson, we extend this terminology to mathematics itself and sometimes say *internal mathematics* instead of "conventional mathematics" and *internal induction* instead of "conventional mathematical induction." Internal induction is the following, which is a theorem of set theory.

(V) If A is any internal property such that $A(0)$ holds and $A(n) \to A(n+1)$ holds for all $n \in \mathbb{N}$, then $A(n)$ holds for all $n \in \mathbb{N}$.

Note that internal induction can only be applied to internal properties (conventional mathematical induction can only be applied to conventional mathematical properties). If (V) could be applied with $A(n)$ meaning "n is standard," then together with axioms (I) and (II) it would imply that every natural number is standard, but that would contradict axiom (III), and our axioms would be inconsistent when combined with those of conventional mathematics (ZFC). Note that external induction (IV) applied with $A(n)$ meaning "n is standard" only produces the tautology that every standard n is standard.

So these axioms produce no immediately obvious contradiction. As was mentioned at the end of the last chapter, our theory is consistent if conventional mathematics is consistent. Thus we need not worry about contradictions, obvious or not.

Theorem 2.1. *The numbers 1, 2, ..., 10 are limited.*

The proof is obvious. Axioms (I) and (II) together imply that 1 is limited. Having established that, another application of axiom (II) shows that 2 is limited. Then another shows that 3 is limited. And so forth. Clearly the process need not stop at 10. The theorem would be just as true if we replaced 10 by a thousand or a million. (There would just be more steps to the proof.) We cannot, however, strengthen Theorem 2.1 by removing the upper bound entirely. That would conflict with axiom (III).

2.2 Arithmetic

If we want to increase the upper bound in Theorem 2.1 we can make our proof more efficient by using external induction.

Theorem 2.2. *If m and n are limited natural numbers, then so is $m + n$.*

Proof. Fix a limited $m \in \mathbb{N}$ and let $A(n)$ be the property "$m + n$ is limited." By axiom (II) and external induction $A(n)$ holds for all limited n. □

Now we can count limited numbers a little faster: $10 + 10 = 20$ is limited, $20 + 20 = 40$ is limited, and so forth.

Theorem 2.3. *If m and n are limited natural numbers, then so is $m \cdot n$.*

Proof. Fix a limited $m \in \mathbb{N}$ and let $A(n)$ be the property "$m \cdot n$ is limited." By Theorem 2.2 and external induction $A(n)$ holds for all standard n. □

Now we can count limited numbers a faster still: $10 \cdot 10 = 100$ is limited, $100 \cdot 100 = 10,000$ is limited, and so forth.

Theorem 2.4. *If m and n are limited natural numbers, then so is m^n.*

Proof. Fix a limited $m \in \mathbb{N}$ and let $A(n)$ be the property "m^n is limited." By Theorem 2.3 and external induction $A(n)$ holds for all standard n. □

Now we can count limited numbers much faster: 10^{10} is limited, $10^{10^{10}}$ is limited, and so forth. We could accelerate this process further with further applications of external induction, but we have run out of familiar mathematical operations (what comes after addition, multiplication, exponentiation?) and so will be content to stop here.

2.3 Order

We now fill in the gaps in our counting.

Theorem 2.5. *If n and ν are natural numbers, n limited and ν unlimited, then $n < \nu$.*

Proof. Fix an unlimited ν, and let $A(n)$ be the property $n < \nu$. Since ν is unlimited, it is not zero, hence $A(0)$ holds. If n is limited and $A(n)$ holds, then $n + 1$ is limited and $n + 1 \leq \nu$, and since equality is impossible when $n + 1$ is limited and ν is unlimited, we actually have $A(n + 1)$. Thus external induction implies $n < \nu$ for all limited n. □

Hence we could now improve Theorem 2.1 to say that 1, 2, ..., $10^{10^{10}}$ are limited, but we won't bother with a theorem number and a formal statement.

2.4 Illegal Set Formation

It is a principle of internal mathematics that properties can be used to define sets (the *subset axiom* or the *axiom of specification* of ZFC).

(VI) For any internal property A and any set S, there exists a set B such that $x \in B$ if and only if $x \in S$ and $A(x)$.

The set B is unique (by the *axiom of extension* of ZFC) and is usually denoted

$$\{\, x \in S : A(x) \,\} \tag{2.1}$$

But nothing in internal mathematics says that *external* properties can be used to define sets in this way (internal mathematics having no external properties).

Moreover, none of our four axioms of nonstandard analysis allows the use of (2.1) when A is an external property. Nelson calls attempts to use (2.1) with A external *illegal set formation*.

This explains why we only distinguish between internal and external *properties* and not between internal and external *objects*. From a foundational point of view, everything is a set. Natural numbers are sets: zero is just another name for the empty set, one is the set $\{0\}$, two is the set $\{0, 1\}$, and so forth. Signs are identified with $\{0, 1\}$ and integers with ordered pairs of a sign and a natural number. Rational numbers are identified with ordered pairs of integers. Real numbers are identified with Dedekind cuts of rationals (pairs of sets of rationals). Functions are identified with their graphs, which are subsets of the Cartesian product of the domain and range. And so forth. The rule against illegal set formation disallows the use of (2.1) when A is external. Thus there is no way to get any new sets that are not already present in internal mathematics, hence no new mathematical objects of any sort. All *objects* are *internal*. Only *properties* can be *internal* or *external*.

This is a bit confusing; one must make an effort to keep this distinction clear. When we say a natural number ν is unlimited, we are asserting $A(\nu)$ holds where A is the external property defined by $A(n)$ means "n is unlimited," but ν itself is an internal object. From a foundational point of view, ν is the set $\{0, \ldots, \nu - 1\}$ and the principles of set theory allow us to form this set for any natural number ν.

The rule about illegal set formation is very important because ignoring it is like money laundering: it destroys the distinction between internal and external properties and internal and external induction. If A is a property and B is a set such that $A(x)$ holds if and only if $x \in B$, then the property A' defined by $A'(x)$ meaning $x \in B$ is equivalent to A and is an internal property, because B is an internal set (all sets being internal) and $\in$ is an internal relation.

Theorem 2.6. *There does not exist a subset B of $\mathbb{N}$ such that $n \in B$ if and only if n is limited.*

Proof. Suppose the set B described by the theorem does exist. Then the property A defined by $A(n)$ means "$n \in B$" is an internal property, and we can apply internal induction to it. Axiom (I) implies $A(0)$, and Axiom (II) implies $A(n)$ implies $A(n) \to A(n + 1)$. We conclude $A(n)$ holds for all $n \in \mathbb{N}$, which says that every natural number is limited, but that contradicts Axiom (III). Hence no such set exists. □

The rule about illegal set formation does not deny existence; it merely denies a particular *justification* of existence. Let $A(n)$ be the property "n is limited and $n < 6$." At first sight A appears to be an external property because it involves the external property "limited." However, we know from Theorem 2.1 that the elements of the set $B = \{0, 1, 2, 3, 4, 5\}$ are all limited and thus we see that A is equivalent to the internal property A' defined by $A'(n)$ meaning either $n \in B$ or $n < 6$.

Thus even though the rule about illegal set formation forbids us to use the subset axiom as a justification of the existence of this set, it nevertheless exists. If we want to show that no set corresponds to an external property, then we need a theorem (like Theorem 2.6). Without a theorem proving either existence or non-existence, we do not know whether any set corresponds to an external property. All we know is that there is no rule that asserts it *automatically* exists and that we can't just blithely use the notation (2.1) as if there were such a rule.

Chapter 3

Real Numbers

Any real number x lies between two integers $\lfloor x \rfloor$ and $\lceil x \rceil$, which are called the *floor* and *ceiling* of x and which are the greatest integer less than or equal to x and the least integer greater than or equal to x, respectively. If x is an integer, then $\lfloor x \rfloor = \lceil x \rceil$. Otherwise, $\lfloor x \rfloor + 1 = \lceil x \rceil$.

Note that any two consecutive natural numbers n and $n+1$ are either both limited or both unlimited. If n is limited, then so is $n+1$ by Axiom II. If n is unlimited, then so is $n+1$, by Theorem 2.5.

Thus we extend the notions of limited and unlimited to real numbers as follows.

- A nonnegative real number x is *limited* if and only if $\lfloor x \rfloor$ and $\lceil x \rceil$ are standard.
- A negative real number x is *limited* if and only if $|x|$ is limited.

Unlimited means not limited.

We use the concept "limited" to define a new concept "infinitesimal."

- 0 is *infinitesimal*.
- A nonzero real number x is *infinitesimal* if and only if $1/x$ is unlimited.

Another useful auxiliary concept is "appreciable."

- A real number is *appreciable* if it is limited and not infinitesimal.

It is important to remember that *limited*, *unlimited*, *infinitesimal*, *non-infinitesimal*, and *appreciable* are *external properties*.

Theorem 3.1. *A nonzero real number x is infinitesimal if and only if $1/x$ is unlimited and is appreciable if and only if $1/x$ is appreciable.*

Proof. The first assertion just restates the definition of infinitesimal. The second assertion then follows because when x is appreciable $1/x$ can be neither infinitesimal nor unlimited. □

With these definitions come new notation. In the following x and y are any real numbers.

- $x \simeq y$ means $x - y$ is infinitesimal.
- $x \ll y$ means $x \leq y$ and $x \not\simeq y$.
- $x \lesssim y$ means $x \leq y$ or $x \simeq y$.
- $x \simeq \infty$ means x is positive and unlimited.
- $x \ll \infty$ means $x \not\simeq \infty$.
- $x \gg y$ means $y \ll x$.
- $x \gtrsim y$ means $y \lesssim x$.
- $x \simeq -\infty$ means $-x \simeq \infty$.
- $x \gg -\infty$ means $-x \ll \infty$.

And it is important to remember that all of the symbolic notations above express *external properties*.

These notations are read as follows.

- $x \simeq y$ is read *x and y are nearly equal*.
- $x \ll y$ is read *x is appreciably less than y* or *x is strongly less than y*.
- $x \lesssim y$ is read *x is weakly less than y*.

Please note, to avoid any misunderstanding, that the notation $x \simeq \infty$ does not (despite appearances) assert that there are two objects x and ∞ that are nearly equal. It is just shorthand for *x is positive and unlimited* and says nothing about an object ∞, not even that such an object exists.

3.1 Order

Theorem 3.2. *If ξ, x, ϵ, y, and η are real numbers, ξ is negative and unlimited, x is negative and appreciable, ϵ is infinitesimal, y is positive and appreciable, and η is positive and unlimited, then $\xi < x < \epsilon < y < \eta$.*

Proof. $y < \eta$ is immediate from Theorem 2.5 and the definition of limited real numbers. Then $\epsilon < y$ follows from the fact that for positive x and y we have $x < y$ if and only if $1/y < 1/x$ (and from nonpositive numbers being less than positive numbers). The other inequalities follow from $0 < x < y$ if and only if $-y < -x < 0$. □

3.2 Arithmetic

Theorem 3.3. *If x and y are real numbers, then $x + y$ is limited if both x and y are limited, and $|x| + |y|$ is limited only if both x and y are limited.*

Proof. The forward part follows from Theorem 2.2 by the order properties of addition and the definitions of limited and unlimited real numbers. The converse part follows from $\max(|x|, |y|) \leq |x| + |y|$. □

Theorem 3.4. *If x and y are real numbers, then $x \cdot y$ is limited if both x and y are limited.*

Proof. This follows from Theorem 2.3 by the order properties of multiplication and the definitions of limited and unlimited real numbers. □

Corollary 3.5. *If $x_1, \ldots, x_n$ are limited real numbers and n is a limited natural number, then $x_1 + \cdots + x_n$ is limited and $x_1 \times \cdots \times x_n$ is limited.*

Proof. Apply external induction to Theorem 3.3 or Theorem 3.4. □

Theorem 3.6. *If x and y are real numbers, then $x \cdot y$ is unlimited if x is non-infinitesimal and y is unlimited.*

Proof. Suppose to get a contradiction that $z = x \cdot y$ is limited. Then, by Theorem 3.4, $y = z \cdot (1/x)$ is limited. □

Theorem 3.7. *If x and y are real numbers, then $x + y$ is infinitesimal if both x and y are infinitesimal, and $|x| + |y|$ is infinitesimal only if both x and y are infinitesimal.*

Proof. The direct part is trivial when either x or y is zero, and because of $|x + y| \leq |x| + |y|$ we may assume without loss of generality that x and y are positive, in which case we have

$$\frac{1}{x+y} \geq \frac{1}{2} \cdot \frac{1}{\max(x,y)} \simeq \infty$$

by the definition of infinitesimal and Theorem 3.6.

The converse part follows from $\max(|x|, |y|) \leq |x| + |y|$ and Theorem 3.2. □

Corollary 3.8. *If $x_1, \ldots, x_n$ are infinitesimal real numbers and n is a limited natural number, then $x_1 + \cdots + x_n$ is infinitesimal.*

Proof. Apply external induction to Theorem 3.7. □

Corollary 3.9. *The external relation $\simeq$ is an equivalence.*

An equivalence relation is symmetric, reflexive, and transitive. The corollary asserts that $\simeq$ has these properties. We emphasize that it is an *external* relation. Hence $\{\, (x, y) \in \mathbb{R} \times \mathbb{R} : x \simeq y \,\}$ and $\{\, x \in \mathbb{R} : x \simeq y \,\}$ are illegal set formation. We can prove facts about $\simeq$ but we cannot consider it a mathematical object (that is, a set), nor can we define objects, such as equivalence classes, in terms of it.

Proof. Symmetry, $x \simeq x$ for all x, is obvious. Reflexivity, $x \simeq y \longrightarrow y \simeq x$, is also obvious. Transitivity, $x \simeq y$ and $y \simeq z \longrightarrow x \simeq z$, follows from the sum of infinitesimals is infinitesimal (Theorem 3.7). □

Theorem 3.10. *If x and y are real numbers, then $x \cdot y$ is infinitesimal if x is infinitesimal and y is limited.*

Proof. This is trivial when either x or y is zero, and because of $|xy| = |x| \cdot |y|$ we may assume without loss of generality that $0 < x \leq y$, in which case we have

$$\frac{1}{x \cdot y} = \frac{1}{x} \cdot \frac{1}{y} \simeq \infty$$

by Theorem 3.6 because $1/x$ is unlimited and $1/y$ is non-infinitesimal. □

Theorem 3.11. $\exp(x) \ll \infty$ *if and only if* $x \ll \infty$.

Proof. Since $\exp(x) \leq 1$ when $x \leq 0$ and $\exp(x) \leq 3^{\lceil x \rceil}$ when $x \geq 0$, the "if" direction follows from Theorems 2.4 and 3.2.

The "only if" direction follows from $\exp(x) \geq 1 + x$ (which is obvious from the Maclaurin series for exp). □

Theorem 3.12. $\exp(x) \simeq 1$ *if and only if* $x \simeq 0$.

Proof. Suppose $x \simeq 0$. By the law of the mean there exists an x^* between 0 and x such that

$$\exp(x) - \exp(0) = \exp(x^*) \cdot x.$$

By Theorems 3.2, 3.10 and 3.11 the right hand side is infinitesimal. That proves one direction.

Suppose $x \simeq 1$. By the law of the mean there exists an x^* between 1 and x such that

$$\log(x) - \log(1) = \frac{1}{x^*} \cdot (x - 1)$$

By theorems 3.1, 3.2, 3.10 and 3.11 the right hand side is infinitesimal. That proves the other direction. □

3.3 Summary of Arithmetic

Robert (1988) summarizes the arithmetic of nonstandard analysis in several tables. This seems like a good idea, and we have copied it, making some modifications.

Let δ and ϵ be infinitesimal real numbers, u and v appreciable real numbers, and X and Y unlimited real numbers. Then we have the following results about addition (and subtraction).

infinitesimal	appreciable	unlimited
δ, ϵ	u, v	X, Y
$\delta + \epsilon$	$\delta + u$, $\lvert u\rvert + \lvert v\rvert$	$\delta + X$, $u + X$
$u + v$		$\lvert X\rvert + \lvert Y\rvert$
$X + Y$		

The wide boxes indicate lack of more specific information. We only know that $u + v$ is limited. It may be infinitesimal (for example, if $v = -u$). And we can't say anything about $X + Y$. It may be infinitesimal, appreciable, or unlimited.

And we have the following results about multiplication and division.

infinitesimal	appreciable	unlimited
δ, ϵ	u, v	X, Y
$\delta \cdot \epsilon$, $\delta \cdot u$	$u \cdot v$	$u \cdot X$, $X \cdot Y$
δ/u, δ/X, u/X	u/v	u/δ, X/u, X/δ
$\delta \cdot X$		
δ/ϵ, X/Y		

(Again the wide box contains results we can't say anything about in general.)

When we consider exponentiation, the identities

$$x^{-y} = \left(\frac{1}{x}\right)^y = \frac{1}{x^y}$$

allow us to calculate x^y for all positive real x and all real y if we are only given x^y for $x \geq 1$ and $y \geq 0$. Thus we make a table only covering that case. If δ, ϵ, u, v, X, and Y are all nonnegative, we have the following results about powers.

infinitesimal	appreciable	unlimited
δ, ϵ	u, v	X, Y
result $\simeq 1$	$1 \ll$ result $\ll \infty$	result $\simeq \infty$
$(1+\delta)^\epsilon$, $(1+\delta)^u$, $(1+u)^\delta$	$(1+u)^v$	$(1+u)^X$, X^u, X^Y
$(1+\delta)^X$, X^δ		

Exercise 3-1. Verify all entries in the summary tables about addition, subtraction, multiplication, and division. For "wide boxes" not only show that the result is in the wide box but also give examples showing that the result can be in each part of the wide box.

Exercise 3-2. Prove the assertions of each row of the following table: x has the external property in the left column if and only if $\exp(x)$ has the external property in the right column and same row.

x	$y = \exp(x)$
$x \geq 0$ and $x \simeq 0$	$y \geq 1$ and $y \simeq 1$
$0 \ll x \ll \infty$	$1 \ll y \ll \infty$
$x \simeq \infty$	$y \simeq \infty$

Also state and prove the analogous theorem about x and $\log(x)$.

Exercise 3-3. Use the results of Exercises 3-1 and 3-2 to verify the summary table about exponentiation. As in Exercise 3-1, provide examples showing that results in the wide box can be either infinitesimal, appreciable, or unlimited.

3.4 Overspill

Illegal set formation is not just a complication that makes nonstandard analysis hard to understand. It is involved in an important proof technique called *overspill*.

We have divided the real numbers into "parts" with external properties, but these parts are not sets.

Theorem 3.13. *There does not exist a set that contains all and only the ____ real numbers, where the blank is filled in with any of the following: limited, unlimited, infinitesimal, non-infinitesimal, appreciable, or any of these modified by positive or negative.*

Proof. If the limited reals constituted a set L, then $L \cap \mathbb{N}$ would be the set B that Theorem 2.6 asserts does not exist. If the unlimited reals constituted a set U, then its complement would be L, which doesn't exist. If the infinitesimals constituted a set I, then $\{\, x \in \mathbb{R} \setminus \{0\} : 1/x \in I \,\}$ would be U, which doesn't exist. If the non-infinitesimal reals constituted a set, then its complement would be I, which doesn't exist. If the appreciable reals constituted a set A, then $[-1, 1] \setminus A$ would be I, which doesn't exist.

If any of these modified by positive or negative constituted a set S, then $S' = \{\, -x : x \in S \,\}$ would be the same modified by negative or positive, respectively, and $S \cup S'$ or (for infinitesimals) $S \cup S' \cup \{0\}$ would be a set that we have already proved does not exist. □

Now we explain *overspill*. Suppose A is an *internal* property such that $A(x)$ holds for all x in any one of the "non-sets" mentioned in the theorem, for concreteness, say the infinitesimals, that is, we are assuming $A(x)$ holds for all infinitesimal x. Then because A is internal,

$$B = \{\, x \in \mathbb{R} : A(x) \,\} \tag{3.1}$$

is not illegal set formation. Now we know by assumption that $A(x)$ holds for all infinitesimal x. If it held *only* for infinitesimal x, then B would be the set I which the proof shows does not exist. We conclude that $A(x)$ holds for some non-infinitesimal x. In picturesque language, A spills over from infinitesimals to non-infinitesimals. In short, if $A(x)$ holds for all infinitesimal x, then *by overspill*, it holds for some non-infinitesimal x (the words "by overspill" alluding to Theorem 3.13).

There are three important things to note about this technique. The key issue is that A must be *internal*. This method of proof is completely bogus

when applied to an external property, because then its starting point (3.1) is illegal set formation.

The second issue is that the scope of the technique is not limited to the infinitesimals. As we hinted above, any of the "non-sets" mentioned in the theorem will work just as well, for example, if $A(x)$ holds for all unlimited x, then, by overspill, it holds for some limited x.

The third issue is that the technique works when the free variable ranges over the natural numbers, in which case "by overspill" alludes to Theorem 2.6 instead of Theorem 3.13. For example, if $A(n)$ is an internal property that holds for all unlimited n, then it holds for some limited n.

As simple examples of overspill, we have the following.

$$x \simeq 0 \longleftrightarrow (\forall \epsilon \gg 0)(|x| \leq \epsilon)$$
$$x \simeq \infty \longleftrightarrow (\forall y \ll \infty)(y \leq x)$$

To see the first one, let $A(\epsilon)$ be the property $|x| \leq \epsilon$. If x is infinitesimal, then $A(\epsilon)$ holds for every $\epsilon \gg 0$ by Theorem 3.2. Conversely, if $A(\epsilon)$ holds for every $\epsilon \gg 0$, then, *by overspill*, it also holds for some infinitesimal ϵ, which implies x is infinitesimal (by Theorem 3.2 every number smaller than an infinitesimal is infinitesimal).

Chapter 4

Calculus

This chapter isn't really about calculus, despite its title. What it is about, is concepts from nonstandard analysis that compete with concepts from calculus and real analysis.[1]

4.1 Convergence

A sequence $x_1, x_2, \ldots$ of nonrandom real numbers *nearly converges* to a real number x if

$$x_n \simeq x, \qquad n \simeq \infty \tag{4.1a}$$

(Nelson, 1987, Chapter 6).

At first sight, this appears very different from the conventional notion of convergence. But the following definition, which looks more like the conventional notion, is equivalent. A sequence ... *nearly converges* ... if

$$(\forall \epsilon \gg 0)(\exists N \ll \infty)(\forall n \geq N)(|x_n - x| \leq \epsilon). \tag{4.1b}$$

The reason these two definitions are equivalent is because they are both equivalent to a sequence ... *nearly converges* ... if

$$(\forall n \geq N)(|x_n - x| \leq \epsilon), \qquad \epsilon \gg 0,\ N \simeq \infty. \tag{4.1c}$$

That (4.1a) implies (4.1c) is clear. For the opposite direction, fix an $n \simeq \infty$. Then (4.1c) implies $|x_n - x| \leq \epsilon$ holds for all $\epsilon \gg 0$ and hence for some $\epsilon \simeq 0$ by overspill. Hence (4.1a) holds for this n and (since $n \simeq \infty$ was arbitrary) for all unlimited n.

[1] It is possible to redo calculus and real and functional analysis using nonstandard analysis, but "radically elementary nonstandard analysis" is not the right vehicle. In order to define derivatives and integrals as mathematical objects, one needs what we have called the "full theory" Nelson's IST. Nelson (1977) and Gordon, et al. (2002, Chapter 2) give brief sketches of that theory. Robert (1988) gives a more complete discussion covering all of elementary calculus.

That (4.1b) implies (4.1c) is also clear. For the opposite direction, fix an $\epsilon \gg 0$. Then (4.1c) implies $(\forall n \geq N)(|x_n - x| \leq \epsilon)$ holds for all $N \simeq \infty$ and hence for some $N \ll \infty$ by overspill. Hence (4.1b) holds.

Two more issues of note should be mentioned. First, the limit is not unique. Indeed (4.1a) and $x \simeq y$ implies by Corollary 3.9 that x_n nearly converges to y. Second, and this is what is important for Nelson-style probability theory, the concept applies quite nicely to *finite* sequences.

Nelson (1987), after pointing out that near convergence is not exactly the same as ordinary convergence, drops the "near" because ordinary convergence is of no interest in Nelson-style probability theory. We will follow his lead. When we say "convergent," we always mean "nearly convergent" unless it is clear from the context that we are talking about the notion from conventional mathematics, although sometimes we say "nearly convergent" for emphasis.

An important aspect of this notion of convergence is that in many contexts it makes sequences irrelevant. A sequence x_n is (nearly) convergent if $x_m \simeq x_n$ for all unlimited m and n. So it is enough to understand (the external equivalence relation) near equality. We never need to deal with the whole sequence; we always deal with two elements at a time (is $x_m \simeq x_n$ or not?)

In some respects, near convergence is very different from conventional convergence. Consider a double sequence x_{ij} and suppose

$$\begin{aligned} x_{ij} &\to y_j, \qquad i \to \infty \\ x_{ij} &\to z_i, \qquad j \to \infty \end{aligned}$$

If we are talking about the conventional notion of convergence, this does not imply anything about joint convergence, the behavior of $x_{i_n j_n}$ as $n \to \infty$ for arbitrary subsequences i_n and j_n. But if we are talking about near convergence, then the situation is very different;

$$\begin{aligned} x_{ij} &\simeq y_j, \qquad i \simeq \infty \\ x_{ij} &\simeq z_i, \qquad j \simeq \infty \end{aligned}$$

implies

$$x_{ij} \simeq x_{mn}, \qquad i, j, m, n \simeq \infty$$

because

$$x_{ij} \simeq y_j \simeq x_{mj} \simeq z_m \simeq x_{mn}$$

whenever all of the indices are unlimited: that's just how equivalence relations (Corollary 3.9) behave.

Thus we see that near convergence and conventional convergence are in some respects similar, but in other respects near convergence is a much stronger and more useful property.

4.2 Continuity

If S is a subset of $\mathbb{R}$, then a function $g : S \to \mathbb{R}$ is *nearly continuous at the point* $x \in S$ if

$$(\forall y \in S)\big(x \simeq y \longrightarrow g(x) \simeq g(y)\big) \tag{4.2}$$

and *nearly continuous on a set* $T \subset S$ if (4.2) holds for all $x \in T$.

