Appendix A

A REVIEW OF BASIC PROBABILITY THEORY

In statistics we study techniques for obtaining and using information in the presence of uncertainty. A prerequisite for such a study is a mathematical model for randomness and some knowledge of its properties. The Kolmogorov model and the modern theory of probability based on it are what we need. The reader is expected to have had a basic course in probability theory. The purpose of this appendix is to indicate what results we consider basic and to introduce some of the notation that will be used in the rest of the book. Because the notation and the level of generality differ somewhat from that found in the standard textbooks in probability at this level, we include some commentary. Sections A.14 and A.15 contain some results that the student may not know, which are relevant to our study of statistics. Therefore, we include some proofs as well in these sections.

In Appendix B we will give additional probability theory results that are of special interest in statistics and may not be treated in enough detail in some probability texts.

A.1 THE BASIC MODEL

Classical mechanics is built around the principle that like causes produce like effects. Probability theory provides a model for situations in which like or similar causes can produce one of a number of unlike effects. A coin that is tossed can land heads or tails. A group of ten individuals selected from the population of the United States can have a majority for or against legalized abortion. The intensity of solar flares in the same month of two different years can vary sharply.

The situations we are going to model can all be thought of as *random experiments*. Viewed naively, an experiment is an action that consists of observing or preparing a set of circumstances and then observing the outcome of this situation. We add to this notion the requirement that to be called an experiment such an action must be repeatable, at least conceptually. The adjective *random* is used only to indicate that we do not, in addition, *require* that every repetition yield the same outcome, although we do not exclude this case. What we expect and observe in practice when we repeat a random experiment many times is that the relative frequency of each of the possible outcomes will tend to stabilize. This long-term relative frequency n_A/n, where n_A is the number of times the possible outcome A occurs in n repetitions, is to many statisticians, including the authors, the operational interpretation of the mathematical concept of probability. In this sense, almost any kind of activity involving uncertainty, from horse races to genetic experiments, falls under the vague heading of "random experiment."

Another school of statisticians finds this formulation too restrictive. By interpreting probability as a subjective measure, they are willing to assign probabilities in any situation involving uncertainty, whether it is conceptually repeatable or not. For a discussion of this approach and further references the reader may wish to consult Savage (1954), Raiffa and Schlaiffer (1961), Savage (1962), Lindley (1965), de Groot (1970), and Berger (1985). We now turn to the mathematical abstraction of a random experiment, the probability model.

In this section and throughout the book, we presume the reader to be familiar with elementary set theory and its notation at the level of Chapter 1 of Feller (1968) or Chapter 1 of Parzen (1960). We shall use the symbols $\cup$, $\cap$, ^{c}, $-$, $\subset$ for union, intersection, complementation, set theoretic difference, and inclusion as is usual in elementary set theory.

A random experiment is described mathematically in terms of the following quantities.

A.1.1 The *sample space* is the set of all possible outcomes of a random experiment. We denote it by Ω. Its complement, the *null set* or *impossible event*, is denoted by $\emptyset$.

A.1.2 A *sample point* is any member of Ω and is typically denoted by ω.

A.1.3 Subsets of Ω are called *events*. We denote events by A, B, and so on or by a description of their members, as we shall see subsequently. The relation between the experiment and the model is given by the correspondence "A occurs if and only if the actual outcome of the experiment is a member of A." The set operations we have mentioned have interpretations also. For example, the relation $A \subset B$ between sets considered as events means that the occurrence of A implies the occurrence of B. If $\omega \in \Omega$, $\{\omega\}$ is called an *elementary* event. If A contains more than one point, it is called a *composite* event.

A.1.4 We will let $\mathcal{A}$ denote a class of subsets of Ω to which we can assign probabilities. For technical mathematical reasons it may not be possible to assign a probability P to every subset of Ω. However, $\mathcal{A}$ is always taken to be a sigma field, which by definition is a nonempty class of events closed under countable unions, intersections, and complementation (cf. Chung, 1974; Grimmett and Stirzaker, 1992; and Loeve, 1977). A *probability distribution* or *measure* is a nonnegative function P on $\mathcal{A}$ having the following properties:

(i) $P(\Omega) = 1$.

(ii) If $A_1, A_2, \ldots$ are pairwise disjoint sets in $\mathcal{A}$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Recall that $\cup_{i=1}^{\infty} A_i$ is just the collection of points that are in any one of the sets A_i and that two sets are disjoint if they have no points in common.

A.1.5 The three objects Ω, $\mathcal{A}$, and P together describe a random experiment mathematically. We shall refer to the triple $(\Omega, \mathcal{A}, P)$ either as a *probability model* or identify the model with what it represents as a (random) *experiment*. For convenience, when we refer to *events* we shall automatically exclude those that are not members of $\mathcal{A}$.

References

Gnedenko (1967) Chapter 1, Sections 1–3, 6–8
Grimmett and Stirzaker (1992) Sections 1.1–1.3
Hoel, Port, and Stone (1971) Sections 1.1, 1.2
Parzen (1960) Chapter 1, Sections 1–5
Pitman (1993) Sections 1.2 and 1.3

A.2 ELEMENTARY PROPERTIES OF PROBABILITY MODELS

The following are consequences of the definition of P.

A.2.1 If $A \subset B$, then $P(B - A) = P(B) - P(A)$.

A.2.2 $P(A^c) = 1 - P(A)$, $P(\emptyset) = 0$.

A.2.3 If $A \subset B$, $P(B) \geq P(A)$.

A.2.4 $0 \leq P(A) \leq 1$.

A.2.5 $P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n)$.

A.2.6 If $A_1 \subset A_2 \subset \cdots \subset A_n \ldots$, then $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$.

A.2.7 $P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(A_i^c)$ (Bonferroni's inequality).