As with near convergence, these concepts are analogous to some conventional concepts. Near continuity at a point x is analogous to conventional continuity at x, and near continuity on a set T is analogous to conventional *uniform continuity on* T.

For example, the function $g : x \mapsto 1/x$ is continuous (but not uniformly continuous) on $(0, \infty)$, and for infinitesimal positive ϵ we have $g(\epsilon) - g(2\epsilon) = 1/2\epsilon$, which is unlimited, so g is not nearly continuous on $(0, \infty)$.

Conversely, if g is nearly continuous on T, then for any $\epsilon \gg 0$

$$(\forall x \in T)(\forall y \in T)\big(|x - y| \le \delta \longrightarrow |g(x) - g(y)| \le \epsilon\big) \tag{4.3}$$

holds for all infinitesimal δ, hence, by overspill, for some non-infinitesimal δ. Thus for every $\epsilon \gg 0$ there exists a $\delta \gg 0$ such that (4.3) holds, which looks like the conventional notion of uniform continuity (but isn't exactly because it has $\gg$ where the conventional notion has $>$).

We need to know that familiar functions from calculus are nearly continuous. The following lemma says they are.

Lemma 4.1. *Suppose g is a differentiable function $T \to \mathbb{R}$, where T is an open interval, and g' is limited on T. Then g is nearly continuous on T.*

Proof. For x and y in T, by the mean value theorem,

$$g(y) - g(x) = g'(\xi)(y - x)$$

for some ξ between x and y. By assumption $g'(\xi)$ is limited, so if $y - x \simeq 0$ we have $g(y) - g(x) \simeq 0$ by Theorem 3.10. □

This lemma is not sharp, because functions that are nearly continuous but not differentiable are easily defined. Moreover, even if g is differentiable, we do not need g' to be everywhere limited in order for g to be continuous. The lemma does, however, handle familiar "nice" functions. For example, it implies that the sine and cosine functions and the identity function $x \mapsto x$ are nearly continuous on all of $\mathbb{R}$.

Some care is required. For example, the exponential function is its own derivative, and Theorem 3.11 says $e^x \ll \infty$ if and only if $x \ll \infty$. Hence the lemma only implies that $x \mapsto e^x$ is nearly continuous at points x such that $x \ll \infty$. This particular application of the lemma is sharp, as direct calculation shows. For unlimited x by

$$e^y - e^x = e^x(e^{y-x} - 1) \ge e^x(y - x)$$

we see that this cannot be infinitesimal when $y - x$ is a sufficiently large infinitesimal, say, $y - x = e^{-x/2}$ (the inequality here is the same as the one in the proof of Theorem 3.11).

Thus we see that the exponential function is *not* (nearly) continuous on all of $\mathbb{R}$, and again we see the analogy between near continuity and conventional uniform continuity (the exponential function is not uniformly continuous on $\mathbb{R}$ in conventional mathematics).

4.3 Summation

The "radically elementary" analog of infinite series and integrals studied in calculus is a sum with an unlimited number of terms. Consider two sums

$$\sum_{i=1}^{\nu} x_i \qquad \text{and} \qquad \sum_{i=1}^{\nu} y_i$$

When will they be nearly equal? If ν is unlimited, then $x_i \simeq y_i$ for all i is *not* a sufficient condition. Consider $x_i = 1/\nu$ and $y_i = 2/\nu$.

A sufficient condition is given by Nelson (1987, Chapter 5). It uses the following notion. If x and y are real numbers with y nonzero, we say x is *asymptotic to* y, written $x \sim y$, when $x/y \simeq 1$, in which case x is also nonzero.

Lemma 4.2. *The external relation $\sim$ on $\mathbb{R} \setminus \{0\}$ is an equivalence.*

Proof. Symmetry, $x \sim x$, is obvious. Reflexivity is also obvious, because $x/y \simeq 1$ implies $y/x \simeq 1$. Transitivity is also obvious, because $x/y \simeq 1$ and $y/z \simeq 1$ imply $x/z \simeq 1$. □

Then we have the following, which is Theorem 5.3 in Nelson (1987).

Theorem 4.3. *If $x_i > 0$, $y_i > 0$, and $x_i \sim y_i$ for each i, then*

$$\sum_{i=1}^{\nu} x_i \sim \sum_{i=1}^{\nu} y_i. \tag{4.4}$$

Nelson's proof is instructive, so we copy it here.

Proof. Fix $\epsilon \gg 0$. Then we have $1 - \epsilon \le x_i/y_i \le 1 + \epsilon$ for each i and

$$(1 - \epsilon) \sum_{i=1}^{\nu} y_i \le \sum_{i=1}^{\nu} x_i \le (1 + \epsilon) \sum_{i=1}^{\nu} y_i.$$

Hence

$$1 - \epsilon \le \frac{\sum_{i=1}^{\nu} x_i}{\sum_{i=1}^{\nu} y_i} \le 1 + \epsilon \tag{4.5}$$

Since $\epsilon \gg 0$ was arbitrary, this implies the fraction in (4.5) is nearly equal to one, which implies (4.4). □

The conclusion of the theorem is not exactly what is wanted. We wanted $\simeq$ in place of $\sim$ in (4.4). The following simple lemma, which is Theorem 5.2 in Nelson (1987), usually gives us what we want.

Lemma 4.4. *Suppose x or y is appreciable. Then $x \sim y$ if and only if $x \simeq y$.*

Proof. Suppose y is appreciable. First, suppose $x \sim y$. Then $x/y - 1$ is infinitesimal, and multiplying by y leaves it infinitesimal. Hence $x \simeq y$. Conversely, suppose $x \simeq y$. Then $x - y$ is infinitesimal, and dividing by y leaves it infinitesimal. Hence $x \sim y$. □

Thus, so long as one side of (4.4) is appreciable, the other side is appreciable too, and there is no difference between $\sim$ and $\simeq$.

4.4 Integration

The title of this section should be in quotation marks. In "radically elementary" nonstandard analysis and probability theory, we replace integrals with finite sums. But Nelson (1987) often writes sums in a form that makes them look like integrals for easy comparison with conventional mathematics.

Following Nelson (1987, Chapter 9), we use the following notation. Let T be a finite subset of $\mathbb{R}$. For $t \in T$ such that $t \neq \max(T)$, we write dt to mean the difference between t and its successor in T, that is,

$$dt = \min\{\, u \in T : u > t \,\} - t. \tag{4.6}$$

For $t = \max(T)$, we adopt the convention $dt = 0$. We call dt the *spacing of T at t* and collectively refer to these dt as the *spacings of T*.

If all the spacings dt are infinitesimal and each $t \in T$ is limited, then we say T is a *near interval*. If all the spacings dt are infinitesimal and $\min(T) \simeq -\infty$ and $\max(T) \simeq \infty$, then we say T is a *near line*.

This "dt" notation allows us to write things like

$$\sum_{t \in T} e^{-t^2/2}\, dt \simeq \sqrt{2\pi} \tag{4.7}$$

when T is a near line. (This will be proved in this section.)

Theorem 4.5. *Suppose T is a near interval and suppose f and g are limited-valued functions on T such that $f(t) \simeq g(t)$ for all $t \in T$. Then*

$$\sum_{t \in T} f(t)\, dt \simeq \sum_{t \in T} g(t)\, dt. \tag{4.8}$$

Proof. Let L be the maximum of all $|f(t)|$ and $|g(t)|$ for $t \in T$. The maximum is achieved because T is finite and hence is limited by assumption. Fix $\epsilon \gg 0$

and define

$$\begin{aligned} T_+ &= \{\, t \in T : f(t) \geq \epsilon \,\} \\ T_0 &= \{\, t \in T : |f(t)| < \epsilon \,\} \\ T_- &= \{\, t \in T : f(t) \leq -\epsilon \,\} \end{aligned}$$

Then by Lemma 4.4 we have

$$f(t) \sim g(t), \qquad t \in T_+ \cup T_-$$

and hence by Theorem 4.3 and Lemma 4.4 again we have

$$\sum_{t \in T_+} f(t)\,dt \simeq \sum_{t \in T_+} g(t)\,dt$$

and the same with T_+ replaced by T_-. Also

$$\left| \sum_{t \in T_0} f(t)\,dt \right| \leq \sum_{t \in T_0} |f(t)|\,dt \leq \epsilon(b-a)$$

where a and b are the endpoints of T, and the same with f replaced by g and ϵ replaced by 2ϵ, because $|f(x)| \leq \epsilon$ implies $|g(x)| \leq 2\epsilon$. Thus from the triangle inequality and the sum of infinitesimals being infinitesimal we obtain

$$\left| \sum_{t \in T} f(t)\,dt - \sum_{t \in T} g(t)\,dt \right| \lesssim 3\epsilon(b-a)$$

Since $\epsilon \gg 0$ was arbitrary and a and b are limited (by the near interval assumption), this establishes (4.8). □

Theorem 4.6. *Suppose T is a near interval having endpoints a and b, and suppose f is a function $[a,b] \to \mathbb{R}$ that is Riemann integrable, has a limited bound, and is nearly continuous on $[a,b]$. Then*

$$\sum_{t \in T} f(t)\,dt \simeq \int_a^b f(t)\,dt.$$

Proof. Fix $\epsilon \gg 0$. By definition of Riemann integrability, there exists a subset S of $\mathbb{R}$ with endpoints a and b such that

$$\left| \sum_{s \in S} f(s)\,ds - \int_a^b f(t)\,dt \right| \leq \epsilon$$

(where ds is the spacing of S at s), and the same holds when S is replaced by a finer partition, in particular,

$$\left| \sum_{u \in U} f(u)\,du - \int_a^b f(t)\,dt \right| \leq \epsilon$$

where $U = S \cup T$ (and where du is the spacing of U at u).

Define $h : U \to \mathbb{R}$ by

$$h(u) = f(t), \qquad t \in T \text{ and } t \le u < t + dt.$$

By near continuity of f we have $f(u) \simeq h(u)$ for $u \in U$, and hence by Theorem 4.5

$$\sum_{u \in U} f(u)\, du \simeq \sum_{u \in U} h(u)\, du = \sum_{t \in T} f(t)\, dt.$$

Hence by the triangle inequality, we have

$$\left| \sum_{t \in T} f(t)\, dt - \int_a^b f(t)\, dt \right| \lesssim \epsilon.$$

Since $\epsilon \gg 0$ was arbitrary, this finishes the proof. □

Corollary 4.7. *Suppose T is a near line and suppose f and g are limited-valued functions on T such that $f(t) \simeq g(t)$ for all $t \in T$. Also suppose there exist $M \ll \infty$ and $\alpha \gg 1$ such that*

$$|f(t)| \le M|t|^{-\alpha}, \qquad |t| \simeq \infty \tag{4.9}$$

and similarly with f replaced by g. Then (4.8) *holds.*

Proof. For any appreciable δ we have

$$\sum_{\substack{t \in T \\ t < -a}} |f(t)|\, dt \le M \int_{-\infty}^{-a} |t|^{-\alpha}\, dt$$

$$\sum_{\substack{t \in T \\ t > a}} |f(t)|\, dt \le M \int_{a}^{\infty} |t - \delta|^{-\alpha}\, dt$$

hence

$$\sum_{\substack{t \in T \\ |t| > a}} |f(t)|\, dt \le \frac{2M(a - \delta)^{-(\alpha-1)}}{(\alpha - 1)} \tag{4.10}$$

and the right hand side is infinitesimal when $a \simeq \infty$ because $\alpha - 1$ is non-infinitesimal so $(a - \delta)^{\alpha-1}$ is the case X^u or X^Y in the table for powers in Section 3.3, hence $(a - \delta)^{-(\alpha-1)}$ is infinitesimal, M is limited, and $1/(\alpha - 1)$ is limited, and the product is infinitesimal. Similarly (4.10) holds with f replaced by g. Fix $\epsilon \gg 0$. Then

$$\sum_{\substack{t \in T \\ |t| > a}} |f(t)|\, dt \le \epsilon \quad \text{and} \quad \sum_{\substack{t \in T \\ |t| > a}} |g(t)|\, dt \le \epsilon$$

holds for every $a \simeq \infty$ hence, by overspill, for some limited a. Hence by Theorem 4.5 and the triangle inequality

$$\left| \sum_{t \in T} f(t)\,dt - \sum_{t \in T} g(t)\,dt \right| \lesssim 2\epsilon.$$

Since $\epsilon \gg 0$ was arbitrary, this finishes the proof. □

Corollary 4.8. *Suppose T is a near line and suppose f is a function $\mathbb{R} \to \mathbb{R}$ that is absolutely Riemann integrable, has a limited bound, and is nearly continuous at each point of T. Also suppose there exist $M \ll \infty$ and $\alpha \gg 1$ such that* (4.9) *holds. Then*

$$\sum_{t \in T} f(t)\,dt \simeq \int_{-\infty}^{\infty} f(t)\,dt.$$

Proof. As in the preceding proof, condition (4.9) implies

$$\int_{-\infty}^{-a} |f(t)|\,dt + \int_{a}^{\infty} |f(t)|\,dt \le \frac{2Ma^{-(\alpha-1)}}{(\alpha-1)}$$

and the right hand side is infinitesimal when $a \simeq \infty$. Fix $\epsilon \gg 0$. Then

$$\int_{-\infty}^{-a} |f(t)|\,dt + \int_{a}^{\infty} |f(t)|\,dt \le \epsilon \tag{4.11}$$

holds for every $a \simeq \infty$ hence, by overspill, for some limited a.

Also by the argument in the preceding proof

$$\sum_{\substack{t \in T \\ |t| > a}} |f(t)|\,dt \le \epsilon \tag{4.12}$$

holds for some limited a. Moreover we can choose one limited a so that (4.11) and (4.12) both hold.

By Theorem 4.6

$$\sum_{\substack{t \in T \\ |t| \le a}} f(t)\,dt \simeq \int_{-a}^{a} f(t)\,dt.$$

Now apply the triangle inequality and the arbitrariness of ϵ. □

The condition (4.9) is not sharp, but a sharp condition, such as asserting that the left hand sides of (4.11) and (4.12) are infinitesimal for all $a \simeq \infty$, would leave a lot of work to be done in applying the corollary.

The inequality $\exp(x) \ge 1 + x$ implies $\exp(-t^2/2) \le 2|t|^{-2}$. By Lemma 4.1 $t \mapsto \exp(-t^2/2)$ is nearly continuous, and it is bounded by 1. We know from calculus that $\int_{-\infty}^{\infty} \exp(-t^2/2)\,dt = \sqrt{2\pi}$. Applying the corollary gives (4.7),

4.5 Derivability

If T is a subset of $\mathbb{R}$, then a function $g : T \to \mathbb{R}$ is *derivable at the point* $x \in T$ if there exists a limited real number L such that

$$\frac{g(x_2) - g(x_1)}{x_2 - x_1} \simeq L, \qquad \text{whenever } x_1 \simeq x \simeq x_2 \text{ and } x_1 \neq x_2 \tag{4.13}$$

(it being understood that x_1 and x_2 are elements of T).[2]

It is clear that an analogous characterization is the following: g is *derivable at* x if there exists a limited real number L such that whenever $x_1 \simeq x$ and $h \simeq 0$ there exists an $\alpha \simeq 0$ such that

$$g(x_1 + h) - g(x_1) = Lh + \alpha h, \tag{4.14}$$

(it being understood that x_1 and $x_1 + h$ are elements of T). To see the connection between the two characterizations take $x_2 = x_1 + h$ and α the difference between the two sides of (4.13).

Like other external notions, derivability is different from its internal analog (differentiability) and is actually a stronger and more useful property. Suppose g is derivable at every point of T and we define a limited-real-valued function L on T such that

$$\frac{g(x_2) - g(x_1)}{x_2 - x_1} \simeq L(x), \qquad \text{whenever } x_1 \simeq x \simeq x_2 \text{ and } x_1 \neq x_2 \tag{4.15}$$

(it being understood that x, x_1, and x_2 are elements of T). Clearly the function L is not unique, since it can be changed by an infinitesimal amount at any x without affecting the validity of (4.15). However, it is clear from (4.15) that L is nearly continuous on T and from (4.14) that g is nearly continuous on T.

Lemma 4.9. *Suppose g is a differentiable function $I \to \mathbb{R}$, where I is an open interval, and g' is limited and nearly continuous on I. Then g is derivable on I and we may take $L(x) = g'(x)$ in* (4.15).

Proof. For x, x_1, and x_2 in I such that $x_1 \simeq x \simeq x_2$ and $x_1 \neq x_2$, by the mean value theorem

$$\frac{g(x_2) - g(x_1)}{x_2 - x_1} = g'(\xi)$$

where ξ is between x_1 and x_2 and hence $\xi \simeq x$ and $g'(\xi) \simeq g'(x)$ by the near continuity assumption. □

[2]This notion is taken from Nelson (1977, Section 5). It does not appear in Nelson (1987). In Nelson (1977), the formula (4.13) actually defines the derivative $g'(x)$ because for standard g and x there is a unique standard L (by the standardization axiom of IST), which we denote $g'(x)$, that satisfies (4.13) and this implicitly defines derivability for nonstandard g or x (by the transfer axiom of IST). Our "radically elementary" nonstandard analysis is too weak to do that (having neither standardization nor transfer).

Applying both Lemma 4.1 and Lemma 4.9 we see that a function that has two derivatives, both limited on I, is derivable on I. And so most familiar functions from calculus are derivable, at least on limited intervals.

Another way to look at derivability, is to note that when $L(x)$ in (4.15) is appreciable then $\simeq$ can be replaced by $\sim$ by Lemma 4.4 and the result is equivalent to

$$g(x_2) - g(x_1) \sim L(x)(x_2 - x_1), \qquad \text{whenever } x_1 \simeq x \simeq x_2 \tag{4.16}$$

(which, we reiterate, only holds when $L(x)$ is appreciable).

As an application of this theory, we prove the following useful identity.

Theorem 4.10. *For limited real numbers x and unlimited natural numbers n*

$$\left(1 + \frac{x}{n}\right)^n \simeq e^x. \tag{4.17}$$

Proof. By (near) continuity of the exponential function, it is enough to show

$$n \log\left(1 + \frac{x}{n}\right) \simeq x. \tag{4.18}$$

By derivability of the logarithm function and because its derivative at one is appreciable, we have by (4.16) and Lemma 4.9

$$\log(1 + h) \sim h, \qquad h \simeq 0.$$

Hence for x and n as in the statement of the theorem,

$$n \log\left(1 + \frac{x}{n}\right) \sim n \cdot \frac{x}{n} = x$$

and this implies (4.18) by another application of Lemma 4.4. □

Part II

Probability

Chapter 5

Radically Elementary Probability Theory

5.1 Introduction

Nelson (1987) invented a new formalism for probability theory in which all random variables are defined on probability spaces that have

(i) finite sample space and

(ii) no nonempty events of probability zero.

The main point of (ii) is to assure that conditional probabilities are always well defined. When not using conditional probability, it need not be imposed.

Point (i) implies two other restrictions.

(iii) We only use finite collections of random variables. Every stochastic process has a finite index set.

(iv) We only use finite families of probability models. Every statistical model has a finite parameter space.

Nelson (1987) uses (iii). Our justification of (iv) is that a likelihood is a stochastic process indexed by the parameter. Hence to obey (iii) the parameter must take values in a finite (though perhaps unlimited) set. The need for (iv) is even more obvious if one is a Bayesian. If the parameter is a random variable, then (i) implies (iv).

One might think (i) leaves no room for interesting advanced probability theory. Under (i) every probability distribution is discrete. There are no truly continuous random variables. Also under (i) it is not possible to have an infinite sequence of independent random variables (except for the trivial special case where the random variables are constant).

But Nelson combined this "radical" simplification with an innovation, the use of nonstandard analysis. In "Nelson-style" probability theory, a discrete

distribution in which every point has infinitesimal probability can behave much like a continuous distribution in conventional "Kolmogorov-style" probability theory.

In "Nelson-style" probability theory, a *finite* sequence $X_1, \ldots, X_n$ of independent random variables (which can be defined on a finite sample space), where n is unlimited can behave much like an infinite sequence in "Kolmogorov-style" probability theory. For example, the law of large numbers and the central limit theorem can hold for such sequences, where, of course, we are referring to "Nelson-style" analogs of the conventional theorems (Nelson, 1987, Chapters 16 and 18).

5.2 Unconditional Probability

5.2.1 Probability

Probability theory on finite sample spaces having no nonempty events of probability zero is very simple. Probability models consist of a finite set Ω (the *sample space*) and a strictly positive function pr (the *probability mass function*) on Ω such that $\sum_{\omega\in\Omega} \operatorname{pr}(\omega) = 1$.

Every subset of Ω is an *event*, and the probability of an event A is given by

$$\Pr(A) = \sum_{\omega\in A} \operatorname{pr}(\omega). \tag{5.1a}$$

There are never any questions of measurability.

Note that we distinguish between pr and Pr, the relationship being (5.1a) going one way and

$$\operatorname{pr}(\omega) = \Pr(\{\omega\}) \tag{5.1b}$$

going the other. We can think of Pr as a probability measure. It certainly is the Nelson-style analog of a Kolmogorov-style probability measure. However, it is much simpler. There is no sigma-algebra (every subset of Ω is an event). And countable additivity is vacuous (since Ω is finite, there are only finitely many events).

5.2.2 Expectation

Random Scalars

A real-valued function on the sample space is called a *random variable*, and the *expectation* of a random variable X is given by

$$E(X) = \sum_{\omega\in\Omega} X(\omega) \operatorname{pr}(\omega). \tag{5.2}$$

There are never any questions of existence of expectations. The set of all random variables is the finite-dimensional vector space $\mathbb{R}^\Omega$.

The *indicator function* of the set A is the function $I_A : \Omega \to \mathbb{R}$ defined by

$$I_A(\omega) = \begin{cases} 0, & \omega \in \Omega \setminus A \\ 1, & \omega \in A \end{cases} \tag{5.3}$$

The set $\Omega \setminus A$ is called the *complement of* A and is also denoted A^c. Using this notation

$$\Pr(A) = E(I_A)$$

(probability is expectation of indicator functions).

Random Vectors

A function from the sample space Ω to a vector space V is called a *random vector* and the *expectation* of a random vector X is given by (5.2) where $X(\omega)\,\mathrm{pr}(\omega)$ is interpreted as multiplication of the vector $X(\omega)$ by the scalar $\mathrm{pr}(\omega)$ and the sum is interpreted as vector addition, which is well defined because Ω is finite.

The set of all (V-valued) random vectors is V^Ω. Note that even if V is infinite-dimensional, it has a finite-dimensional subspace that contains all the $X(\omega)$ for $\omega \in \Omega$ because Ω is finite. Even if we are interested in a finite sequence $X_1, \ldots, X_n$ of random variables, there exists a finite dimensional subspace that contains all the $X_i(\omega)$ for $1 \le i \le n$ and $\omega \in \Omega$. We shall never be interested in infinite sequences or any other infinite collection of random vectors — restriction (iii) discussed in Section 5.1. Thus without loss of generality we may assume V is finite-dimensional.

Random Elements

A function from the sample space Ω to an arbitrary set S is called a *random element of* S. The set of all random elements (of S) is S^Ω. The same sort of argument as in the last paragraph of the preceding section says that without loss of generality we may take S to be finite.

Random elements need not have expectations. The addition and multiplication in (5.2) are not defined for an element X of an abstract set S. But if X is a random element of S and f is a function $S \to \mathbb{R}$, then $f \circ X$, which is usually written $f(X)$, is a random variable and does have expectation.

5.3 Conditional Probability

For any family $\mathcal{X}$ of random variables define a relation $\overset{\mathcal{X}}{\sim}$ on Ω by

$$\omega_1 \overset{\mathcal{X}}{\sim} \omega_2 \quad \longleftrightarrow \quad (\forall X \in \mathcal{X})\big(X(\omega_1) = X(\omega_2)\big) \tag{5.4}$$

Clearly, this is an internal equivalence relation, and hence defines a partition $\mathcal{S}$ of Ω. We write $\mathcal{S} = \mathrm{at}(\mathcal{X})$ and call the elements of $\mathcal{S}$ the *atoms* of $\mathcal{X}$. By

definition, every element of $\mathcal{X}$ is constant on each element of $\mathcal{S}$, and $\mathcal{S}$ is the coarsest partition of Ω that has this property.

The *algebra* generated by $\mathcal{X}$ is the largest family of random variables $\mathcal{A}$ such that $\text{at}(\mathcal{A}) = \text{at}(\mathcal{X})$. Clearly, it is the set of all random variables that are constant on each element of $\text{at}(\mathcal{X})$.[1]

An algebra $\mathcal{A}$ is closed under arbitrary operations. If f is a real-valued function with d real arguments and X_1, ..., X_d are elements of $\mathcal{A}$, then $Y = f(X_1, \ldots, X_d)$ is also an element of $\mathcal{A}$, where this notation is defined (as is usual in probability theory) by

$$Y(\omega) = f\big(X_1(\omega), \ldots, X_d(\omega)\big), \qquad \omega \in \Omega.$$

(This is obvious from the definition.)

5.3.1 Conditional Expectation

Let $\mathcal{A}$ be a family of random variables (not necessarily an algebra), and let A_ω denote the element of $\text{at}(\mathcal{A})$ containing ω. The *conditional expectation* of a random variable X given the family $\mathcal{A}$ is the random variable Y defined by

$$Y(\omega) = \frac{1}{\Pr(A_\omega)} \sum_{\omega' \in A_\omega} X(\omega') \operatorname{pr}(\omega'), \qquad \omega \in \Omega. \tag{5.5}$$

There are never any questions of existence or uniqueness; $\Pr(A_\omega)$ cannot be zero because of the assumption that pr is strictly positive.

Nelson mostly uses the notation $E_{\mathcal{A}} X$ to denote the random variable defined by (5.5) but also uses the notation $E(X|\mathcal{A})$ more common in Kolmogorov-style theory and also uses the notation $E(X|Z_1, \ldots, Z_d)$ when $\mathcal{A} = \{Z_1, \ldots, Z_d\}$.

Theorem 5.1. *Suppose $\mathcal{A}$ and $\mathcal{B}$ are algebras of random variables and $\mathcal{A} \subset \mathcal{B}$. Then*

$$E_{\mathcal{A}}(X + Y) = E_{\mathcal{A}} X + E_{\mathcal{A}} Y, \qquad X, Y \in \mathbb{R}^\Omega \tag{5.6a}$$

$$E_{\mathcal{A}}(XY) = X E_{\mathcal{A}} Y, \qquad X \in \mathcal{A},\ Y \in \mathbb{R}^\Omega \tag{5.6b}$$

$$E_{\mathcal{A}} E_{\mathcal{B}} = E_{\mathcal{A}} \tag{5.6c}$$

$$E E_{\mathcal{A}} = E \tag{5.6d}$$

The meaning of (5.6c) or (5.6d) is that the two sides of the equation are equal when applied to any random variable. The proofs are straightforward verifications directly from the definition (5.5).

[1] Nelson (1987, p. 6) defines *algebra* differently: a family of random variables containing the constant random variables and closed under addition and multiplication. His definition is more "mathematical" because it justifies the name "algebra." But he then immediately proves that his definition characterizes the same notion as ours. As closure under addition and multiplication are not particularly interesting in light of the comments immediately following the footnoted text, we just take the characterization more relevant to probability theory as our definition.

5.3.2 Conditional Probability

As with unconditional probability, conditional probability is expectation of indicator functions: we define conditional probability by

$$\Pr(B|\mathcal{A}) = \Pr_{\mathcal{A}}(B) = E_{\mathcal{A}}(I_B). \tag{5.7}$$

Consider the special case where $\mathcal{A} = \{I_A\}$, so $\text{at}(\mathcal{A}) = \{A, A^c\}$. Then for $\omega \in A$ so $A_\omega = A$, the definition (5.5) gives

$$\frac{1}{\Pr(A)} \sum_{\omega \in A} I_B(\omega) \operatorname{pr}(\omega) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

for the value of $\Pr(B|\mathcal{A})$ at such an ω.

As in undergraduate probability theory we write this

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{5.8}$$

and regard it as an independent definition of what $\Pr(B|\cdot)$ means when the thingy behind the bar is an event rather than a family of random variables. Clearly, (5.8) is well defined whenever A is nonempty (because of our rule that nonempty events have nonzero probability).

It is clear that $\Pr(B|\mathcal{A})$ evaluated at an $\omega \in A^c$ is $\Pr(B|A^c)$. Moreover, for any family of random variables $\mathcal{A}$ (not just ones generated by a single indicator function), we have

$$\Pr(B|\mathcal{A})(\omega) = \Pr(B|A_\omega)$$

where the notation on the left hand side means the random variable $\Pr(B|\mathcal{A})$ evaluated at the point ω and, as in the definition (5.5), A_ω is the element of $\text{at}(\mathcal{A})$ containing ω. And this connects the two notions of conditional probability: the undergraduate level (5.8) and the PhD level (5.7), at least "PhD level" when done Kolmogorov-style, our Nelson-style definition using (5.5) being actually only undergraduate level in difficulty.

Although one does not usually see the "PhD level" definition (5.7) of conditional probability in undergraduate courses, one does see the so-called *regression function* $E(Y|X_1, \ldots, X_p)$ without which one cannot understand multiple regression. Of course, it is usually introduced without having the PhD level Kolmogorov-style definition that makes it rigorous. Our Nelson-style definition serves just as well to make the concept rigorous and is much lower in level of difficulty.

5.4 Distribution Functions

A random variable X induces a distribution function F defined by

$$F(x) = \Pr(X \le x), \qquad x \in \mathbb{R} \tag{5.9a}$$

and probability mass function f defined by

$$f(x) = \Pr(X = x), \qquad x \in \mathbb{R}. \tag{5.9b}$$

These are related by

$$f(x) = F(x) - \max_{y<x} F(y) \tag{5.9c}$$

and

$$F(x) = \sum_{\substack{y\in\mathbb{R}\\ y\le x}} f(y), \tag{5.9d}$$

a relation denoted by $f = dF$.

The sum in (5.9d) is always well defined because of the assumption (i) of Section 5.1 that all random variables are defined on finite sample spaces so $f(y)$ is zero for all but finitely many y. This implies that F is a step function with only finitely many jumps. We say that the set of y such that $f(y)$ is positive is the *support of* f and also the *support of* F.

If h is a real-valued function on $\mathbb{R}$ and X is a random variable having distribution function F, then we can write

$$E\{h(X)\} = \sum_{x\in\mathbb{R}} h(x)\, dF(x). \tag{5.9e}$$

Here too, the sum is always well defined because of the assumption of finite support.

Recall the notation introduced in Section 4.4 in which T is a finite subset of $\mathbb{R}$ and dt denotes the spacings of T.

Lemma 5.2. *If F is the distribution function of a nonnegative random variable and T is a subset of $[0,\infty)$ containing zero and the support of F (and perhaps other points), then*

$$E(X) = \sum_{t\in T} \bigl[1 - F(t)\bigr]\, dt. \tag{5.10}$$

Proof.

$$\begin{aligned} E(X) &= \sum_{x\in\mathbb{R}} x\, dF(x) \\ &= \sum_{x\in\mathbb{R}} \sum_{\substack{t\in T\\ t<x}} dt\, dF(x) \\ &= \sum_{t\in T} \sum_{\substack{x\in\mathbb{R}\\ x>t}} dF(x)\, dt \\ &= \sum_{t\in T} \bigl[1 - F(t)\bigr]\, dt \end{aligned}$$

□

5.5 Probability Measures

A random element X of a set S induces a probability measure P defined by

$$P(A) = \Pr(X \in A), \qquad A \subset S \tag{5.11a}$$

and probability mass function p defined by

$$p(x) = \Pr(X = x), \qquad x \in S. \tag{5.11b}$$

These are related by

$$p(x) = P(\{x\}) \tag{5.11c}$$

and

$$P(A) = \sum_{x \in S} p(x), \tag{5.11d}$$

a relation denoted by $p = dP$. This relationship between X and P is denoted $P = \mathcal{L}(X)$, and we say P is the *law of* X.

The sum in (5.11d) is always well defined, even if S is an infinite set, because of the assumption (i) of Section 5.1 that all random elements are defined on finite sample spaces so $p(x)$ is zero for all but finitely many x. We say that the set of x such that $p(x)$ is positive is the *support of* p and also the *support of* P. We shall never be interested in measures that do not have finite support.

If h is a real-valued function on S, X is a random element of S, and $P = \mathcal{L}(X)$, then we can write

$$E\{h(X)\} = \sum_{x \in S} h(x)\, dP(x). \tag{5.11e}$$

As with (5.11d), the sum in (5.11e) is always well defined because of the assumption that P has finite support.

Chapter 6

More Radically Elementary Probability Theory

6.1 Almost Surely

The definition of "almost surely" appropriate in Nelson-style probability theory (Nelson, 1987, Chapter 7) goes as follows: a property A holds almost surely if for every $\epsilon \gg 0$ there exists an event N (which may depend on ϵ) such that $\Pr(N) \leq \epsilon$ and $A(\omega)$ is true except for $\omega \in N$.

If the property A in question is *internal*, then the event

$$\{\, \omega \in \Omega : A(\omega) \,\} \tag{6.1}$$

has probability nearly equal to one, which more resembles the conventional (Kolmogorov-style) definition.

But if the property A is *external*, then (6.1) is an instance of *illegal set formation* (Section 2.4). It need not define a set, and hence need not have a probability. Thus we need the more complicated definition involving a different exception set N for every $\epsilon \gg 0$ when we want to say an external property holds almost surely.

Lemma 6.1. *Suppose A_1, ..., A_n are properties that hold almost surely and n is limited. Then A_1, ..., A_n hold simultaneously almost surely.*

Proof. For every $\epsilon \gg 0$, we have $\epsilon/n \gg 0$, hence there exist events N_i such that $\Pr(N_i) \geq \epsilon/n$ and $A_i(\omega)$ holds except for $\omega \in N_i$. But then $\Pr(\bigcup_{i=1}^{n} N_i) \leq \epsilon$ and and $A_i(\omega)$ holds for $i = 1, \ldots, n$, except for $\omega \in \bigcup_{i=1}^{n} N_i$. □

6.1.1 Infinitesimal Almost Surely

When the property A in question is $X \simeq 0$, the last bit of the definition of almost surely becomes: and $X(\omega) \simeq 0$ except for $\omega \in N$.

Lemma 6.2. *The following three conditions are equivalent.*

(i) *X is infinitesimal almost surely.*

(ii) *If $\lambda \gg 0$, then $\Pr(|X| \geq \lambda) \simeq 0$.*

(iii) *There exists a $\lambda \simeq 0$ such that $\Pr(|X| \geq \lambda) \simeq 0$.*

This is Theorem 7.1 in Nelson (1987). We do not give a proof here. The proof is very similar to that of the lemma in the next section.

6.1.2 Limited Almost Surely

When the property A in question is $|X| \ll \infty$, the last bit of the definition of almost surely becomes: and $|X(\omega)| \ll \infty$ except for $\omega \in N$.

Lemma 6.3. *The following three conditions are equivalent.*

(i) *X is limited almost surely.*

(ii) *If $x \simeq \infty$, then $\Pr(|X| \geq x) \simeq 0$.*

(iii) *For every $\epsilon \gg 0$ there exists a limited x such that $\Pr(|X| \geq x) \leq \epsilon$.*

Proof. Assume (i). Then for every $\epsilon \gg 0$ there exists an event N such that $X(\omega)$ is limited except when $\omega \in N$. Hence if $x \simeq \infty$ the event $|X| \geq x$ is contained in N, and $\Pr(|X| \geq x) \leq \epsilon$. Since ϵ was arbitrary, (ii) holds. Thus (i) $\Rightarrow$ (ii).

Assume (ii). Fix $\epsilon \gg 0$. Then $\Pr(|X| \geq x) \leq \epsilon$ holds for all unlimited positive x, hence by overspill for some limited x. Thus (ii) $\Rightarrow$ (iii).

(iii) $\Rightarrow$ (i) is obvious. □

The concept in Kolmogorov-style probability theory analogous to "limited almost surely" is "tight." In Kolmogorov-style theory, tightness is an uninteresting concept when applied to single random variables, because by countable additivity *every* random variable is tight.

In conventional finitely-additive probability theory, non-tight random variables exist, although proof of their existence involves fancy mathematics like the Hahn-Banach theorem. In Nelson-style probability theory, random variables that are not limited almost surely are easily constructed. For example, take the (discrete) uniform distribution on the integers 1, ..., n, where n is unlimited.

6.2 L^1 Random Variables

In Nelson-style theory, every random variable has a well defined expectation given by (5.2). But, unlike the situation in Kolmogorov-style probability theory, the mere existence of expectation proves nothing (since every random variable has expectation, existence is vacuous).

One might guess that the Nelson-style property analogous to Kolmogorov-style existence of expectation is *limited absolute expectation*, but it turns out that an even stronger property is needed for a random variable to be "well behaved."

A random variable X is L^1 if $E\bigl(|X| I_{(a,\infty)}(|X|)\bigr)$ nearly converges to zero as a goes to infinity, meaning

$$E\bigl(|X| I_{(a,\infty)}(|X|)\bigr) \simeq 0, \qquad a \simeq \infty. \tag{6.2}$$

The analogous Kolmogorov-style definition is easily seen to define L^1. If X is a Kolmogorov-style random variable that has expectation, then the left hand side of (6.2) converges to zero as $a \to \infty$ by dominated convergence. On the other hand, if X is a Kolmogorov-style random variable that does not have expectation, then the left hand side of (6.2) is equal to $+\infty$ for all a, because if there existed an a for which the left hand side was finite, then we would have

$$\begin{aligned} E\bigl(|X|\bigr) &= E\bigl(|X| I_{[0,a]}(|X|)\bigr) + E\bigl(|X| I_{(a,\infty)}(|X|)\bigr) \\ &\le a + E\bigl(|X| I_{(a,\infty)}(|X|)\bigr) \end{aligned}$$

finite, contrary to assumption.

6.2.1 The Radon-Nikodym Theorem

We don't usually use the definition of L^1 directly. More often we use the following characterization, which is Theorem 8.1 in Nelson (1987).

Theorem 6.4 (Radon-Nikodym and converse)**.** *A random variable X is L^1 if and only if $E\bigl(|X|\bigr)$ is limited and, for all events M, if $\Pr(M) \simeq 0$, then $E\bigl(|X| I_M\bigr) \simeq 0$.*

Thus we see that L^1 is a stronger property than limited absolute expectation (however, see Theorem 6.8 below for a criterion based on limitedness of higher moments). Here is an example of a random variable X that has limited absolute expectation, but is not L^1. Let X be Bernoulli(p) and $Y = bX$ with $b > 0$. Then Y is nonnegative, so we may omit absolute values.

$$E\bigl(Y I_{(a,\infty)}(Y)\bigr) = \begin{cases} bp, & a < b \\ 0, & a \ge b \end{cases}$$

so, if $b \simeq \infty$ and $bp \gg 0$, then Y is not L^1. But $E(Y) = bp$, so if $bp \ll \infty$, then Y does have limited absolute expectation. For example, take $p = 1/b$.

The name of the theorem comes from Nelson (1987). He often labels his theorems with names of theorems from Kolmogorov-style theory. The names don't mean that his theorem is the same as the Kolmogorov-style theorem (or even a corollary of it). Rather they mean that his theorem is the analog in his theory of the named theorem in Kolmogorov-style theory.

In this case the analogy is the following. Suppose $(\Omega, \mathcal{A}, P)$ is a Kolmogorov-style probability space and X a Kolmogorov-style $L^1(P)$ random variable on this probability space. Define a positive measure ν on $(\Omega, \mathcal{A})$ by

$$\nu(A) = \int_A |X| \, dP.$$

Then $P(M) = 0$ implies $\nu(M) = 0$, a property called *absolute continuity of ν with respect to P* in Kolmogorov-style theory, which is the condition in the Radon-Nikodym theorem. The condition in the Nelson-style Radon-Nikodym theorem is $P(M) \simeq 0$ implies $\nu(M) \simeq 0$.

Unlike the situation in conventional mathematics (Lebesgue-style measure theory as well as Kolmogorov-style probability theory) L^1 is not a vector space. It is not even a set. The following

$$\{ X \in \mathbb{R}^\Omega : X \text{ is } L^1 \} \tag{6.3}$$

is illegal set formation because L^1 is an *external property*. In fact, we can prove that there does not exist a subset S of $\mathbb{R}^\Omega$ such that $X \in S$ if and only if X is L^1, because if this set did exist, then by Theorem 6.4, letting Y be the constant random variable everywhere equal to one,

$$\{ a \in \mathbb{R} : aY \in S \}$$

would be the set of limited real numbers, which does not exist (Theorem 3.13).

However, it immediately follows from Theorem 6.4 that

$$X \text{ and } Y \text{ are } L^1 \longrightarrow X + Y \text{ is } L^1 \tag{6.4a}$$

$$X \text{ is } L^1 \text{ and } |a| \ll \infty \longrightarrow aX \text{ is } L^1 \tag{6.4b}$$

$$Y \text{ is } L^1 \text{ and } |X| \leq |Y| \longrightarrow X \text{ is } L^1 \tag{6.4c}$$

so Nelson-style L^1 behaves much like Kolmogorov-style L^1. The differences are that Kolmogorov-style theory would allow unlimited a in (6.4b) and would allow $|X| \leq |Y|$ to hold only almost surely in (6.4c).

In Nelson-style theory we cannot insert "almost surely" in (6.4c) if there exists any point ω having infinitesimal probability (that is, whenever "almost surely" is not vacuous), because if X is L^1 and we define Y_a to be equal to X everywhere except at ω where we have $Y_a(\omega) = a$, then $EY_a \to \infty$ as $a \to \infty$ and hence Y_a is not L^1 for sufficiently large a, but $Y_a \simeq X$ almost surely.

6.2.2 The Lebesgue Theorem

Another important theorem about L^1 is the following, which it gets its name because it is the Nelson-style analog of the Lebesgue dominated convergence theorem, monotone convergence theorem, and Fatou's lemma. It is Theorem 8.2 in Nelson (1987).

Theorem 6.5 (Lebesgue)**.** *If X and Y are L^1 and $X \simeq Y$ almost surely, then $E(X) \simeq E(Y)$.*

Note that what is a convergence in Kolmogorov-style theory has become an external equivalence relation in Nelson-style theory. This is typical. As we shall see, the same thing happens with convergence in probability and convergence in distribution.

How can this one simple Nelson-style theorem replace all of that Lebesgue-Kolmogorov-style theory? In Nelson-style theory, we only look at finite sequences, and existence and uniqueness of limits are not at issue: we can always take the limit to be the last element of the sequence but the limit is never unique. Nor can it be an issue to prove that the limit is L^1. By the comment following and concerning (6.4c), we can never prove a limit to be L^1 because all random variables almost surely equal to the limit are also limits but many of them are not L^1. Thus the only part of the dominated, monotone, and Fatou convergence theorems that is of interest is that almost sure convergence implies convergence of expectations, and this is assured by Theorem 6.5. Nelson-style theory is in some respects much simpler than the competition.

Lemma 6.6. *Let Z be a nonnegative random variable. If $E(Z) \simeq 0$, then $Z \simeq 0$ almost surely. Conversely, if Z is L^1 and $Z \simeq 0$ almost surely, then $E(Z) \simeq 0$.*

Proof. One direction is the Lebesgue theorem. The other direction is Markov's inequality.

$$\Pr(Z > \lambda) \le \frac{E(Z)}{\lambda}. \tag{6.5}$$

If $E(Z) \simeq 0$, then for every $\lambda \gg 0$ the right hand side of (6.5) is infinitesimal over non-infinitesimal equals infinitesimal. Hence by criterion (ii) of Lemma 6.2 $Z \simeq 0$ almost surely. □

This lemma is the Nelson-style analog of the Kolmogorov-style theorem that a nonnegative random variable Z is zero almost surely if and only if $E(Z) = 0$.

For any random variable X and any positive real number a define another random variable

$$X^{(a)}(\omega) = \begin{cases} -a, & X(\omega) < -a \\ X(\omega), & -a \le X(\omega) \le a \\ a, & X(\omega) > a \end{cases} \tag{6.6}$$

Lemma 6.7 (Approximation)**.** *Suppose X and Y are L^1 random variables and $EX^{(a)} \simeq EY^{(a)}$ for all limited a. Then $EX \simeq EY$.*

Proof. For every $\epsilon \gg 0$ we have

$$|EX - EX^{(a)}| \le \epsilon \quad \text{and} \quad |EY - EY^{(a)}| \le \epsilon \tag{6.7}$$

for all unlimited a and hence by overspill for some limited a. But for this a we have $EX^{(a)} \simeq EY^{(a)}$ and hence by the triangle inequality

$$|EX - EY| \le 3\epsilon. \tag{6.8}$$

Since (6.8) holds for every $\epsilon \gg 0$, the left hand side must be infinitesimal. □

6.2.3 L^p Random Variables

For $1 \leq p < \infty$, we say a random variable X is L^p if $|X|^p$ is L^1, and we say a random variable X is L^∞ if it is limited, that is, if $|X(\omega)| \ll \infty$ for all ω (Nelson, 1987, p. 31). And for such variables we define the norms

$$\|X\|_p = E\{|X|^p\}^{1/p} \tag{6.9a}$$

when $1 \leq p < \infty$ and

$$\|X\|_\infty = \max_{\omega \in \Omega} |X(\omega)|. \tag{6.9b}$$

In (6.9b) we can write "max" instead of "sup" (the supremum is achieved) because Ω is a finite set. Hence, if X is L^∞, then $\|X\|_\infty$ is limited. The analogous property, if X is L^p, then $\|X\|_p$ is limited, holds for $1 \leq p \ll \infty$ because the expectation in (6.9a) is limited by the Radon-Nikodym theorem.

Theorem 6.8. *Suppose $1 \leq p \ll q \ll \infty$ and $E(|X|^q) \ll \infty$, then X is L^p.*

Proof. First

$$|X|^p = \frac{|X|^q}{|X|^{p(q/p-1)}}$$

from which we obtain

$$|X|^p I_{(a,\infty)}(|X|^p) \leq \frac{|X|^q}{a^{q/p-1}}$$

and

$$E\{|X|^p I_{(a,\infty)}(|X|^p)\} \leq \frac{E\{|X|^q\}}{a^{q/p-1}}.$$

Now $q/p - 1 = (q-p)/p$ is appreciable over appreciable equals appreciable, hence $a^{q/p-1} \simeq \infty$ whenever $a \simeq \infty$. Since $E\{|X|^q\} \ll \infty$, we have

$$E\{|X|^p I_{(a,\infty)}(|X|^p)\} \simeq 0, \qquad a \simeq \infty,$$

so $|X|^p$ is L^1 and X is L^p. □

6.2.4 Conditional Expectation

There are two important theorems about L^1 and conditional expectation (Theorems 8.3 and 8.4 in Nelson, 1987) which are repeated here below.

Theorem 6.9. *If $1 \leq p \leq \infty$ and X is L^p and $\mathcal{A}$ is a family of random variables, then $E_{\mathcal{A}}X$ is L^p.*

Theorem 6.10. *If X is L^1 and $\mathcal{A}$ is a family of random variables, then X is L^1 on almost every atom of $\mathcal{A}$.*

The meaning of Theorem 6.10 is not entirely obvious. It means that for every $\epsilon \gg 0$ there exists a set N, which in this case may be taken to be a union of atoms of $\mathcal{A}$, such that $\Pr(N) \leq \epsilon$ and

$$E_{\mathcal{A}}|X|I_{(a,\infty)}(|X|) \simeq 0, \qquad a \simeq \infty \tag{6.10}$$

except on N (the conditional expectation is an element of $\mathcal{A}$, hence constant on atoms of $\mathcal{A}$ and the assertion is that (6.10) holds except on those atoms of $\mathcal{A}$ that are contained in N).

Theorem 6.11 (Conditional Lebesgue)**.** *If X and Y are L^1 random variables such that $X \simeq Y$ almost surely and $\mathcal{A}$ is a family of random variables, then $E(X \mid \mathcal{A}) \simeq E(Y \mid \mathcal{A})$ almost surely.*

Proof. Let $Z = X - Y$. Then Z is L^1 and $Z \simeq 0$ almost surely. We are to show that $E(Z \mid \mathcal{A}) \simeq 0$ almost surely. Define $W = E(Z \mid \mathcal{A})$. W is L^1 by Theorem 6.9. By the conditional Jensen inequality (Nelson, 1987, p. 8) and iterated conditional expectation (5.6d)

$$E(|W|) = E\big[|E(Z \mid \mathcal{B})|\big] \leq E\big[E(|Z| \mid \mathcal{B})\big] = E(|Z|). \tag{6.11}$$

Now Z is L^1 if and only if $|Z|$ is by definition of L^1 and Z is infinitesimal almost surely if and only if $|Z|$ is because a number z is infinitesimal if and only if $|z|$ is. Hence $|Z|$ is L^1 and infinitesimal almost surely, so by Lemma 6.6 $E(|Z|) \simeq 0$. Hence (6.11) implies $E(|W|) \simeq 0$, and by the other direction of Lemma 6.6 $|W|$ is infinitesimal almost surely, so W is infinitesimal almost surely. □

6.2.5 The Fubini Theorem

Why do we have a section with this title? If expectations are always finite sums, isn't it obvious that we can interchange the order of summation? Yes, it is. But the Fubini theorem in Kolmogorov-style probability theory also makes measurability and integrability assertions (some authors put these in a preliminary lemma). The measurability assertions are vacuous in Nelson-style probability theory, but the integrability assertions are still important (Corollary to Theorem 8.4 in Nelson, 1987, repeated below).

Theorem 6.12 (Fubini)**.** *Suppose X is L^1 on a probability space with $\Omega = \Omega_1 \times \Omega_2$ and $\operatorname{pr}\big((\omega_1, \omega_2)\big) = \operatorname{pr}_1(\omega_1) \operatorname{pr}_2(\omega_2)$, then for pr_2 almost all ω_2 in Ω_2 the random variable $\omega_1 \mapsto X(\omega_1, \omega_2)$ on $(\Omega_1, \operatorname{pr}_1)$ is L^1.*

Chapter 7

Stochastic Convergence

7.1 Almost Sure Convergence

A sequence $X_1, \ldots, X_\nu$ of random variables *converges almost surely* (Nelson, 1987, p. 26) if almost surely

$$X_n \simeq X_\nu, \qquad n \simeq \infty. \tag{7.1}$$

More precisely, for every $\epsilon \gg 0$ there exists an event N (which may depend on ϵ) such that $\Pr(N) \leq \epsilon$ and

$$X_n(\omega) \simeq X_\nu(\omega), \qquad n \simeq \infty,\ \omega \notin N.$$

7.2 Convergence in Probability

A sequence $X_1, \ldots, X_\nu$ of random variables *converges in probability* (Nelson, 1987, p. 26) if

$$X_n \simeq X_\nu \text{ almost surely}, \qquad n \simeq \infty. \tag{7.2}$$

More precisely, for every $n \simeq \infty$ and $\epsilon \gg 0$ there exists an event N (which may depend on n and ϵ) such that $\Pr(N) \leq \epsilon$ and

$$X_n(\omega) \simeq X_\nu(\omega), \qquad \omega \notin N.$$

7.3 Almost Sure Near Equality

Let us define a notation for $X \simeq Y$ almost surely. Random variables X and Y are *nearly equal almost surely*, written $X \overset{\text{as}}{\simeq} Y$, if $X \simeq Y$ holds almost surely.

Using this notation, we can redefine convergence in probability, replacing (7.2): a sequence ... *converges in probability* if

$$X_n \overset{\text{as}}{\simeq} X_\nu, \qquad n \simeq \infty.$$

Lemma 7.1. *The external relation* $\overset{\text{as}}{\simeq}$ *is an equivalence on* $\mathbb{R}^\Omega$.

Proof. Symmetry and reflexivity come from the corresponding properties of $\simeq$. If $X \overset{\text{as}}{\simeq} Y$ and $Y \overset{\text{as}}{\simeq} Z$, then $X \overset{\text{as}}{\simeq} Z$ by Corollary 3.9 and Lemma 6.1. □

Note that the situation here is much like what we saw with (near) convergence of non-random sequences in Section 4.1. The sequence does very little work. It is enough to understand the external equivalence relation almost sure near equality. We never need to deal with the whole sequence; we always deal with two elements at a time (is $X_m \overset{\text{as}}{\simeq} X_\nu$ or not?)