References

Gnedenko, (1967) Chapter 1, Section 8
Grimmett and Stirzaker (1992) Section 1.3
Hoel, Port, and Stone (1992) Section 1.3
Parzen (1960) Chapter 1, Sections 4–5
Pitman (1993) Section 1.3

A.3 DISCRETE PROBABILITY MODELS

A.3.1 A probability model is called *discrete* if Ω is finite or countably infinite and every subset of Ω is assigned a probability. That is, we can write $\Omega = \{\omega_1, \omega_2, \ldots\}$ and $\mathcal{A}$ is the collection of subsets of Ω. In this case, by axiom (ii) of (A.1.4), we have for any event A,

$$P(A) = \Sigma_{\omega_i \in A} P(\{\omega_i\}). \qquad \text{(A.3.2)}$$

An important special case arises when Ω has a finite number of elements, say N, all of which are equally likely. Then $P(\{\omega\}) = 1/N$ for every $\omega \in \Omega$, and

$$P(A) = \frac{\text{Number of elements in } A}{N}. \tag{A.3.3}$$

A.3.4 Suppose that $\omega_1, \dots, \omega_N$ are the members of some population (humans, guinea pigs, flowers, machines, etc.). Then selecting an individual from this population in such a way that no one member is more likely to be drawn than another, *selecting at random*, is an experiment leading to the model of (A.3.3). Such selection can be carried out if N is small by putting the "names" of the ω_i in a hopper, shaking well, and drawing. For large N, a random number table or computer can be used.

References

Gnedenko (1967) Chapter 1, Sections 4–5
Parzen (1960) Chapter 1, Sections 6–7
Pitman (1993) Section 1.1

A.4 CONDITIONAL PROBABILITY AND INDEPENDENCE

Given an event B such that $P(B) > 0$ and any other event A, we define the *conditional probability* of A *given* B, which we write $P(A \mid B)$, by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{A.4.1}$$

If $P(A)$ corresponds to the frequency with which A occurs in a large number of repetitions of the experiment, then $P(A \mid B)$ corresponds to the frequency of occurrence of A relative to the class of trials in which B does occur. From a heuristic point of view $P(A \mid B)$ is the chance we would assign to the event A if we were told that B has occurred.

If $A_1, A_2, \dots$ are (pairwise) disjoint events and $P(B) > 0$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \mid B\right) = \sum_{i=1}^{\infty} P(A_i \mid B). \tag{A.4.2}$$

In fact, for fixed B as before, the function $P(\cdot \mid B)$ is a probability measure on $(\Omega, \mathcal{A})$ which is referred to as the *conditional probability measure* given B.

Transposition of the denominator in (A.4.1) gives the *multiplication rule*,

$$P(A \cap B) = P(B)P(A \mid B). \tag{A.4.3}$$

If $B_1, B_2, \dots, B_n$ are (pairwise) disjoint events of positive probability whose union is Ω, the identity $A = \bigcup_{j=1}^{n}(A \cap B_j)$, (A.1.4)(ii) and (A.4.3) yield

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j)P(B_j). \tag{A.4.4}$$

If $P(A)$ is positive, we can combine (A.4.1), (A.4.3), and (A.4.4) and obtain *Bayes rule*

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}. \tag{A.4.5}$$

The *conditional probability* of A given $B_1, \dots, B_n$ is written $P(A \mid B_1, \dots, B_n)$ and defined by

$$P(A \mid B_1, \dots, B_n) = P(A \mid B_1 \cap \cdots \cap B_n) \tag{A.4.6}$$

for any events $A, B_1, \dots, B_n$ such that $P(B_1 \cap \cdots \cap B_n) > 0$.

Simple algebra leads to the *multiplication rule*,

$$P(B_1 \cap \cdots \cap B_n) = P(B_1)P(B_2 \mid B_1)P(B_3 \mid B_1, B_2) \dots P(B_n \mid B_1, \dots, B_{n-1}) \tag{A.4.7}$$

whenever $P(B_1 \cap \cdots \cap B_{n-1}) > 0$.

Two events A and B are said to be *independent* if

$$P(A \cap B) = P(A)P(B). \tag{A.4.8}$$

If $P(B) > 0$, the relation (A.4.8) may be written

$$P(A \mid B) = P(A). \tag{A.4.9}$$

In other words, A and B are independent if knowledge of B does not affect the probability of A.

The events $A_1, \dots, A_n$ are said to be *independent* if

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}) \tag{A.4.10}$$

for any subset $\{i_1, \dots, i_k\}$ of the integers $\{1, \dots, n\}$. If all the $P(A_i)$ are positive, relation (A.4.10) is equivalent to requiring that

$$P(A_j \mid A_{i_1}, \dots, A_{i_k}) = P(A_j) \tag{A.4.11}$$

for any j and $\{i_1, \dots, i_k\}$ such that $j \notin \{i_1, \dots, i_k\}$.

References

Gnedenko (1967) Chapter 1, Sections 9
Grimmett and Stirzaker (1992) Section 1.4
Hoel, Port, and Stone (1971) Sections 1.4, 1.5
Parzen (1960) Chapter 2, Section 4; Chapter 3, Sections 1,4
Pitman (1993) Section 1.4

A.5 COMPOUND EXPERIMENTS

There is an intuitive notion of independent experiments. For example, if we toss a coin twice, the outcome of the first experiment (toss) reasonably has nothing to do with the outcome of the second. On the other hand, it is easy to give examples of dependent experiments: If we draw twice at random from a hat containing two green chips and one red chip, and if we do not replace the first chip drawn before the second draw, then the probability of a given chip in the second draw will depend on the outcome of the first draw. To be able to talk about independence and dependence of experiments, we introduce the notion of a compound experiment.

Informally, a compound experiment is one made up of two or more component experiments. There are certain natural ways of defining sigma fields and probabilities for these experiments. These will be discussed in this section. The reader not interested in the formalities may skip to Section A.6 where examples of compound experiments are given.