It is also much like what we saw with the Lebesgue theorem in Section 6.2.2 replacing conventional theorems about sequences (monotone convergence, dominated convergence, Fatou).

The only forms of stochastic convergence in which necessarily involve sequences are almost sure convergence (Section 7.1 above) and the invariance principle (Nelson, 1987, Theorem 18.1), both of which are *sample path* limit theorems. The point is that in (7.1) the exception sets must work for all unlimited n, whereas in (7.2) one may use different exception sets for each n. So (7.1) is a statement about the whole sequence and (7.2) isn't.

7.4 Near Equivalence

A real-valued function g is *limited* if every value is limited. Random variables X and Y defined on possibly different probability spaces are *nearly equivalent* (Nelson, 1987, Chapter 17), written $X \overset{\text{w}}{\simeq} Y$, if

$$E\{g(X)\} \simeq E\{g(Y)\}$$

for every limited (nearly) continuous function g.

Lemma 7.2. *The external relation* $\overset{\text{w}}{\simeq}$ *is an equivalence.*

Proof. Obvious from $\simeq$ being an equivalence. □

7.5 Convergence in Distribution

A sequence $X_1, \ldots, X_\nu$ of random variables defined on possibly different probability spaces *converges in distribution* (also called "weak convergence" or "convergence in law"), if

$$X_n \overset{\text{w}}{\simeq} X_\nu, \qquad n \simeq \infty. \tag{7.3}$$

The situation here is much like what we have seen with every other form of convergence of sequences with the sole exception of almost sure convergence. The sequence does very little work. It is enough to understand the external equivalence relation near equivalence. We never need to deal with the whole

sequence; we always deal with two elements at a time (is $X_m \overset{\text{w}}{\simeq} X_\nu$ or not?) There is so little point to convergence in distribution (over and above near equivalence) that Nelson (1987) does not even bother to define it.

As we saw with (near) convergence of non-random sequences in Section 4.1, and for exactly the same reasons, Nelson-style convergence in distribution is a much stronger property than Kolmogorov-style convergence in distribution. We have, for instance, the same behavior of double sequences

$$X_{ij} \overset{\text{w}}{\simeq} Y_j, \qquad i \simeq \infty$$
$$X_{ij} \overset{\text{w}}{\simeq} Z_i, \qquad j \simeq \infty$$

implies

$$X_{ij} \overset{\text{w}}{\simeq} X_{mn}, \qquad i, j, m, n \simeq \infty.$$

That's just the way equivalence relations behave.

The analogous property does not hold for Kolmogorov-style convergence in distribution (it doesn't even hold when all the random variables are constant random variables and the convergence in distribution is convergence of non-random sequences in disguise).

Now it might be that Nelson-style convergence in distribution is *too strong?* Maybe it is hard to get? But it turns out this is not the case. We get near equivalence when we expect to get it in situations analogous to when Kolmogorov-style convergence in distribution occurs. Thus it seems that Kolmogorov-style convergence in distribution is *too weak*. Nelson-style arguments are often simpler and easier.

Theorem 7.3. *The following are the only implications that hold between the various modes of stochastic convergence.*

(i) *A sequence of random variables that converges almost surely also converges in probability.*

(ii) *A sequence of random variables that converges in probability also converges in distribution.*

(iii) $X \overset{\text{as}}{\simeq} Y$ *implies* $X \overset{\text{w}}{\simeq} Y$.

(The reverse implications are, in general, false.)

Proof. (i) is obvious from the definitions. (iii) obviously implies (ii). Every limited function is L^1 by the Radon-Nikodym theorem (our Theorem 6.4) and hence (iii) holds by the Lebesgue theorem (our Theorem 6.5).

The converses to (ii) and (iii) need not hold, because $X \overset{\text{w}}{\simeq} Y$ does not even require that X and Y be defined on the same probability space, and $X \overset{\text{as}}{\simeq} Y$ does. Moreover, consider X having the uniform distribution on the two-point set $\{-1, 1\}$ and $Y = -X$, then $X \overset{\text{w}}{\simeq} Y$ is true, but $X \overset{\text{as}}{\simeq} Y$ is false.

Nelson (1987, p. 26) gives a counterexample to the converse to (i). Let X_1, ..., X_ν be independent and identically distributed Bernoulli random variables

with $\Pr(X_n = 1) = c/\nu$, where ν is unlimited and c is appreciable. Note that, since X_n is zero-or-one-valued, we have $X_n \simeq 0$ if and only if $X_n = 0$. Since c/ν is infinitesimal, we have $X_n \overset{\text{as}}{\simeq} 0$ for all n. Hence X_n converges in probability to zero.

Let A_μ denote the event

$$X_m = 0, \qquad \mu < m \leq \nu$$

Then

$$\Pr(A_\mu) = \left(1 - \frac{c}{\nu}\right)^{\nu-\mu}$$

and if $\nu - \mu$ is unlimited, we have

$$\Pr(A_\mu) \simeq \exp\left(-c \cdot \frac{\nu - \mu}{\nu}\right) \tag{7.4}$$

by Theorem 4.10.

This makes it impossible for X_n to converge almost surely to zero, because this requires that for every unlimited μ we have $\Pr(A_\mu) \geq 1 - \epsilon$ for every $\epsilon \gg 0$, hence $\Pr(A_\mu) \simeq 1$, which happens only if the argument of the exponential function in (7.4) is infinitesimal for all unlimited μ, and this is not so. □

Chapter 8

The Central Limit Theorem

8.1 Independent and Identically Distributed

Nelson (1987, Chapter 18 and also the discussion on p. 57) gives a theorem that has the following obvious corollary.

Corollary 8.1 (The Central Limit Theorem)**.** *Suppose X_1, X_2, ..., X_ν are independent and identically distributed L^2 random variables with mean μ and variance σ^2, and suppose $\sigma^2 \gg 0$ and $\nu \simeq \infty$. Define*

$$\overline{X}_\nu = \frac{1}{\nu} \sum_{i=1}^{\nu} X_i.$$

Then the random variable

$$Z = \frac{\overline{X}_\nu - \mu}{\sigma / \sqrt{\nu}} \tag{8.1}$$

is L^2 and nearly equivalent to every other such random variable.

The assertion of the theorem is that, no matter what independent and identically distributed L^2 sequence with appreciable variance is chosen, the distribution of (8.1) is the same up to near equivalence.

We say any random variable nearly equivalent to (8.1) is *standard normal*. Note that in Nelson-style theory the term "standard normal" does not name a distribution. It is an *external property* that distributions may or may not have. As with any external property, it is illegal set formation to try to form the set of all standard normal distributions (this set does not exist).

8.2 The De Moivre-Laplace Theorem

In this section we find out more about the limiting distribution in the central limit theorem. The Bernoulli distribution with success probability p has

probability mass function

$$f(k) = \begin{cases} q, & k = 0 \\ p, & k = 1 \end{cases}$$

where $q = 1 - p$. This distribution is abbreviated Bernoulli(p).

The binomial distribution for n trials with success probability p has probability mass function

$$f(k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, \ldots, n. \tag{8.2}$$

This is the distribution of the sum of n IID Bernoulli(p) random variables. This distribution is abbreviated Binomial(n, p).

Theorem 8.2 (De Moivre-Laplace)**.** *Suppose X has the Binomial(n, p) distribution with*

$$n \simeq \infty \tag{8.3a}$$

$$0 \ll p \ll 1 \tag{8.3b}$$

and

$$Z = \frac{X - np}{\sqrt{npq}}.$$

Then there exists a near line T such that

$$\Pr(Z \le z) \simeq \frac{1}{\sqrt{2\pi}} \sum_{\substack{t \in T \\ t \le z}} e^{-t^2/2}\, dt. \tag{8.4}$$

Our proof closely follows Feller (1950, Section 7.2).

Proof. Stirling's approximation for $n!$ is

$$(2\pi)^{1/2} n^{n+1/2} e^{-n} \le n! \le (2\pi)^{1/2} n^{n+1/2} e^{-n+1/(12n)} \tag{8.5}$$

(Feller, 1950, pp. 41–44)[1] Dividing through by the left hand term in (8.5) gives

$$1 \le \frac{n!}{(2\pi)^{1/2} n^{n+1/2} e^{-n}} \le e^{1/(12n)} \tag{8.6}$$

From the continuity of the exponential function for limited values of its argument, we conclude $1 \simeq e^{1/(12n)}$ whenever $1/(12n)$ is infinitesimal, which is whenever n is unlimited. Thus a nonstandard analysis version of Stirling's approximation is

$$n! \sim (2\pi)^{1/2} n^{n+1/2} e^{-n}, \qquad n \simeq \infty. \tag{8.7}$$

[1] Actually, Feller's argument at this point in his book only establishes (8.5) with the factor $(2\pi)^{1/2}$ replaced by an unknown constant. It is only toward the end of the proof of the De Moivre-Laplace theorem that we find out, by comparison with the normalizing constant for the normal distribution, what this unknown constant is. (See footnote 3.)

Interestingly, Feller uses the same $\sim$ notation in his equation for Stirling's approximation, but, of course, he means something conventional (that the ratio of the two sides converges to one as n goes to infinity).

Plugging Stirling's approximation into (8.2) gives

$$\begin{aligned} f(k) &= \binom{n}{k} p^k q^{n-k} \\ &= \frac{n! p^k q^{n-k}}{k!(n-k)!} \\ &\sim \frac{(2\pi)^{1/2} n^{n+1/2} e^{-n} p^k q^{n-k}}{(2\pi)^{1/2} k^{k+1/2} e^{-k} (2\pi)^{1/2} (n-k)^{n-k+1/2} e^{-(n-k)}} \\ &= \frac{1}{(2\pi)^{1/2}} \cdot \frac{n^{n+1/2} p^k q^{n-k}}{k^{k+1/2} (n-k)^{n-k+1/2}} \\ &= \left(\frac{n}{2\pi k(n-k)} \right)^{1/2} \cdot \left(\frac{np}{k} \right)^k \cdot \left(\frac{nq}{n-k} \right)^{n-k} \end{aligned}$$

whenever $k \simeq \infty$ and $n-k \simeq \infty$ [this is our analog of equation (2.5) in Chapter 7 of Feller (1950)].

Now (still following Feller) we define $\delta = k - np$ so $k = np + \delta$ and $n - k = nq - \delta$ and

$$f(k) \sim \left(\frac{n}{2\pi(np+\delta)(nq-\delta)} \right)^{1/2} \cdot \frac{1}{\left(1 + \frac{\delta}{np}\right)^{np+\delta} \left(1 - \frac{\delta}{nq}\right)^{nq-\delta}} \tag{8.8}$$

[this is our analog of equation (2.7) in Chapter 7 of Feller (1950)].

For $\delta = 0$ the right hand side is exactly $1/\sqrt{2\pi npq}$. For $\delta \neq 0$ we use the two-term Taylor series in $\eta = \delta/n$ with remainder to expand the logarithm of the denominator of the second term

$$\begin{aligned} &\log \left[\left(1 + \frac{\delta}{np}\right)^{np+\delta} \left(1 - \frac{\delta}{nq}\right)^{nq-\delta} \right] \\ &\quad = \frac{\delta^2}{2npq} - \left(\frac{1}{p^{*2}} - \frac{1}{q^{*2}} \right) \cdot \frac{\delta^3}{6n^2} \end{aligned}$$

where $p^* = p + \delta^*/n$, $q^* = 1 - p^*$, and $|\delta^*| \leq |\delta|$.

By (8.3b) the term $1/p^{*2} - 1/q^{*2}$ on the right hand side is limited when δ/n is infinitesimal and hence we have

$$\log \left[\left(1 + \frac{\delta}{np}\right)^{np+\delta} \left(1 - \frac{\delta}{nq}\right)^{nq-\delta} \right] \sim \frac{\delta^2}{2npq} \tag{8.9a}$$

whenever

$$\frac{\delta}{n} \simeq 0 \tag{8.9b}$$

[these are our analogs of equations (2.9) and (2.10) in Chapter 7 of Feller (1950)].[2]. Equation (8.9b) together with $n \simeq \infty$ implies both $k \simeq \infty$ and $n - k \simeq \infty$. Thus the only conditions required for (8.8) and (8.9a) to hold are (8.3a), (8.3b), and (8.9b).

Still following Feller, (8.9b) together with (8.3b) implies $np + \delta \sim np$ and $nq - \delta \sim nq$. Hence the first term on the right hand side in (8.8) is asymptotic to $1/\sqrt{2\pi npq}$, and

$$f(k) \sim \left(\frac{1}{2\pi npq}\right)^{1/2} \exp\left(-\frac{\delta^2}{2npq}\right) \tag{8.10}$$

[this is our analog of equation (2.11) in Chapter 7 of Feller (1950)].

We now leave Feller and do some "calculus" nonstandard analysis style. We know from conventional probability theory that we should be interested in the standardized variable

$$z = \frac{k - np}{\sqrt{npq}}$$

in terms of which

$$k = np + z\sqrt{npq}$$

and

$$\delta = z\sqrt{npq}$$

Note that z takes values in the near line

$$T = \{\, (k - np)\epsilon : k = 0, \ldots, n \,\}$$

with regular spacing $\epsilon = 1/\sqrt{npq}$, which by (8.3a) and (8.3b) is infinitesimal. Rewriting (8.10) in terms of z gives

$$f(k) \sim \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \phi(z)\, dz \tag{8.11}$$

where $dz = \epsilon$ is the spacing of T and ϕ is the standard normal density function from conventional probability theory.[3]

Now for any limited numbers a and b with $a < b$ we have by Theorem 4.3

$$\Pr(a < Z < b) \sim \sum_{\substack{z \in T \\ a < z < b}} \phi(z)\, dz \tag{8.12}$$

because, if z is limited and (8.3a) and (8.3b) hold, then

$$\frac{\delta}{n} = z\sqrt{\frac{pq}{n}}$$

[2]Our (8.9b) is simpler than Feller's (2.10) because our $\sim$ relation is simpler than his (because we are using nonstandard analysis).

[3]As mentioned in footnote 1 of this chapter, Feller's argument establishing (8.7) leaves the constant $(2\pi)^{1/2}$ undetermined, and this is the same constant as the $\sqrt{2\pi}$ in (8.11). Now we see that, since the $f(k)$ must sum to one, by (4.7), which was proved by Corollary 4.8 and the discussion following it, the unknown constant must be nearly equal to $\sqrt{2\pi}$.

is infinitesimal, hence (8.9b) holds and hence also (8.11). Actually, we have

$$\Pr(a < Z < b) \simeq \sum_{\substack{z \in T \\ a < z < b}} \phi(z)\, dz \tag{8.13}$$

because the left hand side of (8.12), being less than one, is limited and, if appreciable, nearly equal to the right hand side by Lemma 4.4, and, if infinitesimal, nearly equal to the right hand side by the definition of $\sim$.

Now we know from conventional probability theory (or can easily calculate) that $E(Z) = 0$ and $\operatorname{var}(Z) = 1$. It then follows from Chebyshev's inequality (Nelson, 1987, p. 5) that

$$\Pr(|Z| \geq a) \leq \frac{1}{a^2}.$$

Hence Z is limited almost surely, and by Lemma 6.3 for any $\epsilon \gg 0$ there exists a limited a such that $\Pr(|Z| \geq a) \leq \epsilon$. Hence for limited b

$$\left| \Pr(Z < b) - \sum_{\substack{z \in T \\ z < b}} \phi(z)\, dz \right| \lesssim \epsilon$$

and, since $\epsilon \gg 0$ was arbitrary, we have (8.4) for all limited z. Because Z is limited almost surely, both sides of (8.4) are infinitesmal when $z \simeq -\infty$ and nearly equal to one when $z \simeq +\infty$. □

Corollary 8.3. *For any near line T, the distribution having distribution function F defined by*

$$F(z) = \frac{1}{c} \sum_{\substack{t \in T \\ t \leq z}} e^{-t^2/2}\, dt,$$

where

$$c = \sum_{t \in T} e^{-t^2/2}\, dt,$$

is standard normal, and

$$F(z) \simeq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

for all $z \in \mathbb{R}$.

Proof. Apply Corollary 4.8. □

Later, after we have learned more about distribution functions and near equivalence, we will improve this corollary to an "if and only if" statement (Corollary 10.6).

Chapter 9

Near Equivalence in Metric Spaces

9.1 Metric Spaces

Let (S, d) be a metric space, that is, S is a set and d is a metric for S. We say points x and y of S are *nearly equal* and write $x \simeq y$ when $d(x, y)$ is infinitesimal. Note that our definition of $x \simeq y$ in $\mathbb{R}$ is a special case of this generalization when we use the usual metric for $\mathbb{R}$ defined by $d(x, y) = |x - y|$.

Let (S, d) and (S', d') be metric spaces. A function $h : S \to S'$ is *nearly continuous at a point* $x \in S$ if

$$y \in S \text{ and } x \simeq y \longrightarrow h(x) \simeq h(y), \tag{9.1}$$

and *nearly continuous on a set* $T \subset S$ if (9.1) holds at all $x \in T$.

Random elements X and Y of a metric space (S, d) defined on possibly different probability spaces are *nearly equivalent* written $X \overset{\text{w}}{\simeq} Y$, if

$$E\{g(X)\} \simeq E\{g(Y)\}$$

for every limited (nearly) continuous function $g : S \to \mathbb{R}$.

Let d_1 and d_2 be two different metrics for the same set S. We say that d_1 and d_2 are *equivalent* if they agree as to the meaning of $x \simeq y$, that is, if

$$d_1(x, y) \simeq 0 \longleftrightarrow d_2(x, y) \simeq 0.$$

In this case, a function $h : S \to S'$ where (S', d') is another metric space is (nearly) continuous when d_1 is the metric for S if and only if it is continuous when d_2 is the metric. In this sense, near equivalence of random elements of metric spaces does not depend on the metric but only on the external equivalence relation $\simeq$ induced by the metric.

9.2 Probability Measures

A random element X of a metric space (S, d) induces a probability measure P defined by (5.11a) and probability mass function p defined by (5.11b), a relation denoted by $p = dP$. This relationship between X and P is denoted $P = \mathcal{L}(X)$, and we say P is the *law of* X.

Since near equivalence is determined by expectations, and expectations are determined by measures, near equivalence really depends only on measures not on random elements (except through their measures). Thus we make the definition: measures P and Q on a metric space (having finite support) are *nearly equivalent*, written $P \overset{\text{w}}{\simeq} Q$ if $Ph \simeq Qh$ for every limited nearly continuous function $h : S \to \mathbb{R}$, where

$$Ph = \sum_{x \in S} h(x) dP(x)$$

is a shorthand for the expectation of the random variable $h(X)$ when $P = \mathcal{L}(X)$.

9.3 The Prohorov Metric

If d is a metric on S we define the notation

$$d(x, A) = \inf_{y \in A} d(x, y)$$

for any nonempty subset A of S. Then we define the notation

$$A^\epsilon = \{\, x \in S : d(x, A) < \epsilon \,\}$$

for any nonempty subset A of S and any $\epsilon > 0$. For completeness, we define $\varnothing^\epsilon = \varnothing$ for $\epsilon > 0$. Note that the "open" ball of radius ϵ centered at x can be denoted $\{x\}^\epsilon$, and our definition for the empty set makes

$$A^\epsilon = \bigcup_{x \in A} \{x\}^\epsilon$$

hold for all A (rather than just nonempty A). The set A^ϵ is called the *ϵ-dilation of A*.

The triangle inequality implies $(A^\epsilon)^\eta \subset A^{\epsilon+\eta}$ because $z \in (A^\epsilon)^\eta$ when there exist $x \in A$ and $y \in S$ such that $d(x, y) < \epsilon$ and $d(y, z) < \eta$.

Let S be a finite set with metric d and let $\mathcal{P}(S, d)$ denote the set of all probability measures on S. The *Prohorov metric* on $\mathcal{P}(S, d)$ is the function $\pi : \mathcal{P}(S, d) \times \mathcal{P}(S, d) \to \mathbb{R}$ defined so that $\pi(P, Q)$ is the infimum of all $\epsilon > 0$ such that

$$P(A) \le Q(A^\epsilon) + \epsilon \text{ and } Q(A) \le P(A^\epsilon) + \epsilon, \qquad A \subset S. \tag{9.2}$$

(When P and Q have finite support the infimum is achieved.)

Note that (9.2) holds for every $\epsilon > \pi(P, Q)$ because if (9.2) holds for some ϵ, then for any $\eta > 0$

$$P(A) \le P(A^\eta) \le Q\big((A^\eta)^\epsilon\big) + \epsilon \le Q(A^{\eta+\epsilon}) + \epsilon,$$

and similarly with P and Q swapped, so (9.2) also holds with ϵ replaced by $\epsilon + \eta$. Hence the set of ϵ for which (9.2) holds is the union of intervals of the form $[\epsilon, \infty)$ and hence has one of the forms (δ, ∞) or $[\delta, \infty)$, only the latter being possible when P and Q have finite support. In either case δ is the Prohorov distance between P and Q.

In the following theorem and the rest of this chapter we restrict our attention to Nelson-style probability theory: all probability measures have finite support.

Theorem 9.1. *π is a metric.*

Proof. The properties symmetry, nonnegativity, and $\pi(P, P) = 0$ are obvious. If $P \ne Q$ then $p = dP \ne q = dQ$ and there exists an x such that $p(x) \ne q(x)$. Define $B = \operatorname{supp} P \cup \operatorname{supp} Q$, where $\operatorname{supp} P$ denotes the support of P (defined in Section 5.5). If $B \ne \{x\}$, then define

$$\epsilon_1 = \min\{\, d(x, y) : y \in B \setminus \{x\} \,\},$$

otherwise (when $B = \{x\}$) define $\epsilon_1 = 1$. Then $\epsilon_1 > 0$ because B is a finite set, and for $0 < \epsilon < \epsilon_1$ we have

$$P(\{x\}^\epsilon) = p(x) \qquad \text{and} \qquad Q(\{x\}^\epsilon) = q(x),$$

hence $\pi(P, Q) \ge |p(x) - q(x)|$. The triangle inequality follows from

$$\begin{aligned} P(A) &\le Q(A^\epsilon) + \epsilon, \qquad \epsilon > \pi(P, Q) \\ Q(A^\epsilon) &\le R(A^{\epsilon+\eta}) + \eta, \qquad \eta > \pi(Q, R) \end{aligned}$$

hence

$$P(A) \le R(A^{\epsilon+\eta}) + \epsilon + \eta, \qquad \epsilon + \eta > \pi(P, Q) + \pi(Q, R)$$

and similarly with P and R swapped. □

Another useful fact about the Prohorov metric is the following (copied essentially verbatim from Billingsley (1999, p. 72), because the argument uses no measure theory so Kolmogorov-style and Nelson-style argument is the same).

Lemma 9.2. *The Prohorov distance between P and Q is the infimum over all ϵ such that*

$$P(A) \le Q(A^\epsilon) + \epsilon, \qquad A \subset S. \tag{9.3}$$

Proof. First note that $A \subset S \setminus B^\epsilon$ if and only if $B \subset S \setminus A^\epsilon$ because either is the same as $(\forall x \in A)(\forall y \in B)\big(d(x, y) \ge \epsilon\big)$. If (9.3) holds, let $B = S \setminus A^\epsilon$, and then

$$P(A^\epsilon) = 1 - P(B) \ge 1 - Q(B^\epsilon) - \epsilon = Q(S \setminus B^\epsilon) - \epsilon \ge Q(A) - \epsilon$$

so (9.3) also holds with P and Q swapped, and hence (9.2) holds. □

9.4 Near Equivalence and the Prohorov Metric

Let (S, d) and (S', d') be metric spaces. A function $h : S \to S'$ is *nearly Lipschitz continuous* if there exists a limited number L such that

$$d'\big(h(x), h(y)\big) \le L \cdot d(x, y), \qquad x, y \in S.$$

Note that, the product of limited and infinitesimal numbers being infinitesimal, a (nearly) Lipschitz continuous function is, as the name suggests, (nearly) continuous.

Lemma 9.3. *Let (S, d) be a metric space. For any nonempty $A \subset S$ and any $\epsilon \gg 0$, define $h : S \to \mathbb{R}$ by*

$$h(x) = \max\big(0, 1 - d(x, A)/\epsilon\big).$$

Then h is limited and nearly Lipschitz continuous.

Proof. We first establish

$$|h(x) - h(y)| \le \frac{|d(x, A) - d(y, A)|}{\epsilon}. \tag{9.4}$$

(Case I) Suppose $d(x, A) = 0$ so $h(x) = 1$. Then

$$h(x) - h(y) = 1 - h(y) = \begin{cases} 1, & d(y, A) \ge \epsilon \\ d(y, A)/\epsilon & d(y, A) < \epsilon \end{cases}$$

and in either case (9.4) holds.

(Case II) Suppose $d(x, A) \ge \epsilon$ so $h(x) = 0$. Then

$$h(y) - h(x) = h(y) = \begin{cases} 0, & d(y, A) \ge \epsilon \\ 1 - d(y, A)/\epsilon, & d(y, A) < \epsilon \end{cases}$$

and in either case (9.4) holds.

(Case III) Suppose $0 < d(x, A) < \epsilon$ so $0 < h(x) < 1$ and similarly with y replacing x. Then

$$h(y) - h(x) = \frac{d(y, A) - d(x, A)}{\epsilon}$$

and again (9.4) holds. This finishes the proof of (9.4) (the other cases being like I and II with x and y swapped).