A.5.1 Recall that if $A_1, \dots, A_n$ are events, the *Cartesian product* $A_1 \times \cdots \times A_n$ of $A_1, \dots, A_n$ is by definition $\{(\omega_1, \dots, \omega_n) : \omega_i \in A_i,\ 1 \le i \le n\}$. If we are given n experiments (probability models) $\mathcal{E}_1, \dots, \mathcal{E}_n$ with respective sample spaces $\Omega_1, \dots, \Omega_n$, then the *sample space* Ω *of the* n *stage compound experiment* is by definition $\Omega_1 \times \cdots \times \Omega_n$. The ($n$ stage) compound experiment consists in performing *component experiments* $\mathcal{E}_1, \dots, \mathcal{E}_n$ and recording all n outcomes. The interpretation of the sample space Ω is that $(\omega_1, \dots, \omega_n)$ is a sample point in Ω if and only if ω_1 is the outcome of $\mathcal{E}_1$, ω_2 is the outcome of $\mathcal{E}_2$ and so on. To say that $\mathcal{E}_i$ has had outcome $\omega_i^0 \in \Omega_i$ corresponds to the occurrence of the compound event (in Ω) given by $\Omega_1 \times \cdots \times \Omega_{i-1} \times \{\omega_i^0\} \times \Omega_{i+1} \times \cdots \times \Omega_n = \{(\omega_1, \dots, \omega_n) \in \Omega : \omega_i = \omega_i^0\}$. More generally, if $A_i \in \mathcal{A}_i$, the sigma field corresponding to $\mathcal{E}_i$, then A_i corresponds to $\Omega_1 \times \cdots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \cdots \times \Omega_n$ in the compound experiment. If we want to make the $\mathcal{E}_i$ independent, then intuitively we should have all classes of events $A_1, \dots, A_n$ with $A_i \in \mathcal{A}_i$, independent. This makes sense in the compound experiment. If P is the probability measure defined on the sigma field $\mathcal{A}$ of the compound experiment, that is, the subsets $\mathcal{A}$ of Ω to which we can assign probability$^{(1)}$, we should have

$$\begin{aligned}
&P([A_1 \times \Omega_2 \times \cdots \times \Omega_n] \cap [\Omega_1 \times A_2 \times \cdots \times \Omega_n] \cap \dots)\\
&= P(A_1 \times \cdots \times A_n)\\
&= P(A_1 \times \Omega_2 \times \cdots \times \Omega_n) P(\Omega_1 \times A_2 \times \cdots \times \Omega_n) \dots P(\Omega_1 \times \cdots \times \Omega_{n-1} \times A_n).
\end{aligned} \tag{A.5.2}$$

If we are given probabilities P_1 on $(\Omega_1, \mathcal{A}_1)$, P_2 on $(\Omega_2, \mathcal{A}_2), \dots, P_n$ on $(\Omega_n, \mathcal{A}_n)$, then (A.5.2) *defines* P for $A_1 \times \cdots \times A_n$ by

$$P(A_1 \times \cdots \times A_n) = P_1(A_1) \dots P_n(A_n). \tag{A.5.3}$$

It may be shown (Billingsley, 1995; Chung, 1974; Loeve, 1977) that if P is defined by (A.5.3) for events $A_1 \times \cdots \times A_n$, it can be uniquely extended to the sigma field $\mathcal{A}$ specified in note (1) at the end of this appendix. We shall speak of *independent experiments* $\mathcal{E}_1, \dots, \mathcal{E}_n$ if the n stage compound experiment has its probability structure specified by (A.5.3). In the discrete case (A.5.3) holds provided that

$$P(\{(\omega_1, \dots, \omega_n)\}) = P_1(\{\omega_1\}) \dots P_n(\{\omega_n\}) \text{ for all } \omega_i \in \Omega_i,\ 1 \le i \le n. \tag{A.5.4}$$

Specifying P when the $\mathcal{E}_i$ are dependent is more complicated. In the discrete case we know P once we have specified $P(\{(\omega_1 \dots, \omega_n)\})$ for each $(\omega_1, \dots, \omega_n)$ with $\omega_i \in \Omega_i$, $i = 1, \dots, n$. By the multiplication rule (A.4.7) we have, in the discrete case, the following.

A.5.5 $P(\{(\omega_1, \dots, \omega_n)\}) = P(\mathcal{E}_1$ has outcome $\omega_1)$ $P(\mathcal{E}_2$ has outcome $\omega_2 \mid \mathcal{E}_1$ has outcome $\omega_i) \dots P(\mathcal{E}_n$ has outcome $\omega_n \mid \mathcal{E}_1$ has outcome $\omega_1, \dots, \mathcal{E}_{n-1}$ has outcome $\omega_{n-1})$. The probability structure is determined by these conditional probabilities and conversely.

References

Grimmett and Stirzaker (1992) Sections 1.5, 1.6
Hoel, Port, and Stone (1971) Section 1.5
Parzen (1960) Chapter 3

A.6 BERNOULLI AND MULTINOMIAL TRIALS, SAMPLING WITH AND WITHOUT REPLACEMENT

A.6.1 Suppose that we have an experiment with only two possible outcomes, which we shall denote by S (success) and F (failure). If we assign $P(\{S\}) = p$, we shall refer to such an experiment as a *Bernoulli trial* with probability of success p. The simplest example of such a Bernoulli trial is tossing a coin with probability p of landing heads (success). Other examples will appear naturally in what follows. If we repeat such an experiment n times independently, we say we have performed n *Bernoulli trials with success probability* p. If Ω is the sample space of the compound experiment, any point $\omega \in \Omega$ is an n-dimensional vector of S's and F's and,

$$P(\{\omega\}) = p^{k(\omega)}(1-p)^{n-k(\omega)} \tag{A.6.2}$$

where $k(\omega)$ is the number of S's appearing in ω. If A_k is the event [exactly k S's occur], then

$$P(A_k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, \dots, n, \tag{A.6.3}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The formula (A.6.3) is known as the *binomial probability*.

A.6.4 More generally, if an experiment has q possible outcomes $\omega_1, \dots, \omega_q$ and $P(\{\omega_i\}) = p_i$, we refer to such an experiment as a *multinomial trial with probabilities* $p_1, \dots, p_q$. If the experiment is performed n times independently, the compound experiment is called n *multinomial trials with probabilities* $p_1, \dots, p_q$. If Ω is the sample space of this experiment and $\omega \in \Omega$, then

$$P(\{\omega\}) = p_1^{k_1(\omega)} \dots p_q^{k_q(\omega)} \tag{A.6.5}$$

where $k_i(\omega)$ = number of times ω_i appears in the sequence ω. If $A_{k_1,\dots,k_q}$ is the event (exactly $k_1\omega_1$'s are observed, exactly $k_2\omega_2$'s are observed, …, exactly $k_q\omega_q$'s are observed), then

$$P(A_{k_1,\dots,k_q}) = \frac{n!}{k_1!\dots k_q!}p_1^{k_1}\dots p_q^{k_q} \tag{A.6.6}$$

where the k_i are natural numbers adding up to n.

A.6.7 If we perform an experiment given by $(\Omega, \mathcal{A}, P)$ independently n times, we shall sometimes refer to the outcome of the compound experiment as a *sample of size* n *from the population given by* $(\Omega, \mathcal{A}, P)$. When Ω is finite the term, *with replacement* is added to distinguish this situation from that described in (A.6.8) as follows.

A.6.8 If we have a finite population of cases $\Omega = \{\omega_1 \dots, \omega_N\}$ and we select cases ω_i successively at random n times *without replacement*, the component experiments are not independent and, for any outcome $a = (\omega_{i_1}, \dots, \omega_{i_n})$ of the compound experiment,

$$P(\{a\}) = \frac{1}{(N)_n} \tag{A.6.9}$$

where

$$(N)_n = \frac{N!}{(N-n)!}.$$

If the case drawn is replaced before the next drawing, we are sampling *with replacement*, and the component experiments are independent and $P(\{a\}) = 1/N^n$. If Np of the members of Ω have a "special" characteristic S and $N(1-p)$ have the opposite characteristic F and A_k = (exactly k "special" individuals are obtained in the sample), then

$$P(A_k) = \binom{n}{k}\frac{(Np)_k(N(1-p))_{n-k}}{(N)_n} = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}} \tag{A.6.10}$$

for $\max(0, n - N(1-p)) \le k \le \min(n, Np)$, and $P(A_k) = 0$ otherwise. The formula (A.6.10) is known as the *hypergeometric* probability.

References

Gnedenko (1967) Chapter 2, Section 11
Hoel, Port, and Stone (1971) Section 2.4
Parzen (1960) Chapter 3, Sections 1–4
Pitman (1993) Section 2.1

A.7 PROBABILITIES ON EUCLIDEAN SPACE

Random experiments whose outcomes are real numbers play a central role in theory and practice. The probability models corresponding to such experiments can all be thought of as having a Euclidean space for sample space.

We shall use the notation R^k of k-dimensional Euclidean space and denote members of R^k by symbols such as $\mathbf{x}$ or $(x_1, \dots, x_k)'$, where $(\ )'$ denotes transpose.

A.7.1 If $(a_1, b_1), \dots, (a_k, b_k)$ are k open intervals, we shall call the set $(a_1, b_1) \times \cdots \times (a_k, b_k) = \{(x_1, \dots, x_k) : a_i < x_i < b_i,\ 1 \le i \le k\}$ *an open k rectangle*.

A.7.2 The *Borel field* in R^k, which we denote by $\mathcal{B}^k$, is defined to be the smallest sigma field having all open k rectangles as members. Any subset of R^k we might conceivably be interested in turns out to be a member of $\mathcal{B}^k$. We will write R for R^1 and $\mathcal{B}$ for $\mathcal{B}^1$.

A.7.3 *A discrete (probability) distribution* on R^k is a probability measure P such that $\sum_{i=1}^{\infty} P(\{\mathbf{x}_i\}) = 1$ for some sequence of points $\{\mathbf{x}_i\}$ in R^k. That is, only an $\mathbf{x}_i$ can occur as an outcome of the experiment. This definition is consistent with (A.3.1) because the study of this model and that of the model that has $\Omega = \{\mathbf{x}_1, \dots, \mathbf{x}_n, \dots\}$ are equivalent.

The *frequency function* p of a discrete distribution is defined on R^k by

$$p(\mathbf{x}) = P(\{\mathbf{x}\}). \tag{A.7.4}$$

Conversely, any nonnegative function p on R^k vanishing except on a sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_n, \dots\}$ of vectors and that satisfies $\sum_{i=1}^{\infty} p(\mathbf{x}_i) = 1$ defines a unique discrete probability distribution by the relation

$$P(A) = \sum_{\mathbf{x}_i \in A} p(\mathbf{x}_i). \tag{A.7.5}$$

A.7.6 A nonnegative function p on R^k, which is integrable and which has

$$\int_{R^k} p(\mathbf{x}) d\mathbf{x} = 1,$$

where $d\mathbf{x}$ denotes $dx_1 \dots dx_n$, is called a *density function*. Integrals should be interpreted in the sense of Lebesgue. However, for practical purposes, Riemann integrals are adequate.

A.7.7 A *continuous probability distribution* on R^k is a probability P that is defined by the relation

$$P(A) = \int_A p(\mathbf{x}) d\mathbf{x} \tag{A.7.8}$$

for some density function p and all events A. P defined by A.7.8 are usually called absolutely continuous. We will only consider continuous probability distributions that are also absolutely continuous and drop the term absolutely. It may be shown that a function P so defined satisfies (A.1.4). Recall that the integral on the right of (A.7.8) is by definition

$$\int_{R^k} 1_A(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

where $1_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$, and 0 otherwise. Geometrically, $P(A)$ is the volume of the "cylinder" with base A and height $p(\mathbf{x})$ at $\mathbf{x}$. An important special case of (A.7.8) is given by

$$P((a_1, b_1) \times \cdots \times (a_k, b_k)) = \int_{a_k}^{b_k} \cdots \int_{a_1}^{b_1} p(\mathbf{x}) d\mathbf{x}. \tag{A.7.9}$$

It turns out that a continuous probability distribution determines the density that generates it "uniquely."$^{(1)}$

Although in a continuous model $P(\{\mathbf{x}\}) = 0$ for every $\mathbf{x}$, the density function has an operational interpretation close to that of the frequency function. For instance, if p is a continuous density on R, x_0 and x_1 are in R, and h is close to 0, then by the mean value theorem

$$P([x_0 - h, x_0 + h]) \approx 2hp(x_0) \text{ and } \frac{P([x_0 - h, x_0 + h])}{P([x_1 - h, x_1 + h])} \approx \frac{p(x_0)}{p(x_1)}. \tag{A.7.10}$$

The ratio $p(x_0)/p(x_1)$ can, thus, be thought of as measuring approximately how much more or less likely we are to obtain an outcome in a neighborhood of x_0 then one in a neighborhood of x_1.