Now for any $z \in A$ we have

$$d(x, y) + d(y, z) \ge d(x, z) \ge d(x, A)$$

taking the infimum over all $z \in A$ gives

$$d(x, y) + d(y, A) \ge d(x, A)$$

and similarly with x and y swapped. Hence

$$d(x,y) \geq |d(y,A) - d(x,A)|$$

from which we see that (9.4) implies

$$|h(x) - h(y)| \leq \frac{|d(x,A) - d(y,A)|}{\epsilon} \leq \frac{1}{\epsilon} \cdot d(x,y).$$

And, $1/\epsilon$ being limited, this establishes the lemma (and incidentally shows that the "Lipschitz constant" can be taken to be $L = 1/\epsilon$). □

Theorem 9.4. *Assume P and Q are measures on (S,d) having finite support and that $Ph \simeq Qh$ holds for every limited nearly Lipschitz continuous function $h : S \to \mathbb{R}$. Then $\pi(P,Q) \simeq 0$.*

Proof. For any nonempty $A \subset S$ and any $\epsilon \gg 0$, define $h : S \to \mathbb{R}$ as in the lemma. Then h is limited and nearly Lipschitz continuous. So $Ph \simeq Qh$. But

$$P(A) \leq Ph \leq P(A^\epsilon)$$

and similarly with P replaced by Q. Hence $P(A) \leq Ph \simeq Qh \leq Q(A^\epsilon)$ holds for all A. Thus (9.3) holds for every $\epsilon \gg 0$ and hence $\pi(P,Q)$ is infinitesimal. □

Theorem 9.5. *Assume P and Q are measures on (S,d) having finite support and $\pi(P,Q) \simeq 0$. Then $P \overset{\text{w}}{\simeq} Q$.*

Proof. It is enough to prove $Ph \simeq Qh$ for all nearly continuous h satisfying $0 < h < 1$ because for any limited nearly continuous function g there exist limited a and b such that $g = a + bh$ and $0 < h < 1$, in which case $Ph \simeq Qh$ implies $Pg \simeq Qg$.

But for such h we have by Lemma 5.2 the representation

$$Ph = \sum_{t \in T} P\{h > t\}\, dt$$

where T is a finite subset of $[0,1]$ containing 0, 1, and $h(\operatorname{supp} P)$ and where the notation $\{h > t\}$ denotes the event $h(X) > t$ and $P\{h > t\}$ denotes the probability of this event, where X is a random element such that $P = \mathcal{L}(X)$. Note that $t \mapsto 1 - P\{h > t\}$ is the distribution function of $h(X)$.

For arbitrary t in $[0,1]$ define $A_t = \{h > t\}$. Fix an infinitesimal ϵ greater than $\pi(P,Q)$. Then for $x \in A_t^\epsilon$ we have $h(x) \gtrsim t$ and hence $h(x) > t - \delta$ for any $\delta \gg 0$. Hence $A_t \subset A_t^\epsilon \subset \{h > t - \delta\}$. Thus we have

$$\begin{aligned}
&P(A_t) = P\{h > t\} \leq P(A_t^\epsilon) \leq P\{h > t - \delta\} \\
&P(A_t) \lesssim Q(A_t^\epsilon) \\
&Q(A_t) \lesssim P(A_t^\epsilon) \\
&Q(A_t) = Q\{h > t\} \leq Q(A_t^\epsilon) \leq Q\{h > t - \delta\}
\end{aligned}$$

from which we infer

$$P\{h > t\} \lesssim Q\{h > t - \delta\} \quad \text{and} \quad Q\{h > t\} \lesssim P\{h > t - \delta\} \tag{9.5a}$$

holds for all t in $[0, 1]$ and all $\delta \gg 0$. To simplify notation, we define $U(t) = P\{h > t\}$ and $V(t) = Q\{h > t\}$ so (9.5a) becomes

$$U(s) \gtrsim V(t) \quad \text{and} \quad V(s) \gtrsim U(t), \qquad \text{whenever } s \ll t. \tag{9.5b}$$

We also define $U_+(t) = P\{h \geq t\}$ and $V_+(t) = Q\{h \geq t\}$.

Now for any limited n choose s_i such that $s_0 = 0$, $s_n = 1$, and

$$U(s_i) \leq \frac{n - i}{n} < U_+(s_i), \qquad i = 1, \ldots, n - 1.$$

It may be that some of the s_i are equal, so we remove duplicates. Let

$$\{s_0, s_1, \ldots, s_n\} = \{s_0^*, s_1^*, \ldots, s_{m_n}^*\}$$

maintaining $s_0^* < s_1^* < \cdots < s_{m_n}^*$.

Now for any $\delta \gg 0$ we have

$$V(s_i^* - \delta) \gtrsim U(s_i^* - \delta/2) \geq U_+(s_i^*) > U(s_i^*) \gtrsim V(s_i^* + \delta) \tag{9.6}$$

hence

$$V(s_i^* - \delta) + \delta \geq U_+(s_i^*) > U(s_i^*) \geq V(s_i^* + \delta) - \delta \tag{9.7}$$

holds for all $\delta \gg 0$ and hence by overspill (9.7) must hold for some infinitesimal δ. For this infinitesimal δ, define $r_i = s_i^* - \delta$ and $t_i = s_i^* + \delta$ for $i = 1$, …, $m_n - 1$, and also define $r_0 = t_0 = 0$ and $r_{m_n} = t_{m_n} = 1$.

Now

$$Ph \gtrsim \sum_{i=1}^{m_n} (s_i^* - s_{i-1}^*) U_+(s_i^*) \tag{9.8a}$$

and

$$Qh \lesssim \sum_{i=1}^{m_n} (r_i - t_{i-1}) V(t_{i-1}) \tag{9.8b}$$

because the contributions from the infinitesimal intervals (t_i, r_i) are negligible.

For some integer k

$$V(t_{i-1}) \lesssim U(s_{i-1}^*) \leq \frac{n - k + 1}{n}$$
$$\frac{n - k}{n} < U_+(s_i^*)$$

Combining these gives

$$V(t_{i-1}) \lesssim U_+(s_i^*) + \frac{1}{n} \tag{9.8c}$$

And combining (9.8c) with (9.8a) and (9.8b) and using the fact that $s_i^* - s_{i-1}^* \simeq r_i - t_{i-1}$ for all i gives $Qh \lesssim Ph + 1/n$. The same argument with P and Q swapped gives $Ph \lesssim Qh + 1/n$. Since these hold for any limited n, we must have $Ph \simeq Qh$. □

9.5 The Portmanteau Theorem

Let (S, d) be a metric space. Define for $A \subset S$

$$A_\epsilon = S \setminus (S \setminus A)^\epsilon.$$

The set A_ϵ is called the *ϵ-erosion of A*.

Theorem 9.6. *Let P and Q be measures on (S, d) having finite support and π the Prohorov metric. The following are equivalent*

(i) *P and Q are nearly equivalent.*

(ii) *$Ph \simeq Qh$ for every limited Lipschitz continuous function $h : S \to \mathbb{R}$.*

(iii) *$\pi(P, Q)$ is infinitesimal.*

(iv) *For some $\epsilon \simeq 0$ and for all $A \subset S$ we have $P(A) \lesssim Q(A^\epsilon)$.*

(v) *For every $\epsilon \gg 0$ and for all $A \subset S$ we have $P(A) \lesssim Q(A^\epsilon)$.*

(vi) *For some $\epsilon \simeq 0$ and for all $A \subset S$ we have $P(A) \gtrsim Q(A_\epsilon)$.*

(vii) *For every $\epsilon \gg 0$ and for all $A \subset S$ we have $P(A) \gtrsim Q(A_\epsilon)$.*

Proof. The implication (i) $\to$ (ii) is trivial. The implication (ii) $\to$ (iii) is Theorem 9.4. The implication (iii) $\to$ (i) is Theorem 9.5.

By Lemma 9.2 (iii) is equivalent to the existence of an $\epsilon \simeq 0$ such that (9.3) holds, which implies (iv). The implication (iv) $\to$ (v) is trivial. If (v) holds, then (9.3) holds for every $\epsilon \gg 0$, and hence by Lemma 9.2 (iii) holds. At this point we have established the equivalence of (i) through (v).

The implications (iv) $\longleftrightarrow$ (vi) and (v) $\longleftrightarrow$ (vii) are just the complement rule: write $B = S \setminus A$ so

$$P(A) = 1 - P(B) \text{ and } Q(A^\epsilon) = 1 - Q(B_\epsilon)$$

and

$$P(A) \lesssim Q(A^\epsilon) \longleftrightarrow P(B) \gtrsim Q(B_\epsilon)$$

hence if the left hand side holds for all A, then the right hand side holds for all B and vice versa. □

Let S and T be two different sets with $S \subset T$ and d a metric for T and hence also for S. Pedantically, the *restriction* d_r of d to $S \times S$ is a metric for S and defines (S, d_r) as a metric subspace of (T, d). Then elements P and Q of $\mathcal{P}(S, d_r)$ are nearly equivalent if and only if they are nearly equivalent considered as elements of $\mathcal{P}(T, d)$. The enclosing superspace T is irrelevant. This is clear from (iv) of the portmanteau theorem. Since $P(A) = P(A \cap \operatorname{supp} P)$, we need only check $A \subset \operatorname{supp} P$ in establishing (iv). So long as we are only interested in two measures P and Q, we can take $S = \operatorname{supp} P \cup \operatorname{supp} Q$, if we like, but any enclosing metric space does as well.

9.6 Continuous Mapping

Let A be any property that may or may not hold at points of S and let P be a measure on (S, d) having finite support. We say A holds *P-almost everywhere* if for every $\epsilon \gg 0$ there exists a set N such that $P(N) \leq \epsilon$ and $A(x)$ holds except for x in N.

The following lemma is Theorem 17.3 in Nelson (1987). We reprove it here only to see how much shorter the proof gets with our extra apparatus.

Lemma 9.7. *Let P and Q be elements of $\mathcal{P}(S, d)$ such that $P \overset{\text{w}}{\simeq} Q$, and let A be any property (internal or external) such that $x, y \in S$ and $x \simeq y$ implies $A(x) \longleftrightarrow A(y)$. Then A holds P-almost everywhere if and only if it holds Q-almost everywhere.*

Proof. Suppose A holds P-almost everywhere. Fix $\epsilon \gg 0$ and choose $N \subset S$ such that $P(N) \leq \epsilon/2$ and $A(x)$ holds for all $x \in S \setminus N$. By (vi) of the portmanteau theorem there is a $\delta \simeq 0$ such that $P(N) \gtrsim Q(N_\delta)$. So $Q(N_\delta) \leq \epsilon$. By definition $S \setminus N_\delta = (S \setminus N)^\delta$. Hence $y \in S \setminus N_\delta$ if and only if there exists an x in $S \setminus N$ such that $d(x, y) < \delta$. This implies $x \simeq y$, and hence $A(y)$ holds. Hence A holds on $S \setminus N_\delta$. Since $\epsilon \gg 0$ was arbitrary, A holds Q-almost everywhere. □

Let P be a measure on (S, d) having finite support. A function $h : S \to S'$ is *nearly continuous P-almost everywhere* if the property $A(x)$ in the definition of "almost everywhere" is "h is nearly continuous at x." Note that by the lemma, when $P \overset{\text{w}}{\simeq} Q$ we have h nearly continuous P-almost everywhere if and only if h is nearly continuous Q-almost everywhere.

For an arbitrary map $h : S \to S'$ and any measure P on (S, d), the image measure $P' \in (S', d')$ denoted by $P \circ h^{-1}$ is defined by

$$P'(B) = P\big(h^{-1}(B)\big), \qquad B \subset S'.$$

If a random element X has the distribution P, then the random element $h(X)$ has the distribution $P \circ h^{-1}$.

Theorem 9.8 (Continuous Mapping)**.** *Let (S, d) and (S', d') metric spaces, and let $P, Q \in \mathcal{P}(S, d)$. Suppose $P \overset{\text{w}}{\simeq} Q$, and suppose $h : S \to S'$ is nearly continuous P-almost everywhere. Then $P \circ h^{-1} \overset{\text{w}}{\simeq} Q \circ h^{-1}$.*

Proof. Fix $\epsilon \gg 0$ and choose $N \subset S$ such that $Q(N) \leq \epsilon/2$ and h is nearly continuous on $S \setminus N$. Let B' be an arbitrary subset of S', and write $B = h^{-1}(B')$. Also define $P' = P \circ h^{-1}$ and $Q' = Q \circ h^{-1}$.

Then $P(B) = P'(B')$. By (iii) of the portmanteau theorem there exists a $\delta \simeq 0$ such that $P(B) \lesssim Q(B^\delta)$. By near continuity, h maps $B^\delta \setminus N$ into B'^ϵ, which implies $Q(B^\delta \setminus N) \leq Q'(B'^\epsilon)$. Hence

$$P'(B') = P(B) \lesssim Q(B^\delta \setminus N) + Q(N) \leq Q'(B'^\epsilon) + \frac{\epsilon}{2}$$

Reading from end to end we have

$$P'(B') \le Q'(B'^{\epsilon}) + \epsilon$$

and since $\epsilon \gg 0$ and $B' \subset S'$ were arbitrary, this implies by Lemma 9.2 that the Prohorov distance between P' and Q' is infinitesimal. □

Corollary 9.9. *Let (S, d) and (S', d') metric spaces, and let X and Y be random elements of (S, d) such that $X \overset{\text{w}}{\simeq} Y$, and suppose $h : S \to S'$ is nearly continuous P-almost everywhere, where $P = \mathcal{L}(X)$. Then $h(X) \overset{\text{w}}{\simeq} h(Y)$.*

9.7 Product Spaces

Let (S_1, d_1) and (S_2, d_2) be metric spaces, then we make $S_1 \times S_2$ into a metric space by giving it the metric d^* defined by

$$d^*\big((x_1, x_2), (y_1, y_2)\big) = d'\big(d_1(x_1, y_1), d_2(x_2, y_2)\big)$$

where d' is any metric on $\mathbb{R}^2$ that is equivalent to one inducing the standard topology, for example, we may use any of

$$d'(u, v) = |u| + |v| \tag{9.9a}$$

$$d'(u, v) = \sqrt{u^2 + v^2} \tag{9.9b}$$

$$d'(u, v) = \max(|u|, |v|) \tag{9.9c}$$

which are referred to as the L^1, L^2 and L^∞ norms, respectively. We know from the comment at the end of Section 9.5 that it does not matter which d' we use, since they all give the same notion of near equivalence in $\mathcal{P}(S_1 \times S_2, d^*)$.

For our purposes here (9.9c) is the most useful, because of its special property that

$$(B_1 \times B_2)^\epsilon = B_1^\epsilon \times B_2^\epsilon, \qquad B_1 \subset S_1,\ B_2 \subset S_2 \tag{9.10}$$

making ϵ-dilations particularly easy to work with.

Let $p_i : S_1 \times S_2 \to S_i$ denote the coordinate projection $(x_1, x_2) \mapsto x_i$. Then for any $P \in \mathcal{P}(S_1 \times S_2, d^*)$ the *marginals* of P are the distributions $P \circ p_i^{-1}$.

Theorem 9.10 (Slutsky)**.** *Suppose $P, Q \in \mathcal{P}(S_1 \times S_2, d^*)$ and suppose*

$$P \circ p_i^{-1} \overset{\text{w}}{\simeq} Q \circ p_i^{-1}, \qquad i = 1, 2 \tag{9.11}$$

and suppose $\operatorname{supp}(Q \circ p_2^{-1})$ *is a singleton. Then $P \overset{\text{w}}{\simeq} Q$.*

Proof. Suppose without loss of generality that (9.9c) is used to define d^* so (9.10) holds. Write $P \circ p_i^{-1} = P_i$ and $Q \circ p_i^{-1} = Q_i$, and write $\{c\} = \operatorname{supp} Q_2$. Choose an infinitesimal ϵ greater than either of the Prohorov distances between opposite sides of (9.11). For $B \subset S_1 \times S_2$ define

$$B_1 = \{\, x_1 \in S_1 : (x_1, c) \in B \,\}.$$

Then $Q(B) = Q(B_1 \times \{c\}) = Q_1(B_1)$.

Also, $Q_2(\{c\}) = 1$ implies $P_2(\{c\}^\epsilon) \simeq 1$ by the portmanteau theorem and

$$P_2(S_2 \setminus \{c\}^\epsilon) \simeq 0. \tag{9.12}$$

Moreover, $B \supset B_1 \times \{c\}$ implies

$$B^\epsilon \supset (B_1 \times \{c\})^\epsilon = B_1^\epsilon \times \{c\}^\epsilon,$$

hence

$$\begin{aligned} P(B^\epsilon) &\geq P(B_1^\epsilon \times \{c\}^\epsilon) \\ &= P_1(B_1^\epsilon) - P\big(B_1^\epsilon \times (S_2 \setminus \{c\}^\epsilon)\big) \\ &\gtrsim P_1(B_1^\epsilon) \end{aligned}$$

the second term on the middle line being infinitesimal because of (9.12). Hence we have

$$Q(B) = Q_1(B_1) \lesssim P_1(B_1^\epsilon) \lesssim P(B^\epsilon),$$

the middle relation being an application of Lemma 9.2 and the other relations having already been established. Since B was arbitrary, we have (iv) of the portmanteau theorem. □

Corollary 9.11. *Suppose $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ are random elements of $(S_1 \times S_2, d^*)$ and Y_2 is a constant random element, and suppose*

$$\begin{aligned} X_1 &\overset{\text{w}}{\simeq} Y_1 \\ X_2 &\overset{\text{w}}{\simeq} Y_2 \end{aligned}$$

then

$$(X_1, X_2) \overset{\text{w}}{\simeq} (Y_1, Y_2).$$

Chapter 10

Distribution Functions

10.1 The Lévy Metric

The *Lévy metric* on the set of all distribution functions (with finite support) on $\mathbb{R}$ is the function λ defined so that $\lambda(F, G)$ is the infimum of all $\epsilon > 0$ such that

$$F(x) \leq G(x + \epsilon) + \epsilon \text{ and } G(x) \leq F(x + \epsilon) + \epsilon, \qquad x \in \mathbb{R}. \tag{10.1}$$

It is easy to see that λ actually is a metric.

Actually, the Lévy metric can be applied to any nondecreasing functions $\mathbb{R} \to \mathbb{R}$. We shall do this in Corollary 10.6 below, where we consider Lévy distance between the distribution function of a Nelson-style random variable and the distribution function Φ of the normal distribution in Kolmogorov-style probability theory.

We say nondecreasing functions F and G are *nearly equal*, written $F \simeq G$, if $\lambda(F, G) \simeq 0$. Caution: this does not necessarily imply that random variables having distribution functions F and G are nearly equivalent! See Theorem 10.4 below.

Lemma 10.1. *Suppose F and G are nondecreasing functions and G is nearly continuous, Then $F \simeq G$ if and only if $F(x) \simeq G(x)$, for all $x \in \mathbb{R}$.*

Proof. One direction is trivial. Conversely, suppose $\lambda(F, G) = \epsilon \simeq 0$. Then by near continuity of G

$$G(x) \simeq G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon \simeq G(x)$$

holds for all x. □

Lemma 10.2. *If λ is the Lévy metric and π the Prohorov metric, F and G are distribution functions and P and Q are the corresponding measures, then*

$$\pi(P, Q) \geq \lambda(F, G).$$

Proof. Fix $\epsilon > \pi(P, Q)$. Then

$$\begin{aligned} F(x) &= P\{(-\infty, x]\} \\ &\le Q\{(-\infty, x]^\epsilon\} + \epsilon \\ &= Q\{(-\infty, x + \epsilon)\} + \epsilon \\ &= \max_{y<x} G(y + \epsilon) + \epsilon \end{aligned}$$

holds for all x. In particular we have

$$F(x) \le G(x + \epsilon) + \epsilon \tag{10.2}$$

whenever $x + \epsilon$ is not a jump of G. However, even if $x + \epsilon$ is a jump of G, there exists a $\delta > 0$ sufficiently small so that G has no jump in $(x + \epsilon, x + \epsilon + \delta]$, and, applying (10.2) with ϵ replaced by $\epsilon + \delta$, we have

$$F(x) \le G(x + \epsilon + \delta) + \epsilon + \delta = G(x + \epsilon) + \epsilon + \delta,$$

which, since $\delta > 0$ was arbitrary, implies (10.2) even when $x + \epsilon$ is a jump of G. The same argument with F and G swapped finishes the proof. □

10.2 Near Equivalence

Corollary 10.3. *If X and Y are random variables having distribution functions F and G, then $X \overset{\text{w}}{\simeq} Y$ implies $F \simeq G$.*

Theorem 10.4. *If X and Y are random variables having distribution functions F and G, either X or Y is limited almost surely, and $F \simeq G$, then $X \overset{\text{w}}{\simeq} Y$.*

The limited almost surely condition cannot be suppressed. Let X have the uniform distribution on the even integers between 1 and 2ν and Y have the uniform distribution on the odd integers between 1 and 2ν. Then if F and G are the corresponding distribution functions, then we have

$$0 \le G(x) - F(x) \le \frac{1}{\nu}$$

for all x. So $\lambda(F, G)$ is infinitesimal whenever ν is unlimited. But, if h is defined by $h(x) = \sin^2(\pi x)$, then h is a limited continuous function, and $h(X) = 0$ and $h(Y) = 1$ (for all ω).

Proof. Fix arbitrary appreciable ϵ_1 and ϵ_2. By Lemma 9.7, if one of X and Y is limited almost surely and $X \overset{\text{w}}{\simeq} Y$, then the other is also limited almost surely. Hence, by Lemma 6.3, there exist limited a and b such that

$$F(a) \le \epsilon_1 \tag{10.3a}$$

$$F(b) \ge 1 - \epsilon_1 \tag{10.3b}$$

and similarly with F replaced by G. Let h be a nearly continuous function with limited bound M. Then

$$|E\{h(X)I_{(-\infty,a]}(X)\}| \le MF(a) \le M\epsilon_1 \tag{10.4a}$$

and

$$|E\{h(X)I_{(b,\infty)}(X)\}| \le M\big[1 - F(b)\big] \le M\epsilon_1 \tag{10.4b}$$

and similarly with X replaced by Y and F replaced by G.

By near continuity of h there exists $\delta \gg 0$ such that $|h(x) - h(y)| \le \epsilon_2$ whenever $|x - y| \le \delta$. There exists a limited natural number n such that $(b-a)/n \le \delta/2$. Define $c_k = a + (k/n)(b-a)$ for integer k, noting that $c_0 = a$ and $c_n = b$. Then

$$|h(x) - h(c_k)| \le \epsilon_2, \qquad c_{k-2} \le x \le c_{k+2}, \tag{10.5}$$

and this together with (10.4a) and (10.4b) implies

$$\left| E\{h(X)\} - \sum_{k=1}^{n+1} h(c_k)\big[F(c_k) - F(c_{k-1})\big] \right| \le \epsilon_2 + 2M\epsilon_1. \tag{10.6}$$

The assumption $\lambda(F,G) \simeq 0$ implies that there exist b_k and d_k such that $b_k \le c_k \le d_k$ and $b_k \simeq c_k \simeq d_k$ and

$$G(b_k) \lesssim F(c_k) \lesssim G(d_k), \qquad \text{for all } k.$$

The same reasoning that lead to (10.6) implies that (10.6) holds with X replaced by Y and F replaced by G. But

$$\begin{aligned} \sum_{k=1}^{n+1} h(c_k)\big[F(c_k) - F(c_{k-1})\big] - \sum_{k=1}^{n+1} h(c_k)\big[G(b_k) - G(b_{k-1})\big] \\ = \sum_{k=1}^{n} \big[h(c_k) - h(c_{k+1})\big] \cdot \big[F(c_k) - G(b_k)\big] \\ + h(c_{n+1})\big[F(c_{n+1}) - G(b_{n+1})\big] - h(c_1)\big[F(c_0) - G(b_0)\big] \end{aligned} \tag{10.7}$$

and $0 \lesssim F(c_k) - G(b_k) \lesssim G(d_k) - G(b_k)$, and the latter sum to less or equal to one. Hence, a limited sum of infinitesimals being infinitesimal (Corollary 3.5), the sum in (10.7) is weakly less than ϵ_2 and weakly greater than zero. The other terms are less than or equal to $4M\epsilon_1$ in absolute value. That is,

$$\left| \sum_{k=1}^{n+1} h(c_k)\big[F(c_k) - F(c_{k-1})\big] - \sum_{k=1}^{n+1} h(c_k)\big[G(b_k) - G(b_{k-1})\big] \right| \le \epsilon_2 + 4M\epsilon_1.$$

Hence by the triangle inequality

$$|E\{h(X)\} - E\{h(Y)\}| \le 3\epsilon_3 + 8M\epsilon_1.$$

Since ϵ_1 and ϵ_2 were arbitrary appreciable numbers and M is limited, we actually have $E\{h(X)\} \simeq E\{h(Y)\}$. □

10.3 General Normal Distributions

Standard normal random variables were defined in Section 8.1. They are the distributions that arise as limits in the central limit theorem (Corollary 8.1). We found out more about these distributions in Theorem 8.2 and Corollary 8.3. Here we finish the job.

Lemma 10.5. *A standard normal random variable is limited almost surely.*

Proof. It is established near the end of the proof of Theorem 8.2 that the random variable Z defined in the theorem statement, which is standard normal, is limited almost surely. Hence by Lemma 9.7 every standard normal random variable is limited almost surely. □

Corollary 10.6. *A random variable is standard normal if and only if its distribution function is nearly equal to Φ defined by*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \, dt. \tag{10.8}$$

Proof. Theorem 8.2 and Corollary 8.3 assert that one particular standard normal random variable has distribution function nearly equal to Φ. Hence by Corollary 10.3, Theorem 10.4, and Lemma 10.5 a random variable is standard normal if and only if its distribution function is nearly equal to Φ. □

Note that Φ is nearly continuous by Lemma 4.1, so by Lemma 10.1 and the preceding lemma a random variable is standard normal if and only if its distribution function F satisfies

$$F(x) \simeq \Phi(x), \qquad x \in \mathbb{R}.$$

If Z is a standard normal random variable, μ is a limited real number, and σ is a positive appreciable real number, then we say $X = \mu + \sigma Z$ is *general normal* and we also apply this terminology to the distribution of X. Like standard normality, general normality is an external property.

Analogies with Kolmogorov-style probability theory tempt us to call μ the mean and σ the standard deviation, but in Nelson-style probability theory, this is nonsense. A normal distribution, as we have defined the concept, need not have moments anywhere near those of a conventional normal distribution.

Theorem 10.7. *If Z is an L^2 standard normal random variable and $X = \mu + \sigma Z$, where μ and σ are limited and $\sigma \geq 0$, then X is L^2, $E(X) \simeq \mu$, and* $\operatorname{var}(X) \simeq \sigma^2$.