A.7.11 The *distribution function* (d.f.) F is defined by

$$F(x_1, \ldots, x_k) = P((-\infty, x_1] \times \cdots \times (-\infty, x_k]). \tag{A.7.12}$$

The d.f. defines P in the sense that if P and Q are two probabilities with the same d.f., then $P = Q$. When $k = 1$, F is a function of a real variable characterized by the following properties:

$$0 \le F \le 1 \tag{A.7.13}$$

$$x \le y \Rightarrow F(x) \le F(y) \text{ (Monotone)} \tag{A.7.14}$$

$$x_n \downarrow x \Rightarrow F(x_n) \to F(x) \text{ (Continuous from the right)} \tag{A.7.15}$$

$$\begin{array}{l} \lim_{x\to\infty} F(x) = 1 \\ \lim_{x\to-\infty} F(x) = 0. \end{array} \tag{A.7.16}$$

It may be shown that any function F satisfying (A.7.13)–(A.7.16) defines a unique P on the real line. We always have

$$F(x) - F(x-0)^{(2)} = P(\{x\}). \tag{A.7.17}$$

Thus, F is continuous at x if and only if $P(\{x\}) = 0$.

References

Gnedenko (1967) Chapter 4, Sections 21, 22
Hoel, Port, and Stone (1971) Sections 3.1, 3.2, 5.1, 5.2
Parzen (1960) Chapter 4, Sections 1–4, 7
Pitman (1993) Sections 3.4, 4.1 and 4.5

A.8 RANDOM VARIABLES AND VECTORS: TRANSFORMATIONS

Although sample spaces can be very diverse, the statistician is usually interested primarily in one or more numerical characteristics of the sample point that has occurred. For example, we measure the weight of pigs drawn at random from a population, the time to breakdown and length of repair time for a randomly chosen machine, the yield per acre of a field of wheat in a given year, the concentration of a certain pollutant in the atmosphere, and so on. In the probability model, these quantities will correspond to random variables and vectors.

A.8.1 A *random variable* X is a function from Ω to R such that the set $\{\omega : X(\omega) \in B\} = X^{-1}(B)$ is in $\mathcal{A}$ for every $B \in \mathcal{B}$.[(1)]

A.8.2 A *random vector* $\mathbf{X} = (X_1, \dots, X_k)^T$ is k-tuple of random variables, or equivalently a function from Ω to R^k such that the set $\{\omega : \mathbf{X}(\omega) \in B\} = \mathbf{X}^{-1}(B)$ is in $\mathcal{A}$ for every $B \in \mathcal{B}^k$.[(1)] For $k = 1$ random vectors are just random variables. The event $\mathbf{X}^{-1}(B)$ will usually be written $[\mathbf{X} \in B]$ and $P([\mathbf{X} \in B])$ will be written $P[\mathbf{X} \in B]$.

The *probability distribution* of a random vector $\mathbf{X}$ is, by definition, the probability measure $P_{\mathbf{X}}$ in the model $(R^k, \mathcal{B}^k, P_{\mathbf{X}})$ given by

$$P_{\mathbf{X}}(B) = P[\mathbf{X} \in B]. \tag{A.8.3}$$

A.8.4 A random vector is said to have a *continuous* or *discrete distribution* (or to be *continuous* or *discrete*) according to whether its probability distribution is continuous or discrete. Similarly, we will refer to the *frequency function*, *density*, *d.f.*, and so on of a random vector when we are, in fact, referring to those features of its probability distribution. The subscript $\mathbf{X}$ or X will be used for densities, d.f.'s, and so on to indicate which vector or variable they correspond to unless the reference is clear from the context in which case they will be omitted.

The probability of any event that is expressible purely in terms of $\mathbf{X}$ can be calculated if we know only the probability distribution of $\mathbf{X}$. In the discrete case this means we need only know the frequency function and in the continuous case the density. Thus, from (A.7.5) and (A.7.8)

$$\begin{aligned} P[\mathbf{X} \in A] &= \sum_{\mathbf{x} \in A} p(\mathbf{x}), \text{ if } \mathbf{X} \text{ is discrete} \\ &= \int_A p(\mathbf{x}) d\mathbf{x}, \text{ if } \mathbf{X} \text{ is continuous.} \end{aligned} \tag{A.8.5}$$

When we are interested in particular random variables or vectors, we will describe them purely in terms of their probability distributions without any further specification of the underlying sample space on which they are defined.

The study of real- or vector-valued functions of a random vector $\mathbf{X}$ is central in the theory of probability and of statistics. Here is the formal definition of such transformations. Let g be any function from R^k to R^m, $k, m \geq 1$, such that[(2)] $\mathbf{g}^{-1}(B) = \{\mathbf{y} \in R^k : \mathbf{g}(\mathbf{y}) \in$

$B\} \in \mathcal{B}^k$ for every $B \in \mathcal{B}^m$. Then the *random transformation* $\mathbf{g}(\mathbf{X})$ is defined by

$$\mathbf{g}(\mathbf{X})(\omega) = \mathbf{g}(\mathbf{X}(\omega)). \tag{A.8.6}$$

An example of a transformation often used in statistics is $\mathbf{g} = (g_1, g_2)'$ with $g_1(\mathbf{X}) = k^{-1}\sum_{i=1}^k X_i = \bar{X}$ and $g_2(\mathbf{X}) = k^{-1}\sum_{i=1}^k (X_i - \bar{X})^2$. Another common example is $\mathbf{g}(\mathbf{X}) = (\min\{X_i\}, \max\{X_i\})'$.

The probability distribution of $\mathbf{g}(\mathbf{X})$ is completely determined by that of $\mathbf{X}$ through

$$P[\mathbf{g}(\mathbf{X}) \in B] = P[\mathbf{X} \in \mathbf{g}^{-1}(B)]. \tag{A.8.7}$$

If $\mathbf{X}$ is discrete with frequency function $p_{\mathbf{X}}$, then $\mathbf{g}(\mathbf{X})$ is discrete and has frequency function

$$p_{\mathbf{g}(\mathbf{X})}(\mathbf{t}) = \sum_{\{\mathbf{x}:\mathbf{g}(\mathbf{x})=\mathbf{t}\}} p_{\mathbf{X}}(\mathbf{x}). \tag{A.8.8}$$

Suppose that X is continuous with density p_X and g is real-valued and one-to-one$^{(3)}$ on an open set S such that $P[X \in S] = 1$. Furthermore, assume that the derivative g' of g exists and does not vanish on S. Then $g(X)$ is continuous with density given by

$$p_{g(X)}(t) = \frac{p_X(g^{-1}(t))}{|g'(g^{-1}(t))|} \tag{A.8.9}$$

for $t \in g(S)$, and 0 otherwise. This is called the *change of variable formula*.