Proof. Every standard normal random variable that arises in the central limit theorem (Corollary 8.1) is L^2. Moreover, such a random variable (8.1) has mean zero and standard deviation one by standardization. Since the map $x \mapsto x^{(a)}$ defined by (6.6) is limited and continuous for limited a, it follows by the

approximation lemma (Lemma 6.7) that nearly equivalent L^2 random variables have nearly equal mean and variance. Hence $E(Z) \simeq 0$ and $\operatorname{var}(Z) \simeq 1$.

By (6.4a) and (6.4b), X is L^2. It is elementary that $E(X) = \mu + \sigma E(Z)$ and $\operatorname{var}(X) = \sigma^2 \operatorname{var}(Z)$. The assertion about $E(X)$ and $\operatorname{var}(X)$ then follows from Theorems 3.7 and 3.10. □

Hence if we were to take L^2 as part of the definition of "normal" then μ and σ would be nearly equal to the mean and standard deviation. We have decided to not take L^2 as part of the definition, because it is easier to add it when wanted (say "L^2 and normal") than to remove it when not wanted (say "nearly equivalent to a normal random variable").

Theorem 10.8. *If Z is a standard normal random variable and $X = \mu + \sigma Z$, where μ and σ are limited and $\sigma \geq 0$, then the median of the distribution of X is nearly equal to μ and the $\Phi(1)$ quantile of X is nearly equal to $\mu + \sigma$, where Φ is defined by* (10.8).

Proof. Writing $\phi = \Phi'$, we have for $a < b$

$$\Phi(b) - \Phi(a) \geq (b - a)\phi\left(\max(|a|, |b|)\right)$$

by the law of the mean and the unimodality of ϕ. From $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ and Theorems 3.10 and 3.11 we conclude

$$\Phi(a) \ll \Phi(b), \qquad \text{whenever } -\infty \ll a \ll b \ll \infty. \tag{10.9}$$

Since Z is limited almost surely, its p-th quantiles for $0 \ll p \ll 1$ are limited and (10.9) implies these quantiles are unique up to near equality. In particular, the 0.5 and $\Phi(1) \approx 0.8413$ quantiles are unique up to near equality, and nearly equal to zero and one, respectively. Then Theorem 3.10 implies the 0.5 and $\Phi(1)$ quantiles of the distribution of X are nearly equal to μ and $\mu + \sigma$, respectively. □

Chapter 11

Characteristic Functions

11.1 Definitions

As always, $\mathbb{R}$ denotes the real number system. Now we introduce $\mathbb{C}$ for the complex number system.

The *characteristic function* of a random variable X is a function $\varphi : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi(t) = E\{e^{itX}\}, \qquad t \in \mathbb{R}.$$

Complex variables play only a limited, algebraic role in the theory. By the Euler formula

$$e^{iy} = \cos(y) + i\sin(y)$$

characteristic functions are determined by two real-valued functions, which are $t \mapsto E\{\cos(tX)\}$, the real part of φ, and $t \mapsto E\{\sin(tX)\}$, the imaginary part of φ. The only virtue in using complex numbers is that certain identities, such as

$$e^{i(u+v)} = e^{iu}e^{iv}$$

are "obvious" phrased in terms of complex exponentials and not "obvious" when phrased in terms of trigonometric identities.

For x and y in $\mathbb{R}^d$ we denote the standard inner product by $\langle \cdot, \cdot \rangle$, that is,

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i,$$

where $x = (x_1, \ldots, x_d)$ and similarly with x replaced by y.

The *characteristic function* of a random vector X taking values in $\mathbb{R}^d$ is a complex-valued function φ defined on all of $\mathbb{R}^d$ by

$$\varphi(t) = E\{e^{i\langle t, X\rangle}\}, \qquad t \in \mathbb{R}^d.$$

If d is limited, we say $\mathbb{R}^d$ is *limited-dimensional*. When $\mathbb{R}^d$ is limited-dimensional, an element of $\mathbb{R}^d$ is

- *infinitesimal* if and only if all its components are infinitesimal
- and *limited* if and only if all its components are limited.

As usual, *unlimited* means not limited, and *appreciable* means limited and non-infinitesimal.

Clearly, a limited-dimensional random vector is infinitesimal, appreciable, or unlimited if and only if its L^∞ norm is, where this norm is defined by

$$\|x\|_\infty = \max\{\, |x_i| : i = 1, \ldots, d \,\}.$$

As usual, we write $x \simeq y$ to mean $x - y \simeq 0$. If we define the L^p norms by

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$$

for $1 \leq p < \infty$, then the obvious inequality

$$\|x\|_\infty \leq \|x\|_p \leq d \cdot \|x\|_\infty$$

implies that a limited-dimensional random vector is infinitesimal, appreciable, or unlimited if and only if its L^p norm is likewise.

When d is unlimited, the L^p norms no longer agree about which vectors are infinitesimal, appreciable, and unlimited. Hence our original definition based on the behavior of components no longer makes sense either. Thus we shall see that the characteristic function theory we develop here is useful only for limited-dimensional random vectors.

11.2 Convergence I

A famous and very important theorem of Kolmogorov-style probability theory says that a sequence of random variables converges in distribution if and only if the characteristic functions converge pointwise. In this section we start to develop the Nelson-style analog, working on the easy direction of the "if and only if" (the other direction is dealt with in Section 11.6).

Theorem 11.1. *If two limited-dimensional random vectors are nearly equivalent, then their characteristic functions are nearly equal at all limited argument values.*

Proof. For limited t, the function $x \mapsto e^{i\langle t,x\rangle}$ is limited and (nearly) continuous because of

$$e^{i\langle t,x\rangle} - e^{i\langle t,y\rangle} = e^{i\langle t,x\rangle}\big(1 - e^{i\langle t,y-x\rangle}\big),$$

which implies

$$\begin{aligned}\left|e^{i\langle t,x\rangle} - e^{i\langle t,y\rangle}\right| &\le \left|1 - e^{i\langle t,y-x\rangle}\right| \\ &= \sqrt{\left|1-\cos\bigl(\langle t,y-x\rangle\bigr)\right|^2 + \left|\sin\bigl(\langle t,y-x\rangle\bigr)\right|^2},\end{aligned}$$

because $\langle t, y - x\rangle \simeq 0$ whenever $x \simeq y$ by the Cauchy-Schwarz inequality and because sine and cosine are nearly continuous at zero (and everywhere else) by Lemma 4.1. □

We cannot ask for near equality at all t. Consider the random variable concentrated at zero, which has characteristic function identically equal to one, and another random variable concentrated at a nonzero infinitesimal ϵ, which has characteristic function $t \mapsto e^{it\epsilon}$. These two random variables are nearly equivalent, but $e^{it\epsilon}$ is not nearly equal to one for all t.

11.3 The Discrete Fourier Transform

Let

$$T = \{\, k\epsilon : k \in \mathbb{Z},\ |k| \le n \,\} \tag{11.1}$$

where $\epsilon > 0$, $\mathbb{Z}$ is the set of all integers, and n is a positive integer. We call such a set T a *symmetric grid*; we call ϵ the *spacing* of T and write $\epsilon = \text{spac}(T)$; we call $N = 2n + 1$ the *cardinality* of T and write $N = \text{card}(T)$.

If T is defined as in (11.1) then we define

$$T^* = \{\, k\epsilon^* : k \in \mathbb{Z},\ |k| \le n \,\} \tag{11.2}$$

where

$$\epsilon^* = \frac{2\pi}{N\epsilon} \tag{11.3}$$

and call T^* the symmetric grid *conjugate* to T. Of course $\text{spac}(T^*) = \epsilon^*$ and $\text{card}(T^*) = N$.

If $f : T \to \mathbb{C}$ is any function, then

$$f^*(t) = \epsilon \sum_{x \in T} f(x) e^{itx} \tag{11.4}$$

defines a function $T^* \to \mathbb{C}$ that we call the *discrete Fourier transform* of f. The terminology "discrete Fourier transform" is also used for the mapping $\mathbb{C}^T \to \mathbb{C}^{T^*}$ given by $f \mapsto f^*$.

Theorem 11.2. *The discrete Fourier transform $f \mapsto f^*$ is invertible with inverse $f^* \mapsto f$ given by*

$$f(x) = \frac{\epsilon^*}{2\pi} \sum_{t \in T^*} f^*(t) e^{-itx}. \tag{11.5}$$

Proof. Plugging (11.4) into the right hand side of (11.5) we obtain

$$\begin{aligned}\frac{\epsilon \cdot \epsilon^*}{2\pi} \sum_{t\in T^*} \sum_{y\in T} f(y) e^{ity} e^{-itx} &= \frac{1}{N} \sum_{t\in T^*} \sum_{y\in T} f(y) e^{it(y-x)} \\ &= \sum_{y\in T} f(y) \frac{1}{N} \sum_{t\in T^*} e^{it(y-x)}\end{aligned}$$

We now claim that

$$\frac{1}{N} \sum_{t\in T^*} e^{it(y-x)} \tag{11.6}$$

is zero if $y \neq x$ and one if $y = x$, which implies (11.5). To establish this claim, suppose $y - x = k\epsilon$, so k is an integer and $|k| \leq 2n$. Then

$$\begin{aligned}\frac{1}{N} \sum_{t\in T^*} e^{it(y-x)} &= \frac{1}{N} \sum_{m=-n}^{n} e^{i\epsilon^* m\epsilon k} \\ &= \frac{1}{N} \sum_{m=-n}^{n} e^{2\pi imk/N} \\ &= e^{-2\pi ink/N} \frac{1}{N} \sum_{m=0}^{2n} e^{2\pi imk/N}\end{aligned}$$

Now we have two cases. If $k = 0$, which happens if and only if $x = y$, then the exponentials are all equal to one and this reduces to $(1/N)\sum_{m=0}^{2n} 1 = 1$. Otherwise, $k \neq 0$, and, recall, $|k| \leq 2n < N$. Define $\omega = \exp(2\pi ik/N)$. Then $0 < |k| < N$ implies $\omega \neq 1$, and

$$\begin{aligned}\sum_{m=0}^{2n} e^{2\pi imk/N} &= \sum_{m=0}^{2n} \omega^m \\ &= \frac{1 - \omega^N}{1 - \omega}\end{aligned}$$

Now $\omega^N = 1$, so that finishes the proof of the claim that (11.6) is zero if $y \neq x$ and one otherwise. □

If $f : T^d \to \mathbb{C}$ is any function, then

$$f^*(t) = \epsilon^d \sum_{x\in T^d} f(x) e^{i\langle t,x\rangle} \tag{11.7}$$

defines a function $(T^*)^d \to \mathbb{C}$ that we call the *multivariate discrete Fourier transform* of f. This terminology is also used for the mapping $f \mapsto f^*$.

Theorem 11.3. *The multivariate discrete Fourier transform $f \mapsto f^*$ is invertible with inverse $f^* \mapsto f$ given by*

$$f(x) = \left(\frac{\epsilon^*}{2\pi}\right)^d \sum_{t\in (T^*)^d} f^*(t) e^{-i\langle t,x\rangle}. \tag{11.8}$$

Proof. Plugging (11.7) into the right hand side of (11.8) we obtain

$$\begin{aligned}\left(\frac{\epsilon\cdot\epsilon^*}{2\pi}\right)^d \sum_{t\in(T^*)^d}\sum_{y\in T^d} f(y)e^{i\langle t,y\rangle}e^{-i\langle t,x\rangle} &= \frac{1}{N^d}\sum_{t\in(T^*)^d}\sum_{y\in T^d} f(y)e^{i\langle t,y-x\rangle}\\ &= \sum_{y\in T^d} f(y)\frac{1}{N^d}\sum_{t\in(T^*)^d} e^{i\langle t,y-x\rangle}\\ &= \sum_{y\in T^d} f(y)\prod_{k=1}^{d}\frac{1}{N}\sum_{t_k\in T^*} e^{it_k(y_k-x_k)}\end{aligned}$$

where $t=(t_1,\ldots,t_k)$ and similarly for x and y. Now we know from the proof of Theorem 11.2 that the innermost average is zero unless $y_k=x_k$, in which case it is one. Thus the product is zero unless $y=x$, in which case it is one. Thus the only nonzero term in the outermost sum is when $y=x$, in which case it is $f(x)$. □

Notice that if we have a random variable concentrated on T whose probability function is given by

$$\Pr(X=x)=\epsilon f(x),\qquad x\in T,$$

then the restriction of the characteristic function to T^* is the discrete Fourier transform f^*. Similarly, if we have a random vector concentrated on T^d whose probability function is given by

$$\Pr(X=x)=\epsilon^d f(x),\qquad x\in T^d,$$

then the restriction of the characteristic function to $(T^*)^d$ is the discrete Fourier transform f^*.

11.4 The Double Exponential Distribution

Let T be a symmetric grid that is also a near line, so $\mathrm{spac}(T)$ is infinitesimal and $\max(T)$ is unlimited. Let $\alpha>0$ be a real parameter. Then we call the distribution concentrated on T having unnormalized density

$$h_\alpha(x)=e^{-\alpha|x|},\qquad x\in T,$$

double exponential with rate parameter α. The normalizing constant for this distribution is calculated as follows

$$\begin{aligned} c(\alpha) &= \epsilon \sum_{x \in T} e^{-\alpha|x|} \\ &= \epsilon \sum_{k=-n}^{n} e^{-\alpha\epsilon|k|} \\ &= \epsilon \left(2 \sum_{k=0}^{n} e^{-\alpha\epsilon k} - 1 \right) \\ &= \epsilon \left(2 \cdot \frac{1 - e^{-\alpha\epsilon(n+1)}}{1 - e^{-\alpha\epsilon}} - 1 \right) \\ &= \epsilon \cdot \frac{1 - 2e^{-\alpha\epsilon(n+1)} + e^{-\alpha\epsilon}}{1 - e^{-\alpha\epsilon}} \\ &= \epsilon \cdot \frac{1 - 2A^{n+1} + A}{1 - A} \end{aligned} \tag{11.9}$$

where we have introduced $A = e^{-\alpha\epsilon}$, and the normalized density is given by

$$f_\alpha(x) = \frac{h_\alpha(x)}{c(\alpha)}, \qquad x \in T.$$

Lemma 11.4. *A double exponential distribution with unlimited rate parameter is infinitesimal almost surely. A double exponential distribution with non-infinitesimal rate parameter is limited almost surely.*

Proof. Suppose X has the double exponential distribution with rate parameter α. Then for $0 < m < n$.

$$\begin{aligned} \Pr(|X| \geq m\epsilon) &= \frac{2\epsilon}{c(\alpha)} \sum_{k=m}^{n} e^{-\alpha\epsilon k} \\ &= \frac{2\epsilon}{c(\alpha)} \cdot e^{-\alpha\epsilon m} \sum_{k=0}^{n-m} e^{-\alpha\epsilon k} \\ &= \frac{2\epsilon}{c(\alpha)} \cdot e^{-\alpha\epsilon m} \cdot \frac{1 - e^{-\alpha\epsilon(n-m+1)}}{1 - e^{-\alpha\epsilon}} \\ &= \frac{2\epsilon}{c(\alpha)} \cdot \frac{A^m(1 - A^{n-m+1})}{1 - A} \\ &= \frac{2A^m(1 - A^{n-m+1})}{1 - 2A^{n+1} + A} \end{aligned}$$

If $m\epsilon \gg 0$ and α is unlimited, then by the rules of exponentiation (Section 3.3) $A \leq 1$ and $A^m \simeq A^{n+1} \simeq 0$. Since $0 \leq A^{n-m+1} \leq 1$, we have

$$\Pr(|X| \geq m\epsilon) \simeq 0. \tag{11.10}$$

So the first statement follows from condition (ii) of Lemma 6.2.

If $m\epsilon \simeq \infty$ and $\alpha \gg 0$, then $A \leq 1$ and $A^m \simeq A^{n+1} \simeq 0$ again follow from the rules of exponentiation, and we obtain (11.10) again. So the second statement follows from condition (ii) of Lemma 6.3. □

The characteristic function of the double exponential distribution with rate parameter α is given by

$$\begin{aligned}
\varphi_\alpha(t) &= \frac{\epsilon}{c(\alpha)} \sum_{x \in T} e^{-\alpha|x| + itx} \\
&= \frac{\epsilon}{c(\alpha)} \sum_{k=-n}^{n} e^{-\alpha\epsilon|k| + it\epsilon k} \\
&= \frac{\epsilon}{c(\alpha)} \left(\sum_{k=0}^{n} e^{-(\alpha - it)\epsilon k} + \sum_{k=0}^{n} e^{-(\alpha + it)\epsilon k} - 1 \right) \\
&= \frac{\epsilon}{c(\alpha)} \left(\frac{1 - e^{-(\alpha - it)\epsilon(n+1)}}{1 - e^{-(\alpha - it)\epsilon}} + \frac{1 - e^{-(\alpha + it)\epsilon(n+1)}}{1 - e^{-(\alpha + it)\epsilon}} - 1 \right)
\end{aligned}$$

Since a double exponential distribution is symmetric about zero, the characteristic function is real, which is obvious from the fact that the two fractions in the large parentheses in the form above are complex conjugates of each other and hence their sum is real.

Using $e^{iw} = \cos(w) + i \sin(w)$ and clearing complex quantities from the denominator by multiplying numerator and denominator by their complex conjugates gives (after much formula manipulation, which has been checked using a computer algebra system)

$$\begin{aligned}
\varphi_\alpha(t) &= \frac{\epsilon}{c(\alpha)} \cdot \frac{1 - 2e^{-\alpha\epsilon(n+1)} \cos[(n+1)\epsilon t] + 2e^{-\alpha\epsilon(n+2)} \cos(n\epsilon t) - e^{-2\alpha\epsilon}}{1 - 2e^{-\alpha\epsilon} \cos(\epsilon t) + e^{-2\alpha\epsilon}} \\
&= \frac{\epsilon}{c(\alpha)} \cdot \frac{1 - 2A^{n+1} \cos[(n+1)\epsilon t] + 2A^{n+2} \cos(n\epsilon t) - A^2}{1 - 2A \cos(\epsilon t) + A^2} \\
&= \frac{1 + \dfrac{2A^{n+1}[1 - \cos((n+1)\epsilon t) - A + A\cos(n\epsilon t)]}{(1 - A)(1 - 2A^{n+1} + A)}}{1 + \dfrac{2A[1 - \cos(\epsilon t)]}{(1 - A)^2}}
\end{aligned} \tag{11.11}$$

We need the following bounds for sine and cosine

$$x - \frac{x^3}{3!} \leq \sin(x) \leq x, \qquad x \geq 0 \tag{11.12a}$$

$$1 - \frac{x^2}{2} \leq \cos(x) \leq 1 - \frac{x^2}{2} + \frac{x^4}{4!}, \qquad x \geq 0 \tag{11.12b}$$

These are easily proved in order of degree of the bound. The function $f_1(x) = x - \sin(x)$ has derivative $1 - \cos(x)$, which is nonnegative, hence f_1 is nondecreasing and greater than or equal to $f_1(0) = 0$ for $x \geq 0$. The function

$f_2(x) = \cos(x) - 1 + x^2/2$ has derivative $f_1(x)$, which is nonnegative, hence f_2 is nondecreasing and greater than or equal to $f_2(0) = 0$ for $x \geq 0$. The function $f_3(x) = \sin(x) - x + x^3/3!$ has derivative $f_2(x)$, which is nonnegative, hence f_3 is nondecreasing and greater than or equal to $f_3(0) = 0$ for $x \geq 0$. The function $f_4(x) = 1 - x^2/2 - x^4/4! - \cos(x)$ has derivative $f_3(x)$, which is nonnegative, hence f_4 is nondecreasing and greater than or equal to $f_4(0) = 0$ for $x \geq 0$. Since both sides of each inequality in (11.12b) are symmetric functions of x, we actually have

$$1 - \frac{x^2}{2} \leq \cos(x) \leq 1 - \frac{x^2}{2} + \frac{x^4}{4!}, \qquad x \in \mathbb{R}. \tag{11.13}$$

hence

$$1 - \cos(x) \sim \frac{x^2}{2}, \qquad x \simeq 0. \tag{11.14}$$

For non-infinitesimal x we use the bounds, also from (11.13),

$$1 - \frac{x^2}{2} \leq \cos(x) \leq 1 - \frac{x^2}{2}\left(1 - \frac{\pi^4}{12}\right), \qquad |x| \leq \pi. \tag{11.15}$$

Lemma 11.5. *Let φ_α denote the discrete Fourier transform of the double exponential density with appreciable rate parameter α defined on a symmetric grid T that is also a near line and satisfies*

$$\operatorname{spac}(T) \cdot \max(T) \simeq \infty. \tag{11.16}$$

Then

$$\varphi_\alpha(t) \simeq \frac{1}{1 + \frac{t^2}{\alpha^2}}, \qquad |t| \ll \infty \tag{11.17}$$

and

$$0 < \varphi_\alpha(t) \leq \frac{2}{1 + \frac{t^2}{6\alpha^2}}, \qquad t \in T^*. \tag{11.18}$$

The near equality (11.17) is mildly interesting, in that we derive it without reference to the right hand side being the Kolmogorov-style characteristic function of the Kolmogorov-style (continuous) double exponential distribution with rate parameter α. Of course, it could also be derived from Corollary 4.8 and the Kolmogorov-style result.

The bounds (11.18) will be crucial in our proof of the characteristic function convergence theorem in Section 11.6.

The condition (11.16) is inelegant but harmless, because we only use this lemma as a technical tool and can always assure that the condition is satisfied. For example, if we want $\max(T) = M$, then we can choose $n = \lceil M^{3/2} \rceil$ and $\epsilon = \operatorname{spac}(T) = M/n \sim M^{-1/2}$.

Proof. Recall that $\varphi_\alpha(t)$ is given by (11.11), and that $A = e^{-\alpha\epsilon}$. It follows from the rules of exponentiation (Section 3.3) that $A \simeq 1$ but $A^{n+1} \simeq 0$.

We first show that the numerator of (11.11) is nearly equal to one for $t \in T^*$. From $|\cos(x)| \leq 1$ and $A \simeq 1$ we have $|1 - \cos((n+1)\epsilon t) - A + A\cos(n\epsilon t)| \lesssim 2$, and from $A^{n+1} \simeq 0$ we have $1 + A + 2A^{n+1} \simeq 2$.

Since $1 - \exp(-\epsilon\alpha) \sim \alpha\epsilon$ by Lemma 4.9 and the following comments, and since $\exp(x) \geq 1 + x$, for $x \geq 0$ by the Maclaurin series for the exponential function, we have for any appreciable positive η

$$\frac{A^n}{1-A} = \frac{\exp(-n\epsilon\alpha)}{1-\exp(-\epsilon\alpha)} \leq \frac{1+\eta}{\alpha\epsilon(1+n\alpha\epsilon)} \leq n^{-1}\alpha^{-2}\epsilon^{-2}(1+\eta).$$

Hence (α being appreciable) condition (11.16) implies $A^n/(1-A) \simeq 0$. Putting all this together we have the numerator of (11.11) nearly equal to one.

Using $A \simeq 1$ and $1 - A \sim \alpha\epsilon$, which were derived above, and (11.14) we obtain

$$\frac{2A[1-\cos(\epsilon t)]}{(1-A)^2} \sim \frac{t^2}{\alpha^2}$$

and combining this with our result about the numerator of (11.11) we get (11.17), using Lemma 4.4 several times.

Since we only consider $t \in T^*$, we have $|t| \leq n\epsilon^*$, hence $|\epsilon t| \leq n\epsilon\epsilon^* = 2\pi n/N < \pi$. Thus we can apply (11.15) to the term $1 - \cos(\epsilon t)$ in the denominator of (11.11), obtaining

$$\frac{\epsilon^2 t^2}{2}\left(1 - \frac{\pi^2}{12}\right) \leq 1 - \cos(\epsilon t)$$

We know from our earlier analysis that $A/(1-A)^2 \sim 1/(\alpha\epsilon)^2$. Hence for any positive $\delta \ll 1$ we have

$$\frac{t^2}{\alpha^2} \cdot \delta\left(1 - \frac{\pi^2}{12}\right) \leq \frac{2A[1-\cos(\epsilon t)]}{(1-A)^2},$$

and, since $1/6 \ll 1 - \pi^2/12$, we can choose δ so this becomes

$$\frac{t^2}{6\alpha^2} \leq \frac{2A[1-\cos(\epsilon t)]}{(1-A)^2}.$$

Combining this with our result about the numerator of (11.11) gives one inequality in (11.18). The positivity assertion is obvious. □

11.5 Convolution

For x and y in T we define $x \oplus y$ to be the sum "modulo T", that is, the unique element of T having the form $x + y + kN\epsilon$ where $N = \text{card}(T)$ and k is an integer. If $n\epsilon = \max(T)$ so $N = 2n + 1$ (notation we have been using all along), then, for example, $n\epsilon \oplus \epsilon = -n\epsilon$ and $n\epsilon + n\epsilon = -\epsilon$.

For x and y in T and t in T^* we have

$$\begin{aligned}\exp[i(x \oplus y)t] &= \exp[i(x + y + kN\epsilon)t] \\ &= \exp(ixt)\exp(iyt)\exp(ikN\epsilon t) \\ &= \exp(ixt)\exp(iyt)\exp(ikmN\epsilon\epsilon^*) \\ &= \exp(ixt)\exp(iyt)\exp(2\pi ikm) \\ &= \exp(ixt)\exp(iyt)\end{aligned}$$

for some integers k and m.

For any random variable Z concentrated on T let φ_Z denote the restriction of its characteristic function to T^*. This conflicts with our earlier notation, which is still in force. If X is a random variable, then φ_X is its characteristic function. If α is a real number, then φ_α is the characteristic function of the double exponential distribution with rate parameter α.

If X and Y are independent random variables concentrated on T, then

$$\begin{aligned}\varphi_{X\oplus Y}(t) &= E\{e^{it(X\oplus Y)}\} \\ &= E\{e^{itX}e^{itY}\} \\ &= \varphi_X(t)\varphi_Y(t)\end{aligned}$$

Of course, except for $\oplus$ rather than $+$, this fact is familiar from classical probability theory. So long as we are interested only in random variables that are limited almost surely and so long as T is a near line, there is no practical difference between $X + Y$ and $X \oplus Y$. The fact proved in this section will serve just as well as its classical counterpart.

For vectors we define $\oplus$ to act coordinatewise

$$(x_1, \ldots, x_d) \oplus (y_1, \ldots, y_d) = (x_1 \oplus y_1, \ldots, x_d \oplus y_d)$$

Then $\varphi_{X\oplus Y}(t) = \varphi_X(t)\varphi_Y(t)$ also holds where X and Y are random elements of T^d and all three characteristic functions are restricted to $(T^*)^d$.

11.6 Convergence II

11.6.1 One-Dimensional

We continue to let T be a symmetric grid that is also a near line and satisfies (11.16). For any random variable X concentrated on T, we denote its density by f_X, so

$$\Pr\{X = x\} = \epsilon f_X(x).$$

We also continue to let φ_X denote the characteristic function of X.

Lemma 11.6. *Let X and Y be random variables and Z_α a double exponential random variable with appreciable rate parameter α independent of X and Y, all three random variables concentrated on T. Suppose*

$$\varphi_X(t) \simeq \varphi_Y(t), \qquad t \in T^*,\ |t| \ll \infty.$$

Then

$$f_{X\oplus Z_\alpha}(x) \simeq f_{Y\oplus Z_\alpha}(x), \qquad x \in T. \tag{11.19}$$

Proof. By (11.5)

$$\begin{aligned} f_{X\oplus Z_\alpha}(x) &= \frac{\epsilon^*}{2\pi} \sum_{t\in T^*} \varphi_{X\oplus Z_\alpha}(t) e^{-itx} \\ &= \frac{\epsilon^*}{2\pi} \sum_{t\in T^*} \varphi_X(t)\varphi_\alpha(t) e^{-itx} \end{aligned} \tag{11.20}$$

and similarly with X replaced by Y. Since $\varphi_X(t)e^{-itx}$ is bounded in modulus by 1 and similarly with X replaced by Y and by the bound (11.18) on $\varphi_\alpha(t)$ the sum (11.20) satisfies the conditions for Corollary 4.7 and similarly with X replaced by Y. Thus that corollary implies (11.19). □

Lemma 11.7 (Scheffé)**.** *Let X and Y be almost surely limited random variables concentrated on the symmetric grid T that is also a near line, and suppose*

$$f_X(x) \simeq f_Y(x), \qquad x \in T. \tag{11.21}$$

Then

$$\epsilon \sum_{x\in T} |f_X(x) - f_Y(x)| \simeq 0 \tag{11.22}$$

and for any limited function h on T

$$E\{h(X)\} = \epsilon \sum_{x\in T} h(x) f_X(x) \simeq \epsilon \sum_{x\in T} h(x) f_Y(x) = E\{h(Y)\}. \tag{11.23}$$

As usual when we take an eponym from classical probability theory, this is not what is usually called Scheffé's lemma (Lehmann, 1959, p. 351) but the lemma that plays the same role in radically elementary probability theory.

Note that the conclusion (11.21) is stronger than near equivalence of X and Y since it holds for all limited h, not just limited nearly continuous h. In Kolmogorov-style probability theory this type of convergence (what Scheffé's lemma implies) is called convergence in total variation. We shall not use it enough to need a name for it.

Proof. By (iii) of Lemma 6.3 for any $\eta \gg 0$ we can choose a limited a such that

$$\Pr\{|X| > a\} = \epsilon \sum_{\substack{x\in T \\ |x|>a}} f_X(x) \le \frac{\eta}{3}$$

and similarly with X replaced by Y. Hence

$$\epsilon \sum_{\substack{x\in T \\ |x|>a}} |f_X(x) - f_Y(x)| \le \frac{2\eta}{3},$$

and by Theorem 4.5 and (11.21)

$$\epsilon \sum_{\substack{x\in T\\ |x|\le a}} |f_X(x) - f_Y(x)| \simeq 0.$$

Hence by the triangle inequality

$$\epsilon \sum_{x\in T} |f_X(x) - f_Y(x)| \lesssim \frac{2\eta}{3}.$$

Since $\eta \gg 0$ was arbitrary, this gives (11.22). And (11.22) immediately implies (11.23). □

Theorem 11.8. *Let X and Y be almost surely limited random variables, and suppose*

$$\varphi_X(t) \simeq \varphi_Y(t), \qquad \text{for all limited } t. \tag{11.24}$$

Then X and Y are nearly equivalent.

Proof. Choose a symmetric grid T that is also a near line and satisfies (11.16) such that $\max(T)$ is larger than the maximum of the supports of $|X|$ and $|Y|$.

Define $g : \mathbb{R} \to T$ by $g(x) = \max\{\, t \in T : t \le x \,\}$. Then $g(X) \simeq X$ always, hence almost surely, hence $g(X)$ is nearly equivalent to X by part (iii) of Theorem 7.3, and similarly with X replaced by Y. By Theorem 11.1 we have $\varphi_X(t) \simeq \varphi_{g(X)}$ for limited t, and similarly with X replaced by Y, hence we have $\varphi_{g(X)} \simeq \varphi_{g(Y)}$ for limited t, because $\simeq$ is an external equivalence relation (Corollary 3.9).

Let Z_α have the double exponential distribution on T with rate parameter α and be independent of X and Y.

First suppose α is appreciable. It is clear that $g(X)$ is limited almost surely. By Lemma 11.4 so is Z_α. Since $x \oplus y = x + y$ whenever x and y are limited, and sum of limited is limited, it is clear that $g(X) \oplus Z_\alpha$ is also limited almost surely. Then by Lemma 11.6 and by Lemma 11.7 and the comment following it, $g(X) \oplus Z_\alpha$ is nearly equivalent to $g(Y) \oplus Z_\alpha$.

Second suppose α is unlimited. Then by Lemma 11.4 $Z_\alpha \simeq 0$ almost surely. Hence by Slutsky's theorem (Theorem 9.10) the random vector $\big(g(X), Z_\alpha\big)$ is nearly equivalent to the random vector $\big(g(X), 0\big)$, and similarly with X replaced by Y. The map $(x, y) \mapsto x \oplus y$ is continuous at points (x, y) such that x and y are limited. Hence by the continuous mapping theorem (Theorem 9.8), $g(X) \oplus Z_\alpha$ is nearly equivalent to $g(X) \oplus 0 = g(X)$, and similarly with X replaced by Y.

Fix a limited nearly continuous function $h : T \to \mathbb{R}$, and fix $\eta \gg 0$. Then

$$\big|E\big\{h\big(g(X) \oplus Z_\alpha\big)\big\} - E\big\{h\big(g(X)\big)\big\}\big| \le \frac{\eta}{3}$$

for all unlimited α and hence by overspill for some limited α, and similarly with X replaced by Y, and we may use the same limited α for both. Since

$$E\big\{h\big(g(X) \oplus Z_\alpha\big)\big\} \simeq E\big\{h\big(g(Y) \oplus Z_\alpha\big)\big\},$$

we have

$$\left|E\{h(g(X))\} - E\{h(g(Y))\}\right| \lesssim \frac{2\eta}{3}$$

by the triangle inequality. Since $\eta \gg 0$ was arbitrary,

$$E\{h(g(X))\} \simeq E\{h(g(Y))\}.$$

Since h was an arbitrary limited nearly continuous function, $g(X)$ is nearly equivalent to $g(Y)$. Since near equivalence is an external equivalence relation (Lemma 7.2), $X \overset{\text{w}}{\simeq} g(X) \overset{\text{w}}{\simeq} g(Y) \overset{\text{w}}{\simeq} Y$ implies $X \overset{\text{w}}{\simeq} Y$. □

11.6.2 Limited-Dimensional

Theorem 11.9. *Let X and Y be almost surely limited random vectors, taking values in $\mathbb{R}^d$ for limited d, and suppose* (11.24) *holds. Then X and Y are nearly equivalent.*

Proof. The proof is very similar to that of Theorem 11.8. We just need the multivariate analogs of each tool used in that proof. We already have the inversion theorem for the discrete Fourier transform (Theorem 11.3).

The analog of the symmetric grid T in the proof of Theorem 11.8 is T^d where T is a symmetric grid that is also a near line and satisfies (11.16) such that $\max(T)$ is larger than all values of $\|X\|_\infty$ and $\|Y\|_\infty$. As usual, $\epsilon = \text{spac}(T)$ and $\epsilon^* = \text{spac}(T^*)$.

The analog of the function g in the proof of Theorem 11.8 is a map from $\mathbb{R}^d$ to T^d that operates coordinatewise, the action on each coordinate being the g in the proof of Theorem 11.8. We again call this map g. Then again we have $g(X) \simeq X$ almost surely, hence $g(X)$ and X are nearly equivalent, and similarly with X replaced by Y.

For the analog of Z_α in the proof of Theorem 11.8 we use here the random vector, also denoted Z_α having independent and identically distributed components all of which have the double exponential distribution with rate parameter α. Lemma 6.1 and Lemma 11.4 imply Z_α is almost surely zero when $\alpha \simeq \infty$ and limited almost surely when $\alpha \gg 0$.

The multivariate analog of Lemma 11.6 is proved much the same way, the only additional wrinkle being external induction on the dimension. Now by Theorem 11.3 we have

$$\begin{aligned} f_{g(X)\oplus Z_\alpha}(x) &= \left(\frac{\epsilon^*}{2\pi}\right)^d \sum_{t \in (T^*)^d} \varphi_{g(X)}(t) \prod_{j=1}^d \varphi_\alpha(t_j) e^{-it_j x_j} \\ &= \frac{\epsilon^*}{2\pi} \sum_{t_1 \in T^*} \varphi_\alpha(t_1) e^{-it_1 x_1} \frac{\epsilon^*}{2\pi} \sum_{t_2 \in T^*} \varphi_\alpha(t_2) e^{-it_2 x_2} \\ &\qquad \cdots \frac{\epsilon^*}{2\pi} \sum_{t_d \in T^*} \varphi_\alpha(t_d) e^{-it_d x_d} \varphi_{g(X)}(t) \end{aligned}$$

where x and t denote vectors with components x_i and t_i and where the sum in the first line over $(T^*)^d$ is replaced by d iterated sums over T^* in the next two lines, and similarly with X replaced by Y. Hence

$$f_{g(X)\oplus Z_\alpha}(x) - f_{g(Y)\oplus Z_\alpha}(x) = \frac{\epsilon^*}{2\pi} \sum_{t_1 \in T^*} \varphi_\alpha(t_1) e^{-it_1x_1} \frac{\epsilon^*}{2\pi} \sum_{t_2 \in T^*} \varphi_\alpha(t_2) e^{-it_2x_2} \cdots \frac{\epsilon^*}{2\pi} \sum_{t_d \in T^*} \varphi_\alpha(t_d) e^{-it_dx_d} |\varphi_{g(X)}(t) - \varphi_{g(Y)}(t)|$$

Now by assumption $\varphi_X(t) \simeq \varphi_Y(t)$ for limited t, and by Theorem 11.1 we have $\varphi_X(t) \simeq \varphi_{g(X)}$ for limited t and similarly with X replaced by Y. Thus the term in the absolute value above is infinitesimal for limited t. Corollary 4.7 now shows that the bottom line above is infinitesimal when α is appreciable. Then external induction shows that $f_{g(X)\oplus Z_\alpha}(x) - f_{g(Y)\oplus Z_\alpha}(x)$ is infinitesimal (for all $x \in T$).

The multivariate analog of Lemma 11.7 is proved much the same way, the only additional wrinkle being external induction on the dimension. Since X is limited almost surely, so is $g(X)$ by Lemma 9.7 and similarly with X replaced by Y. Fix an appreciable α. Then $g(X) \oplus Z_\alpha$ is also limited almost surely.

For any $\eta \gg 0$ there is a limited a such that the box B^d, where

$$B = \{\, x \in T : |x| \le a \,\}$$

satisfies $\Pr\{g(X) \oplus Z_\alpha \in B^d\} \ge 1 - \eta$, and similarly with X replaced by Y. Hence

$$\epsilon \sum_{x \in T^d} |f_{g(X)\oplus Z_\alpha}(x) - f_{g(Y)\oplus Z_\alpha}(x)| \le 2\eta + \epsilon \sum_{x \in B^d} |f_{g(X)\oplus Z_\alpha}(x) - f_{g(Y)\oplus Z_\alpha}(x)| = 2\eta + \epsilon \sum_{x_1 \in B} \epsilon \sum_{x_2 \in B} \cdots \epsilon \sum_{x_d \in B} |f_{g(X)\oplus Z_\alpha}(x) - f_{g(Y)\oplus Z_\alpha}(x)|$$

where $x = (x_1, \ldots, x_d)$. Theorem 4.5 implies the innermost sum is infinitesimal. By external induction and Theorem 4.5 all sums on the right hand side are infinitesimal. Since $\eta \gg 0$ was arbitrary, the left hand side is infinitesimal. As in the one-dimensional case, it is immediate that $g(X) \oplus Z_\alpha$ and $g(Y) \oplus Z_\alpha$ are nearly equivalent (when α is appreciable).

The rest of the proof follows that of Theorem 11.8 almost without change. The Slutsky argument is the same, and the rest is unchanged except now h maps T^d to $\mathbb{R}$. □

Corollary 11.10 (Cramér-Wold)**.** *Let X and Y be random vectors taking values in $\mathbb{R}^d$ where d is limited. Then X and Y are nearly equivalent if and only if $\langle t, X\rangle$ and $\langle t, Y\rangle$ are nearly equivalent for every limited $t \in \mathbb{R}^d$.*

Proof. For scalar s and vector t

$$\varphi_{\langle t,X\rangle}(s) = E\{e^{is\langle t,X\rangle}\} = \varphi_X(st).$$

Suppose X and Y are nearly equivalent and t is limited. Then st is limited, for all limited s, hence

$$\varphi_{\langle t,X\rangle}(s) \simeq \varphi_{\langle t,Y\rangle}(s), \qquad s \text{ limited.}$$

Conversely, suppose $\langle t, X\rangle$ and $\langle t, Y\rangle$ are limited for all limited t. Then

$$\varphi_X(t) \simeq \varphi_Y(t), \qquad t \text{ limited.}$$

□

Corollary 11.11. *Suppose X_i and Y_i are independent, limited-dimensional random vectors for $i = 1, 2$ and $X_1 \overset{\text{w}}{\simeq} X_2$ and and $Y_1 \overset{\text{w}}{\simeq} Y_2$. Then*

$$(X_1, Y_1) \overset{\text{w}}{\simeq} (X_2, Y_2).$$

The simple proof using characteristic functions and Theorems 11.1 and 11.9 is left as an exercise.

Part III

Statistics

Chapter 12

De Finetti's Theorem

12.1 Exchangeability

Fix a positive integer ν, and let $I = \{1, \ldots, \nu\}$. Fix an arbitrary set S, and let X_1, ..., X_ν be random elements of S. Let Π be the set of all permutations (bijective functions) $\pi : I \to I$. We say the sequence X_1, ..., X_ν is *exchangeable* if

$$\Pr(X_i = s_i,\ i \in I) = \Pr(X_i = s_{\pi(i)},\ i \in I) \tag{12.1}$$

for all $\pi \in \Pi$ and any choice of s_1, ..., s_ν in S.

It simplifies discussion if we consider the sequence as single object: a random tuple $\mathbf{X} = (X_1, \ldots, X_\nu)$, which we take to be a random element of the space S^I. We will also consider S^I, as in set theory, to be the set of all functions $I \to S$. Then, if $\mathbf{s} \in S^I$ and $\pi \in \Pi$, it makes sense to write $\mathbf{s} \circ \pi$, a composition of functions $I \to I \to S$ and hence also an element of S^I.

The *exchangeable algebra on* S^I is the family $\mathcal{F}$ of functions $f : S^I \to \mathbb{R}$ satisfying

$$f(\mathbf{s}) = f(\mathbf{s} \circ \pi), \qquad \mathbf{s} \in S^I,\ \pi \in \Pi.$$

The *atoms* of $\mathcal{F}$ are the equivalence classes

$$[\mathbf{s}] = \{\, \mathbf{s} \circ \pi : \pi \in \Pi \,\}.$$

A random element $\mathbf{X}$ of S^I is *exchangeable* if

$$\Pr(\mathbf{X} = \mathbf{s}) = \Pr(\mathbf{X} = \mathbf{s} \circ \pi), \qquad \mathbf{s} \in S,\ \pi \in \Pi, \tag{12.2}$$

which says the same thing as (12.1) in different notation, or, what is equivalent, if

$$\Pr(\mathbf{X} = \mathbf{s}_1) = \Pr(\mathbf{X} = \mathbf{s}_2), \qquad \mathbf{s}_1 \in S,\ \mathbf{s}_2 \in S,\ [\mathbf{s}_1] = [\mathbf{s}_2]. \tag{12.3}$$

The *exchangeable algebra on the sample space induced by* $\mathbf{X}$ is

$$\mathcal{E} = \{\, f \circ \mathbf{X} : f \in \mathcal{F} \,\}.$$

The atoms of $\mathcal{E}$ are the nonempty events of the form $\mathbf{X}^{-1}([\mathbf{s}])$, the nonempty preimages of atoms of $\mathcal{F}$.

12.2 A Simple De Finetti Theorem

We write X_i for the i-th component of $\mathbf{X}$ defined by $X_i(\omega) = \mathbf{X}(\omega)(i)$ and indicate this relationship by $\mathbf{X} = (X_1, \ldots, X_\nu)$, and this gets us back to the simple notation used in (12.1).

Theorem 12.1 (De Finetti)**.** *Let $\mathbf{X} = (X_1, \ldots, X_\nu)$ be exchangeable with ν unlimited. Let $\mathcal{E}$ be the exchangeable algebra on the sample space induced by $\mathbf{X}$. Then for any limited functions h_1, …, h_n from S to $\mathbb{R}$ with n limited*

$$E\Big\{\prod_{i=1}^{n} h_i(X_i) \;\Big|\; \mathcal{E}\Big\} \simeq \prod_{i=1}^{n} E\big\{h_i(X_i) \mid \mathcal{E}\big\} \tag{12.4a}$$

and

$$E\Big\{\prod_{i=1}^{n} h_i(X_i)\Big\} \simeq E\Big[\prod_{i=1}^{n} E\big\{h_i(X_i) \mid \mathcal{E}\big\}\Big]. \tag{12.4b}$$

The condition that the h_i are *limited* functions means that every value is limited. It also means they have a simultaneous limited bound because, the sample space being finite, the maximum of $\big|h_i\big(X_j(\omega)\big)\big|$ over all i, j, and ω is achieved, hence limited.

In (12.4a) the two sides of the equation are random variables, call them Y_L and Y_R. What the equation means is $Y_L(\omega) \simeq Y_R(\omega)$ for all $\omega \in \Omega$.

Lemma 12.2. *Suppose $\mathbf{X}$ is an exchangeable random element of S^I and P is a uniformly distributed random element of Π independent of $\mathbf{X}$. Then $\mathbf{X}$ and $\mathbf{X} \circ P$ are equal in law.*

Elements of S^I are maps $I \to S$ and elements of Π are maps $I \to I$. Hence $\mathbf{X} \circ P$ is a map $I \to I \to S$, which is an element of S^I. More precisely, $\mathbf{X}(\omega) \circ P(\omega)$ is, for each ω, a map $I \to I \to S$. The equal in law assertion is that

$$\Pr(\mathbf{X} = \mathbf{s}) = \Pr(\mathbf{X} \circ P = \mathbf{s}), \qquad \mathbf{s} \in S^I. \tag{12.5}$$

Proof. Fix $\mathbf{s} \in S^I$ and let K be the number of elements in $[\mathbf{s}]$. Then it is clear from (12.3) that

$$\Pr(\mathbf{X} = \mathbf{s}) = \frac{1}{K} \Pr(\mathbf{X} \in [\mathbf{s}])$$

From the bijectivity of permutations and (12.2) and the observation above

$$\begin{aligned} \Pr(\mathbf{X} \circ P = \mathbf{s}) &= \Pr(\mathbf{X} = \mathbf{s} \circ P^{-1}) \\ &= \Pr(\mathbf{X} = \mathbf{s} \circ P^{-1} \circ \pi) \\ &= \frac{1}{K} \Pr(\mathbf{X} \in [\mathbf{s} \circ P^{-1} \circ \pi]) \end{aligned}$$

but the last term is $\Pr(\mathbf{X} \in [\mathbf{s}])$ by the definition of $[\,\cdot\,]$. □

Lemma 12.3. *With the setup of the theorem, let $\mathcal{E}_n$ denote the algebra generated by X_1, …, X_n and $\mathcal{E}$. Then*

$$E\big\{h_n(X_n) \bigm| \mathcal{E}_{n-1}\big\} \simeq E\big\{h_n(X_n) \bigm| \mathcal{E}\big\} \tag{12.6}$$

holds for each limited integer $n \geq 2$.

Proof. The conditional distributions involved in (12.6) are very simple. Fix an ω in Ω. Write $\mathbf{X}(\omega) = \mathbf{s} = (s_1, \ldots, s_\nu)$. The atom of $\mathcal{E}$ containing ω is

$$A_\omega = \{\, \omega' \in \Omega : \mathbf{X}(\omega') \in [\mathbf{s}] \,\}. \tag{12.7a}$$

On this atom we have

$$E\big\{h_n(X_n) \bigm| \mathcal{E}\big\} = \frac{1}{\nu} \sum_{i=1}^{\nu} h_n(s_i) \tag{12.7b}$$

because by Lemma 12.2 $h_n(X_n)$ has the same distribution on this atom as $h_n(s_{P(i)})$ where P is a uniformly distributed random permutation and $P(i)$ puts equal probability on the points 1, …, ν.

The atom of $\mathcal{E}_{n-1}$ containing ω is

$$B_\omega = \{\, \omega' \in A_\omega : X_i(\omega') = s_i,\ i = 1, \ldots, n-1 \,\}. \tag{12.8a}$$

On this atom,

$$E\big\{h_n(X_n) \bigm| \mathcal{E}_{n-1}\big\} = \frac{1}{\nu - n + 1} \sum_{i=n}^{\nu} h_n(s_i) \tag{12.8b}$$

by an argument similar to that establishing (12.7b). We must show that the right hand sides of (12.7b) and (12.8b) are nearly equal.

Each of the h_i is limited, hence by the comment following the statement of the theorem we may assume there is a simultaneous limited bound a for all of them, and

$$\left| \sum_{i=1}^{n-1} h_n(s_i) \right| \leq (n-1)a,$$

which implies

$$\frac{1}{\nu} \sum_{i=1}^{\nu} h_n(s_i) \simeq \frac{1}{\nu} \sum_{i=n}^{\nu} h_n(s_i) \tag{12.9a}$$

because $(n-1)a/\nu$ is infinitesimal.

Also,

$$\begin{aligned} \left| \frac{1}{\nu - n + 1} \sum_{i=n}^{\nu} h_n(s_i) - \frac{1}{\nu} \sum_{i=n}^{\nu} h_n(s_i) \right| &= \frac{n-1}{\nu(\nu - n + 1)} \left| \sum_{i=n}^{\nu} h_n(s_i) \right| \\ &\leq \frac{(n-1)a}{\nu} \end{aligned}$$

is infinitesimal, which implies

$$\frac{1}{\nu}\sum_{i=n}^{\nu} h_n(s_i) \simeq \frac{1}{\nu - n + 1}\sum_{i=n}^{\nu} h_n(s_i). \tag{12.9b}$$

Putting (12.9a) and (12.9b) together finishes the proof. □

Lemma 12.4. *For any random variables X and Y and any algebra $\mathcal{A}$, $X \simeq Y$ implies $E_{\mathcal{A}}X \simeq E_{\mathcal{A}}Y$ and $EX \simeq EY$.*

Proof. For any $\epsilon \gg 0$ we have $X - \epsilon \le Y \le X + \epsilon$, which implies $E_{\mathcal{A}}X - \epsilon \le E_{\mathcal{A}}Y \le E_{\mathcal{A}}X + \epsilon$ and this can hold for every $\epsilon \gg 0$ only if $E_{\mathcal{A}}X - E_{\mathcal{A}}Y$ is infinitesimal. The proof for unconditional expectation is similar. □

Proof of the theorem.

$$\begin{aligned}
E\left\{\prod_{i=1}^{n} h_i(X_i) \;\middle|\; \mathcal{E}\right\} &= E\left[E\left\{\prod_{i=1}^{n} h_i(X_i) \;\middle|\; \mathcal{E}_{n-1}\right\} \;\middle|\; \mathcal{E}\right] \\
&= E\left[\prod_{i=1}^{n-1} h_i(X_i) \cdot E\{h_n(X_n) \mid \mathcal{E}_{n-1}\} \;\middle|\; \mathcal{E}\right] \\
&\simeq E\left[\prod_{i=1}^{n-1} h_i(X_i) \cdot E\{h_n(X_n) \mid \mathcal{E}\} \;\middle|\; \mathcal{E}\right] \\
&= E\{h_n(X_n) \mid \mathcal{E}\} \cdot E\left\{\prod_{i=1}^{n-1} h_i(X_i) \;\middle|\; \mathcal{E}\right\}
\end{aligned} \tag{12.10}$$

the first equality being $E_{\mathcal{E}}E_{\mathcal{E}_{n-1}} = E_{\mathcal{E}}$, the second and third equalities being $E_{\mathcal{A}}XY = YE_{\mathcal{A}}X$ when $Y \in \mathcal{A}$ for $\mathcal{A} = \mathcal{E}_{n-1}$ and $\mathcal{A} = \mathcal{E}$, respectively, and the near equality being Lemmas 12.3 and 12.4. Then (12.4a) follows from (12.10) by external induction, and (12.4b) follows by another application of Lemma 12.4. □

12.3 Philosophical Interpretation

The philosophical interpretation of the theorem is that the random variables X_1, X_2, ... X_ν are "nearly" conditionally independent given the exchangeable algebra by (12.4a) and (exactly) marginally identically distributed (conditionally and unconditionally) by exchangeability, and their joint conditional distribution is "nearly" determined by the marginal distributions by (12.4b).

Hence, if we adopt a Bayesian statistical model for data X_1, ..., X_n that assumes they are exactly conditionally independent and identically distributed given the exchangeable algebra, then we make only infinitesimal errors.