If $g(X) = \sigma X + \mu$, $\sigma \neq 0$, and X is continuous, then

$$p_{g(X)}(t) = \frac{1}{|\sigma|} p_X\left(\frac{t-\mu}{\sigma}\right). \tag{A.8.10}$$

From (A.8.8) it follows that if $(X, Y)^T$ is a discrete random vector with frequency function $p_{(X,Y)}$, then the frequency function of X, known as the *marginal* frequency function, is given by$^{(4)}$

$$p_X(x) = \sum_y p_{(X,Y)}(x, y). \tag{A.8.11}$$

Similarly, if $(X, Y)^T$ is continuous with density $p_{(X,Y)}$, it may be shown (as a consequence of (A.8.7) and (A.7.8)) that X has a *marginal* density function given by

$$p_X(x) = \int_{-\infty}^{\infty} p_{(X,Y)}(x, y)dy.^{(5)} \tag{A.8.12}$$

These notions generalize to the case $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, a random vector obtained by putting two random vectors together. The (marginal) frequency or density of $\mathbf{X}$ is found as in (A.8.11) and (A.8.12) by summing or integrating out over $\mathbf{y}$ in $p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y})$.

Discrete random variables may be used to approximate continuous ones arbitrarily closely and vice versa.

In practice, all random variables are discrete because there is no instrument that can measure with perfect accuracy. Nevertheless, it is common in statistics to work with continuous distributions, which may be easier to deal with. The justification for this may be theoretical or pragmatic. One possibility is that the observed random variable or vector is obtained by rounding off to a large number of places the true unobservable continuous random variable specified by some idealized physical model. Or else, the approximation of a discrete distribution by a continuous one is made reasonable by one of the limit theorems of Sections A.15 and B.7.

A.8.13 A *convention*: We shall write $\mathbf{X} = \mathbf{Y}$ if the probability of the event $[\mathbf{X} \neq \mathbf{Y}]$ is 0.

References

Gnedenko (1967) Chapter 4, Sections 21–24
Grimmett and Stirzaker (1992) Section 4.7
Hoel, Port, and Stone (1971) Sections 3.3, 5.2, 6.1, 6.4
Parzen (1960) Chapter 7, Sections 1–5, 8, 9
Pitman (1993) Section 4.4

A.9 INDEPENDENCE OF RANDOM VARIABLES AND VECTORS

A.9.1 Two random variables X_1 and X_2 are said to be *independent* if and only if for sets A and B in $\mathcal{B}$, the events $[X_1 \in A]$ and $[X_2 \in B]$ are independent.

A.9.2 The random variables $X_1, \ldots, X_n$ are said to be (*mutually*) *independent* if and only if for any sets $A_1, \ldots, A_n$ in $\mathcal{B}$, the events $[X_1 \in A_1], \ldots, [X_n \in A_n]$ are independent. To generalize these definitions to random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (not necessarily of the same dimensionality) we need only use the events $[\mathbf{X}_i \in A_i]$ where A_i is a set in the range of $\mathbf{X}_i$.

A.9.3 By (A.8.7), if $\mathbf{X}$ and $\mathbf{Y}$ are independent, so are $\mathbf{g}(\mathbf{X})$ and $\mathbf{h}(\mathbf{Y})$, whatever be $\mathbf{g}$ and $\mathbf{h}$. For example, if (X_1, X_2) and (Y_1, Y_2) are independent, so are $X_1 + X_2$ and $Y_1 Y_2$, $(X_1, X_1 X_2)$ and Y_2, and so on.

Theorem A.9.1. *Suppose* $\mathbf{X} = (X_1, \ldots, X_n)$ *is either a discrete or continuous random vector. Then the random variables* $X_1, \ldots, X_n$ *are independent if, and only if, either of the following two conditions hold:*

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \text{ for all } x_1, \ldots, x_n \tag{A.9.4}$$

$$p_{\mathbf{X}}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n) \text{ for all } x_1, \ldots, x_n. \tag{A.9.5}$$

A.9.6 If the X_i are all continuous and independent, then $\mathbf{X} = (X_1, \ldots, , X_n)$ is continuous.

A.9.7 The preceding equivalences are valid for random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

A.9.8 If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent identically distributed k-dimensional random vectors with d.f. $F_{\mathbf{X}}$ or density (frequency function) $p_{\mathbf{X}}$, then $\mathbf{X}_1, \dots, \mathbf{X}_n$ is called a *random sample of size* n from a population with d.f. $F_{\mathbf{X}}$ or density (frequency function) $p_{\mathbf{X}}$. In statistics, such a random sample is often obtained by selecting n members at random in the sense of (A.3.4) from a population and measuring k characteristics on each member.

If A is any event, we define the random variable $1(A)$, *the indicator of the event* A, by

$$\begin{aligned} 1(A)(\omega) &= 1 \text{ if } \omega \in A \\ &= 0 \text{ otherwise.} \end{aligned} \tag{A.9.9}$$

If we perform n Bernoulli trials with probability of success p and we let X_i be the indicator of the event (success on the ith trial), then the X_i form a sample from a distribution that assigns probability p to 1 and $(1-p)$ to 0. Such samples will be referred to as the *indicators of* n *Bernoulli trials with probability of success* p.

References

Gnedenko (1967) Chapter 4, Sections 23, 24
Grimmett and Stirzaker (2001) Sections 3.2, 4.2
Hoel, Port, and Stone (1971) Section 3.4
Parzen (1960) Chapter 7, Sections 6, 7
Pitman (1993) Sections 2.5, 5.3

A.10 THE EXPECTATION OF A RANDOM VARIABLE

Let X be the height of an individual sampled at random from a finite population. Then a reasonable measure of the center of the distribution of X is the average height of an individual in the given population. If $x_1, \dots, x_q$ are the only heights present in the population, it follows that this average is given by $\sum_{i=1}^{q} x_i P[X = x_i]$ where $P[X = x_i]$ is just the proportion of individuals of height x_i in the population. The same quantity arises (approximately) if we use the long-run frequency interpretation of probability and calculate the average height of the individuals in a large sample from the population in question. In line with these ideas we develop the general concept of expectation as follows.