This setup is a bit more abstract than most discussions of Bayesian theory (but no more abstract than most discussions of de Finetti theorems). It is hard to see the exchangeable algebra playing the role of a Bayesian parameter. But at

least in Nelson-style theory there is no heavy probabilistic abstraction to wade through. We know that the exchangeable algebra $\mathcal{E}$ induces a partition $\mathrm{at}(\mathcal{E})$ on the sample space. If $\boldsymbol{\theta}$ is a random element of any space that induces the same partition, then we can write $E\{\,\cdot\,|\boldsymbol{\theta}\}$ in place of $E\{\,\cdot\,|\mathcal{E}\}$ and things will look more Bayesian just from the change of notation.

Thus the philosophical import of de Finetti's theorem is that independent and identically distributed (IID) assumptions arise in quite different ways for the Bayesian and the frequentist. The frequentist must assume that the data actually are IID, which is a very strong assumption. The Bayesian only "assumes" that the data are exchangeable, a much weaker assumption. Moreover, for a *subjective* Bayesian, exchangeability is not really an assumption but merely a lack of knowledge that would allow one to treat $\mathbf{X}$ differently from $\mathbf{X}\circ P$ (using the notation of Lemma 12.2).

12.4 A Fancier De Finetti Theorem

Theorem 12.5 (De Finetti II)**.** *Let $\mathbf{X} = (X_1, \ldots, X_\nu)$ be exchangeable with ν unlimited. Let $\mathcal{E}$ be the exchangeable algebra on the sample space induced by $\mathbf{X}$. Then for any functions h_1, …, h_n from S to $\mathbb{R}$ with n limited such that $h_1(X_1)$, …, $h_n(X_n)$ and $\prod_{i=1}^n h_i(X_i)$ and $\prod_{i=1}^n E\big\{h_i(X_i) \,\big|\, \mathcal{E}\big\}$ are L^1*

$$E\left\{\prod_{i=1}^n h_i(X_i) \,\middle|\, \mathcal{E}\right\} \simeq \prod_{i=1}^n E\big\{h_i(X_i) \,\big|\, \mathcal{E}\big\}, \qquad \textit{almost surely,} \tag{12.11a}$$

and

$$E\left\{\prod_{i=1}^n h_i(X_i)\right\} \simeq E\left[\prod_{i=1}^n E\big\{h_i(X_i) \,\big|\, \mathcal{E}\big\}\right]. \tag{12.11b}$$

In (12.11a) the two sides of the equation are random variables, call them Y_L and Y_R. What the equation means is for every $\epsilon \gg 0$ there exists an event N such that $\Pr(N) \le \epsilon$ and $Y_L(\omega) \simeq Y_R(\omega)$ except for ω in N.

Proof. Equation (12.11b) is implied by (12.4b) and Lemma 6.7. By Theorem 6.10 $h_1(X_1)$, …, $h_n(X_n)$ and $\prod_{i=1}^n h_i(X_i)$ are L^1 on almost every atom of $\mathcal{E}$. On such atoms equation (12.11a) is implied by (12.4a) and Lemma 6.7. □

Whether one prefers Theorem 12.1 or Theorem 12.5 is a matter of taste. The restriction to limited functions (the external analog of bounded functions) makes the statement of Theorem 12.1 much simpler, only involving nonstandard analysis through the basic idea of the existence of infinitesimals. Theorem 12.5 is stronger, but only in a rather trivial way involving Lemma 6.7, and involves the more complicated concepts *almost surely* and L^1 that use much more nonstandard analysis.

Theorem 12.1 only involves master's level calculations. Both the statement and the proof are elementary. The key issue is that (12.7b) is very close to (12.8b) whenever ν is much larger than n. Everything else is just applying

rules of probability theory that are no different in Nelson-style theory from Kolmogorov-style theory (master's level or PhD level).

Chapter 13

Almost Sure Convergence

We introduced almost sure convergence in Section 7.1. Now we apply it to a few things.

13.1 The Law of Large Numbers

Theorem 16.3 in Nelson (1987) is the law of large numbers.

Theorem 13.1 (The Law of Large Numbers)**.** *Suppose X_1, X_2, ..., X_ν are IID L^1 random variables with mean μ. Define*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{13.1}$$

Then $\overline{X}_n$ converges to μ almost surely.

We do not usually deal with almost sure convergence by directly applying the definition. What we usually use is the following theorem, which is Theorem 7.2 in Nelson (1987).

Theorem 13.2. *Let X_1, ..., X_ν be random variables, then X_n converges almost surely to zero if and only if*

$$\Pr\left\{ \sup_{\mu \le n \le \nu} |X_n| \ge \lambda \right\} \simeq 0, \qquad \mu \simeq \infty, \ \lambda \gg 0. \tag{13.2}$$

13.2 The Glivenko-Cantelli Theorem

Suppose X_1, ..., X_ν are IID random variables defined on a finite sample space having common (marginal) distribution function F, and $\widehat{F}_n$ is the empirical distribution function defined by

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i), \qquad x \in \mathbb{R}.$$

Then for each fixed x the random variable $\widehat{F}_n(x)$ is a "sample mean" of the form (13.1) with "population mean" $F(x) = E\{\widehat{F}_n(x)\}$. Also $\widehat{F}_n(x)$ is a limited, hence L^1, random variable. Thus the law of large numbers applies and gives almost surely

$$\left|\widehat{F}_n(x) - F(x)\right| \simeq 0, \qquad n \simeq \infty, \tag{13.3a}$$

which by Theorem 13.2 holds if and only if

$$\Pr\left\{ \sup_{\mu \le n \le \nu} \left|\widehat{F}_n(x) - F(x)\right| \ge \lambda \right\} \simeq 0, \qquad \mu \simeq \infty,\ \lambda \gg 0. \tag{13.3b}$$

These statements (13.3a) and (13.3b) are different statements for each fixed x. In (13.3a) the exception sets may depend on x.

The Glivenko-Cantelli theorem says these statements hold uniformly in x.

Theorem 13.3 (Glivenko-Cantelli)**.** *Suppose X_1, ..., X_ν are IID random variables defined on a finite sample space having common (marginal) distribution function F, and suppose $\widehat{F}_n$ is the empirical distribution function for sample size n. Then almost surely*

$$\sup_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \simeq 0, \qquad n \simeq \infty. \tag{13.4a}$$

By Theorem 13.2 the statement (13.4a) holds almost surely if and only if

$$\Pr\left\{ \sup_{\mu \le n \le \nu} \sup_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \ge \lambda \right\} \simeq 0, \qquad \mu \simeq \infty,\ \lambda \gg 0. \tag{13.4b}$$

As in the usual proof of the Kolmogorov-style Glivenko-Cantelli theorem, the proof starts with the statements (13.3a), one for each fixed x, which are implied by the law of large numbers. These statements do not obviously imply (13.4a) because the exception sets in (13.3a) depend on x. Some device must be used to reduce an uncountable infinity of exception sets to a limited number. Actually we work with the statements that do not explicitly involve "almost surely" and want to derive the one statement (13.4b) from the many statements (13.3b), but here too, the implication is not obvious. As in the usual Kolmogorov-style proof, the device we use is monotonicity of distribution functions and compactness of the unit interval.

Proof. Let T be the support of F. To simplify notation choose $\epsilon > 0$ smaller than any spacing in T so we always have

$$F(x - \epsilon) = \sup_{\substack{y \in T \\ y < x}} F(y)$$

whenever $x \in T$. Fix $\mu \simeq \infty$ and $\lambda \gg 0$, and choose a limited natural number k such that $1/k \le \lambda/4$. Then there are unique points $x_i \in T$ determined by

$$F(x_i - \epsilon) < \frac{i}{k} \le F(x_i), \qquad i = 1, \ldots, k$$

(meaning for each i there is a unique element of T that can be x_i and not meaning that the x_i are distinct—it may be that $x_i = x_{i'}$ when $i < i'$, in which case $x_i = x_j$ for $i \le j \le i'$). Also, since T is finite, there exists a (nonunique) point x_0 satisfying $F(x_0) = 0$.

Now for $i = 1, \ldots, k$ and $x_{i-1} < x < x_i$ we have

$$\begin{aligned}\widehat{F}_n(x) \ge F(x) + \lambda &\longrightarrow \widehat{F}_n(x_i - \epsilon) \ge F(x_{i-1}) + \lambda \\ &\longrightarrow \widehat{F}_n(x_i - \epsilon) \ge F(x_i - \epsilon) + \frac{\lambda}{2}\end{aligned}$$

the last implication holding because of

$$F(x_{i-1}) \ge \frac{i-1}{k} > F(x_i) - \frac{2}{k} > F(x_i - \epsilon) - \frac{2}{k} \ge F(x_i - \epsilon) - \frac{\lambda}{2},$$

the interpretation of these statements being "omega by omega," that is, $\widehat{F}_n(x)$ is a random variable defined by

$$\widehat{F}_n(x)(\omega) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}\big(X_i(\omega)\big)$$

and the implications mean that the set of ω such that $\widehat{F}_n(x)(\omega) \ge F(x) + \lambda$ holds is included in the set of ω such that $\widehat{F}_n(x_i - \epsilon)(\omega) \ge F(x_i - \epsilon) + \frac{\lambda}{2}$ holds. Similarly,

$$\begin{aligned}\widehat{F}_n(x) \le F(x) - \lambda &\longrightarrow \widehat{F}_n(x_{i-1}) \le F(x_i - \epsilon) - \lambda \\ &\longrightarrow \widehat{F}_n(x_{i-1}) \le F(x_{i-1}) - \frac{\lambda}{2}\end{aligned}$$

the last implication holding because of

$$F(x_i - \epsilon) < \frac{i}{k} \le F(x_{i-1}) + \frac{1}{k} \le F(x_{i-1}) + \frac{\lambda}{4}.$$

Putting these together and using monotonicity and subadditivity of probability gives

$$\begin{aligned}\Pr\left\{\sup_{\mu\le n\le\nu}\sup_{\substack{x\in\mathbb{R}\\ x_{i-1}<x<x_i}}\left|\widehat{F}_n(x) - F(x)\right| \ge \lambda\right\} &\\ \le \Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_i - \epsilon) - F(x_i - \epsilon)\right| \ge \frac{\lambda}{2}\right\} &\\ + \Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_{i-1}) - F(x_{i-1})\right| \ge \frac{\lambda}{2}\right\} &\end{aligned}$$

and since we also obviously have

$$\Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_i) - F(x_i)\right| \ge \lambda\right\} \le \Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_i) - F(x_i)\right| \ge \frac{\lambda}{2}\right\}$$

we get

$$\begin{aligned}\Pr\left\{\sup_{\mu\le n\le\nu}\sup_{\substack{x\in\mathbb{R}\\ x_0\le x<x_k}}\left|\widehat{F}_n(x)-F(x)\right|\ge\lambda\right\}&\\ \le\sum_{i=1}^{k}\Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_i-\epsilon)-F(x_i-\epsilon)\right|\ge\frac{\lambda}{2}\right\}&\\ +\sum_{i=1}^{k}\Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{F}_n(x_{i-1})-F(x_{i-1})\right|\ge\frac{\lambda}{2}\right\}&\quad(13.5)\end{aligned}$$

and we are done except for two minor points. First, since $F(x_0) = 0$ and $F(x_k) = 1$ it is never possible to have X_i outside the interval $(x_0, x_k]$ and we have $\widehat{F}_n(x_0) = 0$ and $\widehat{F}_n(x_k) = 1$ for all n (for all ω). Hence the inner supremum on the left hand side of (13.5) can run over all of $\mathbb{R}$ rather than just the interval $[x_0, x_k)$ without changing the event whose probability is being calculated. Second, by Corollary 3.8, if all the terms on the right hand side of (13.5) are infinitesimal, then their sum is infinitesimal. □

Caution: Lemma 10.1 and Theorem 10.4 say that $\widehat{F}_n$ converging to F implies convergence in distribution of (the distribution determined by) $\widehat{F}_n$ to (the distribution determined by) F, *if* (a very big "if") random variables having this limiting distribution (determined by F) are *almost surely limited!* The (Nelson-style analog of the) Glivenko-Cantelli theorem is true regardless of whether or not F has this property, but it doesn't say what you might think it says from the analogy with Kolmogorov-style theory.

13.3 Prohorov Consistency

We want a theorem analogous to the Glivenko-Cantelli theorem that applies when X_1, ..., X_ν are random elements of an arbitrary metric space having common (marginal) probability measure P, and $\widehat{P}_n$ is the empirical measure defined by

$$\widehat{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i), \qquad A \subset S.$$

Then (just like we saw for empirical distribution functions in the preceding section) for each fixed A the random variable $\widehat{P}_n(A)$ is a "sample mean" of the form (13.1) with "population mean" $P(A) = E\{\widehat{P}_n(A)\}$. Also $\widehat{P}_n(A)$ is a limited, hence L^1, random variable Thus the law of large numbers applies and gives almost surely

$$\left|\widehat{P}_n(A) - P(A)\right| \simeq 0, \qquad n \simeq \infty, \tag{13.6a}$$

which holds if and only if

$$\Pr\left\{\sup_{\mu\le n\le\nu}\left|\widehat{P}_n(A)-P(A)\right|\ge\lambda\right\}\simeq 0,\qquad \mu\simeq\infty,\ \lambda\gg 0. \tag{13.6b}$$

These statements (13.6a) and (13.6b) are different statements for each fixed A. In (13.6a) the exception sets may depend on A.

We investigate the conditions under which $\widehat{P}_n$ converges to P. In order to discuss convergence, we need a metric for convergence of probability measures. The natural one to use is the Prohorov metric, which was defined in Section 9.3.

We say $\widehat{P}_n$ *converges almost surely to* P (in the Prohorov metric) if almost surely

$$\pi\big(\widehat{P}_n,P\big)\simeq 0,\qquad n\simeq\infty, \tag{13.7a}$$

which holds if and only if

$$\Pr\left\{\sup_{\mu\le n\le\nu}\pi\big(\widehat{P}_n,P\big)\ge\lambda\right\}\simeq 0,\qquad \mu\simeq\infty,\ \lambda\gg 0. \tag{13.7b}$$

We say P is *nearly tight* if for every $\epsilon\gg 0$ there exists a set F of limited cardinality such that $P(F^\epsilon)\ge 1-\epsilon$, where F^ϵ, the ϵ-dilation of F, was defined in Section 9.3.

Note that a random variable X that is limited almost surely is nearly tight because for any $\epsilon\gg 0$, there exists a limited a such that $\Pr(|X|\ge a)\le\epsilon$. So, if we define

$$F=\{\,k\epsilon : k\in\mathbb{Z},\ |k\epsilon|\le a\,\},$$

then F has limited cardinality and F^ϵ covers $[-a,a]$. Hence $\Pr(F^\epsilon)\ge 1-\epsilon$.

Theorem 13.4. *Suppose X_1, ..., X_ν are IID random elements of a metric space having common (marginal) probability measure P and $\widehat{P}_n$ is the empirical measure for sample size n, then a necessary and sufficient condition for $\widehat{P}_n$ to converge almost surely to P (in the Prohorov metric) is that P is nearly tight.*

Proof. Let π denote the Prohorov metric. Then, by Lemma 9.2 and the comments on page 58 about the behavior of this metric,

$$\begin{aligned}\pi\big(\widehat{P}_n,P\big)\ge\lambda &\longleftrightarrow (\forall\epsilon>0)(\exists A\subset S)\big(\widehat{P}_n(A)>P(A^\lambda)+\lambda-\epsilon\big)\\ &\longleftrightarrow \sup_{A\subset S}\big(\widehat{P}_n(A)-P(A^\lambda)\big)\ge\lambda\end{aligned}$$

the arrows meaning that the statements are equivalent "omega by omega," that is if one holds at some point in the sample space, then so do the others. Hence (13.7b) is equivalent to

$$\Pr\left\{\sup_{\mu\le n\le\nu}\sup_{A\subset S}\big(\widehat{P}_n(A)-P(A^\lambda)\big)\ge\lambda\right\}\simeq 0,\qquad \mu\simeq\infty,\ \lambda\gg 0. \tag{13.8}$$

First, we prove necessity. Assume (13.8), and suppose to get a contradiction that P is not nearly tight. Hence there is an $\epsilon\gg 0$ such that we do not have

$P(F^\epsilon) \geq 1-\epsilon$ for any set F of limited cardinality. Let $A_n = \{ X_1, \ldots, X_n \}$, that is, A_n is the support of $\widehat{P}_n$, so $\widehat{P}_n(A_n) = 1$. Then a necessary condition that (13.8) hold is

$$\Pr\{1 - P(A_n^\epsilon) \geq \epsilon\} \simeq 0, \qquad n \simeq \infty. \tag{13.9}$$

This implies

$$\Pr\{1 - P(A_n^\epsilon) \geq \epsilon\} < \frac{1}{2} \tag{13.10}$$

holds for all unlimited n, hence by overspill for some limited n. But this is a contradiction because A_n has at most n points and by our choice of ϵ, this implies $P(A_n^\epsilon) < 1-\epsilon$ always (for all omega) and hence $\Pr\{1 - P(A_n^\epsilon) > \epsilon\} = 1$. That finishes the proof of the necessity of near tightness.

Now, we prove sufficiency. Assume near tightness. Fix $\lambda \gg 0$, and choose a set $F = \{x_1, \ldots, x_m\}$ of limited cardinality such that $P(F^{\lambda/3}) \geq 1 - \lambda/3$.

Disjointify the balls $\{x_i\}^{\lambda/3}$ defining recursively

$$B_i = \{x_i\}^{\lambda/3} \setminus \bigcup_{j=1}^{i-1} B_j.$$

Then the nonempty B_i partition $F^{\lambda/3}$.

Let $\mathcal{I}$ be the set of all subsets of $\{1, \ldots, m\}$, and define

$$A_I = \bigcup_{i \in I} B_i, \qquad I \in \mathcal{I}.$$

We do not need $A_\varnothing$ as a notation for the empty set, so let us reuse the notation, redefining this to be

$$A_\varnothing = S \setminus F^{\lambda/3}.$$

Note that the diameter of each B_i is less than λ, so if a set A contains even one point of B_i, then A^λ contains all of B_i. Hence

$$A \subset A_I \longrightarrow A_I \subset A^\lambda, \qquad I \in \mathcal{I} \setminus \varnothing.$$

By the law of large numbers (13.3b)

$$\Pr\left\{ \sup_{\mu \leq n \leq \nu} \left| \widehat{P}_n(A_I) - P(A_I) \right| \geq \frac{\lambda}{3} \right\} \simeq 0, \qquad \mu \simeq \infty,\ I \in \mathcal{I}.$$

Consider a point in the sample space (an omega) such that

$$\sup_{\mu \leq n \leq \nu} \sup_{I \in \mathcal{I}} \left| \widehat{P}_n(A_I) - P(A_I) \right| < \frac{\lambda}{3},$$

holds. Then for any $A \subset S$ define

$$I = \left\{ i \in \{1, \ldots, m\} : A \cap B_i \neq \varnothing \right\}$$

so when $\mu \leq n \leq \nu$ we have

$$\begin{aligned}\widehat{P}_n(A) - P(A^\lambda) &= \widehat{P}_n(A \setminus F^{\lambda/3}) + \widehat{P}_n(A \cap A_I) - P(A^\lambda) \\ &\leq \widehat{P}_n(A_\varnothing) + \widehat{P}_n(A_I) - P(A_I) \\ &< \frac{2\lambda}{3} + \widehat{P}_n(A_I) - P(A_I) \\ &< \lambda\end{aligned}$$

Hence for $\mu \simeq \infty$

$$\begin{aligned}&\Pr\left\{\sup_{\mu\leq n\leq\nu}\sup_{A\subset S}\left(\widehat{P}_n(A) - P(A^\lambda)\right) \geq \lambda\right\} \\ &\qquad\qquad \leq \sum_{I\in\mathcal{I}} \Pr\left\{\sup_{\mu\leq n\leq\nu}\left|\widehat{P}_n(A_I) - P(A_I)\right| \geq \frac{\lambda}{3}\right\}\end{aligned}$$

Since $\mathcal{I}$ has limited cardinality (less than 2^m, which is limited by Theorem 2.4), this sum is infinitesimal by Corollary 3.8, and, since $\lambda \gg 0$ was arbitrary, we have (13.8) and are done. □

13.4 Discussion

13.4.1 Kolmogorov-Style Pathology

The concept we call "near tightness" that plays such an important role in Theorem 13.4 is the Nelson-style analog of what van der Vaart and Wellner (1996, p. 17) call *pre-tightness*. Their Lemma 1.3.2 says that a Borel probability measure on a metric space is *pre-tight* if and only if it is *separable*. A Borel probability measure P on a metric space is *separable* when it has a separable support, that is, there exists a Borel measurable set A such that $P(A) = 1$ and A has a countable dense subset.

So far this is fairly simple, but, according to the discussion on p. 24 of van der Vaart and Wellner (1996), it is undecidable under the usual axioms of set theory whether nonseparable Borel probability measures can exist. It is known that one could add to mathematics the axiom that they do not exist without causing inconsistency (that is, if mathematics is inconsistent with this new axiom, then it is already inconsistent before this axiom is added). But it is "apparently unknown" (say van der Vaart and Wellner, 1996) whether nonseparable probability measures can consistently exist (whatever that means).

It is incredible what a morass we sink into with this notion. Countable additivity is supposed to make things simple. Here it seems to make things about as complicated as they can possibly be. Kolmogorov-style probability theory allows an incredible amount of pathology and requires extreme technical virtuosity to navigate around it.

In Nelson-style theory these things are simple. Non-tight (not nearly tight) probability measures exist, and there are very simple examples of them, such as

the discrete uniform distribution on the integers 1, ..., ν with ν unlimited. It is clear in Theorem 13.4 what (near) tightness does and why it is needed.

13.4.2 Near Tightness and Topology

On a slightly different point, why did we choose to make our concept *near tightness* the analog of Kolmogorov-style pre-tightness rather than tightness. The reason is that Kolmogorov-style tightness involves compact sets, and it is not possible to define compact sets in "radically elementary" nonstandard analysis. It is possible to define compact sets in the full theory of nonstandard analysis (Nelson, unpublished, p. 13 for the definition of compact subsets of $\mathbb{R}$ and p. 15 for the definition of compact subsets of general topological spaces), but we don't want to use the full theory. Thus we avoid topology and the notions compact, closed, and open. The same issue is why closed and open sets do not appear in our version of the portmanteau theorem (Theorem 9.6) and are replaced by ϵ-dilations and ϵ-erosions.

13.4.3 Prohorov Consistency and Glivenko-Cantelli

Although we drew an analogy between Theorems 13.3 and 13.4, and they are quite analogous in their conditions and their proofs, the Kolmogorov-style analog of Theorem 13.4 is not what is called a uniform Glivenko-Cantelli theorem (see van der Vaart and Wellner, 1996, Chapter 2.4).

13.4.4 Statistical Inference

Whatever the Prohorov-consistency theorem is called, it does show in what sense statistical estimation is possible with no assumptions about the true unknown distribution except near tightness. We know from Theorem 13.4 that, in the language of statistics, $\widehat{P}_n$ is a consistent estimator of the true unknown distribution P.

This remains true for the "minimum Prohorov distance estimator." If we have a statistical model $\mathcal{P}$, which we take to be a finite family of probability distributions and we let $\widetilde{P}_n$ be any estimator — that is, a function of the data X_1, ..., X_n taking values in $\mathcal{P}$ — that satisfies

$$\pi(\widetilde{P}_n, \widehat{P}_n) \lesssim \min_{P \in \mathcal{P}} \pi(P, \widehat{P}_n),$$

then $\widetilde{P}_n$ is a consistent estimator of the true unknown distribution P by the triangle inequality.

Admittedly, "minimum Prohorov distance estimators" are very hard to calculate and not used in practical applications. We will have to do a lot more to get a useful Nelson-style theory of statistics. But it's a start.

Bibliography

Billingsley, P. (1999). *Convergence of Probability Measures*, second edition. Wiley.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society, Series B*, 35:189–233.

Eaton, M. L. (1992). A statistical diptych: Admissible inferences–recurrence of symmetric Markov Chains. *Annals of Statistics*, 20:1147–1179.

Feller, W. (1950). *An Introduction to Probability Theory and its Applications*, Vol. I. Wiley.

Goldblatt, R. (1998). *Lectures on the Hyperreals: An Introduction to Nonstandard Analysis*. Springer-Verlag.

Gordon, E. I., Kusraev, A. G., and Kutateladze, S. S. (2002). *Infinitesimal Analysis*. Kluwer Academic Publishers.

Halmos, P. R. (1958). *Finite-Dimensional Vector Spaces*, second edition. Van Nostrand (original publisher), Springer-Verlag (current publisher).

Kanovei, V. and Reeken, M. (2004). *Nonstandard Analysis, Axiomatically*. Springer-Verlag.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung* Springer. English translation (1950): *Foundations of the theory of probability*. Chelsea.

Lang, S. (1993). *Real and Functional Analysis*, third edition. Springer-Verlag.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley.

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. Springer.

Loeb, P. (1975). Conversion from nonstandard to standard measure spaces and applications to probability theory. *Transactions of the American Mathematical Society* 211: 113–122.

Moore, D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics*. W. H. Freeman.

Nelson, E. (1977). Internal set theory: a new approach to nonstandard analysis. *Bulletin of the American Mathematical Society*, 83:1165–1198.

Nelson, E. (1987). *Radically Elementary Probability Theory*. Princeton University Press.

Nelson, E. (unpublished). *Internal Set Theory*. Chapter 1 of an unfinished book on nonstandard analysis. Available on the web at the URL `http://www.math.princeton.edu/~nelson/books/1.pdf`.

Pearson, E. S. and Please, N. W. (1975). Relation between the shape of the population distribution and the robustness of four simple test statistics. *Biometrika*, 62:223–224.

Robert, A. M. (1988). *Nonstandard Analysis*. Wiley. Translated by the author from the 1985 French original. Republished by Dover, 2003.

Robinson, A. (1996). *Non-standard Analysis*, rev. ed. Princeton University Press. Originally published by North-Holland, 1974. First edition, 1966.

Solovay, R. M. (1970). A model of set-theory in which every set of reals is Lebesgue measurable. *Annals of Mathematics, Second Series*, 92:1–56.

Sudderth, W. D. (1980). Finitely additive priors, coherence and the marginalization paradox. *Journal of the Royal Statistical Society, Series B*, 42:339–341.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics* Springer-Verlag.

Whittle, P. (2005). *Probability via Expectation*, 4th ed. Springer.