If X is a nonnegative, discrete random variable with possible values $\{x_1, x_2, \dots\}$, we define the *expectation* or *mean* of X, written $E(X)$, by

$$E(X) = \sum_{i=1}^{\infty} x_i p_X(x_i). \tag{A.10.1}$$

(Infinity is a possible value of $E(X)$. Take

$$x_i = i,\ p_X(i) = \frac{1}{i(i+1)},\ i = 1, 2, \dots .)$$

A.10.2 More generally, if X is discrete, decompose $\{x_1, x_2, \dots\}$ into two sets A and B where A consists of all nonnegative x_i and B of all negative x_i. If either $\sum_{x_i \in A} x_i p_X(x_i) <$

∞ or $\sum_{x_i \in B}(-x_i)p_X(x_i) < \infty$, we define $E(X)$ unambiguously by (A.10.1). Otherwise, we leave $E(X)$ undefined.

Here are some properties of the expectation that hold when X is discrete. If X is a constant, $X(\omega) = c$ for all ω, then

$$E(X) = c. \tag{A.10.3}$$

If $X = 1(A)$ (cf. (A.9.9)), then

$$E(X) = P(A). \tag{A.10.4}$$

If $\mathbf{X}$ is an n-dimensional random vector, if g is a real-valued function on R^n, and if $E(|g(\mathbf{X})|) < \infty$, then it may be shown that

$$E(g(\mathbf{X})) = \sum_{i=1}^{\infty} g(\mathbf{x}_i)p_{\mathbf{X}}(x_i). \tag{A.10.5}$$

As a consequence of this result, we have

$$E(|X|) = \sum_{i=1}^{\infty} |x_i|p_X(x_i). \tag{A.10.6}$$

Taking $g(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i x_i$ we obtain the fundamental relationship

$$E\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i E(X_i) \tag{A.10.7}$$

if $\alpha_1, \dots, \alpha_n$ are constants and $E(|X_i|) < \infty$, $i = 1, \dots, n$.

A.10.8 From (A.10.7) it follows that if $X \leq Y$ and $E(X)$, $E(Y)$ are defined, then $E(X) \leq E(Y)$.

If X is a continuous random variable, it is natural to attempt a definition of the expectation via approximation from the discrete case. Those familiar with Lebesgue integration will realize that this leads to

$$E(X) = \int_{-\infty}^{\infty} x p_X(x) dx \tag{A.10.9}$$

as the definition of the *expectation* or *mean* of X whenever $\int_0^{\infty} x p_X(x) dx$ or $\int_{-\infty}^{0} x p_X(x) dx$ is finite. Otherwise, $E(X)$ is left undefined.

A.10.10 A random variable X is said to be *integrable* if $E(|X|) < \infty$.

It may be shown that if $\mathbf{X}$ is a continuous k-dimensional random vector and $g(\mathbf{X})$ is any random variable such that

$$\int_{R^k} |g(\mathbf{x})|p_{\mathbf{X}}(\mathbf{x})d\mathbf{x} < \infty,$$

then $E(g(\mathbf{X}))$ exists and

$$E(g(\mathbf{X})) = \int_{R^k} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \tag{A.10.11}$$

In the continuous case expectation properties (A.10.3), (A.10.4), (A.10.7), and (A.10.8) as well as continuous analogues of (A.10.5) and (A.10.6) hold. It is possible to define the expectation of a random variable in general using discrete approximations. The interested reader may consult an advanced text such as Chung (1974), Chapter 3.

The formulae (A.10.5) and (A.10.11) are both sometimes written as

$$E(g(\mathbf{X})) = \int_{R^k} g(\mathbf{x}) dF(\mathbf{x}) \text{ or } \int_{R^k} g(\mathbf{x}) dP(\mathbf{x}) \tag{A.10.12}$$

where F denotes the distribution function of $\mathbf{X}$ and P is the probability function of $\mathbf{X}$ defined by (A.8.5).

A convenient notation is $dP(x) = p(x) d\mu(x)$, which means

$$\begin{aligned} \int g(\mathbf{x}) dP(\mathbf{x}) &= \sum_{i=1}^{\infty} \mathbf{g}(\mathbf{x}_i) p(\mathbf{x}_i), \text{ discrete case} \\ &= \int_{-\infty}^{\infty} g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \text{ continuous case.} \end{aligned} \tag{A.10.13}$$

We refer to $\mu = \mu_P$ as the *dominating measure* for P. In the discrete case μ assigns weight one to each of the points in $\{\mathbf{x} : p(\mathbf{x}) > 0\}$ and it is called *counting measure*. In the continuous case $d\mu(\mathbf{x}) = d\mathbf{x}$ and $\mu(\mathbf{x})$ is called *Lebesgue measure*. We will often refer to $p(\mathbf{x})$ as the *density* of $\mathbf{X}$ in the discrete case as well as the continuous case.

References

Chung (1974) Chapter 3
Gnedenko (1967) Chapter 5, Section 26
Grimmett and Stirzaker (1992) Sections 3.3, 4.3
Hoel, Port, and Stone (1971) Sections 4.1, 7.1
Parzen (1960) Chapter 5; Chapter 8, Sections 1–4
Pitman (1993) Sections 3.3, 3.4, 4.1

A.11 MOMENTS

A.11.1 If k is any natural number and X is a random variable, *the* kth *moment of* X is defined to be the expectation of X^k. We assume that all moments written here exist.

By (A.10.5) and (A.10.11),

$$\begin{aligned} E(X^k) &= \sum_x x^k p_X(x) \text{ if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} x^k p_X(x) dx \text{ if } X \text{ is continuous.} \end{aligned} \tag{A.11.2}$$

In general, the moments depend on the distribution of X only.

A.11.3 The distribution of a random variable is typically uniquely specified by its moments. This is the case, for example, if the random variable possesses a moment generating function (cf. (A.12.1)).

A.11.4 The *kth central moment* of X is by definition $E[(X - E(X))^k]$, the kth moment of $(X - E(X))$, and is denoted by μ_k.

A.11.5 The second central moment is called the *variance* of X and will be written Var X. The nonnegative square root of Var X is called the *standard deviation* of X. The standard deviation measures the spread of the distribution of X about its expectation. It is also called a measure of *scale*. Another measure of the same type is $E(|X - E(X)|)$, which is often referred to as the *mean deviation*.

The variance of X is finite if and only if the second moment of X is finite (cf. (A.11.15)).

If a and b are constants, then by (A.10.7)

$$\text{Var}(aX + b) = a^2 \text{ Var } X. \tag{A.11.6}$$

(One side of the equation exists if and only if the other does.)

A.11.7 If X is any random variable with well-defined (finite) mean and variance, the *standardized version* or *Z-score* of X is the random variable $Z = (X - E(X))/\sqrt{\text{Var } X}$. By (A.10.7) and (A.11.6) it follows then that

$$E(Z) = 0 \text{ and Var } Z = 1. \tag{A.11.8}$$

A.11.9 If $E(X^2) = 0$, then $X = 0$. If Var $X = 0$, $X = E(X)$ (a constant). These results follow, for instance, from (A.15.2).

A.11.10 The third and fourth central moments are used in the *coefficient of skewness* γ_1 and the *kurtosis* γ_2, which are defined by

$$\gamma_1 = \mu_3/\sigma^3, \ \gamma_2 = (\mu_4/\sigma^4) - 3$$

where $\sigma^2 = \text{Var } X$. See also Section A.12 where γ_1 and γ_2 are expressed in terms of cumulants. These descriptive measures are useful in comparing the shapes of various frequently used densities.

A.11.11 If $Y = a + bX$ with $b > 0$, then the coefficient of skewness and the kurtosis of Y are the same as those of X. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\gamma_1 = \gamma_2 = 0$.

A.11.12 It is possible to generalize the notion of moments to random vectors. For simplicity we consider the case $k = 2$. If X_1 and X_2 are random variables and i, j are natural numbers, then the *product moment* of order (i, j) of X_1 and X_2 is, by definition, $E(X_1^i X_2^j)$. The *central product moment* of order (i, j) of X_1 and X_2 is again by definition $E[(X_1 - E(X_1))^i (X_2 - E(X_2))^j]$. The central product moment of order $(1, 1)$ is called the *covariance* of X_1 and X_2 and is written $\text{Cov}(X_1, X_2)$. By expanding the product $(X_1 - E(X_1))(X_2 - E(X_2))$ and using (A.10.3) and (A.10.7), we obtain the relations,

$$\begin{aligned}&\text{Cov}(aX_1 + bX_2, cX_3 + dX_4)\\&= ac\,\text{Cov}(X_1, X_3) + bc\,\text{Cov}(X_2, X_3) + ad\,\text{Cov}(X_1, X_4) + bd\,\text{Cov}(X_2, X_4)\end{aligned} \tag{A.11.13}$$

and

$$\text{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2). \tag{A.11.14}$$

If X_1' and X_2' are distributed as X_1 and X_2 and are independent of X_1 and X_2, then

$$\text{Cov}(X_1, X_2) = \frac{1}{2}E(X_1 - X_1')(X_2 - X_2').$$

If we put $X_1 = X_2 = X$ in (A.11.14), we get the formula

$$\text{Var}\,X = E(X^2) - [E(X)]^2. \tag{A.11.15}$$

The covariance is defined whenever X_1 and X_2 have finite variances and in that case

$$|\text{Cov}(X_1, X_2)| \leq \sqrt{(\text{Var}\,X_1)(\text{Var}\,X_2)} \tag{A.11.16}$$

with equality holding if and only if

(1) X_1 or X_2 is a constant

or

(2) $(X_1 - E(X_1)) = \frac{\text{Cov}(X_1, X_2)}{\text{Var}\,X_2}(X_2 - E(X_2))$.

This is the *correlation inequality*. It may be obtained from the *Cauchy–Schwarz inequality*,

$$|E(Z_1Z_2)| \leq \sqrt{E(Z_1^2)E(Z_2^2)} \tag{A.11.17}$$

for any two random variables Z_1, Z_2 such that $E(Z_1^2) < \infty$, $E(Z_2^2) < \infty$. Equality holds if and only if one of Z_1, Z_2 equals 0 or $Z_1 = aZ_2$ for some constant a. The correlation inequality corresponds to the special case $Z_1 = X_1 - E(X_1)$, $Z_2 = X_2 - E(X_2)$. A proof of the Cauchy–Schwarz inequality is given in Remark 1.4.1.

The *correlation* of X_1 and X_2, denoted by $\text{Corr}(X_1, X_2)$, is defined whenever X_1 and X_2 are not constant and the variances of X_1 and X_2 are finite by

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{(\text{Var}\,X_1)(\text{Var}\,X_2)}}. \tag{A.11.18}$$

The correlation of X_1 and X_2 is the covariance of the standardized versions of X_1 and X_2. The correlation inequality is equivalent to the statement

$$|\text{Corr}(X_1, X_2)| \leq 1. \tag{A.11.19}$$

Equality holds if and only if X_2 is linear function $(X_2 = a + bX_1, b \neq 0)$ of X_1.

If $X_1, \dots, X_n$ have finite variances, we obtain as a consequence of (A.11.13) the relation

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var } X_i + 2\sum_{i<j} \text{Cov}(X_i, X_j). \qquad \text{(A.11.20)}$$

If X_1 and X_2 are independent and X_1 and X_2 are integrable, then

$$E(X_1 X_2) = E(X_1)E(X_2) \qquad \text{(A.11.21)}$$

or in view of (A.11.14),

$$\text{Cov}(X_1, X_2) = \text{Corr}(X_1, X_2) = 0 \text{ when Var}(X_i) > 0,\ i = 1, 2. \qquad \text{(A.11.22)}$$

This may be checked directly. It is not true in general that X_1 and X_2 that satisfy (A.11.22) (i.e., are *uncorrelated*) need be independent.

The correlation coefficient roughly measures the amount and sign of linear relationship between X_1 and X_2. It is -1 or 1 in the case of perfect relationship ($X_2 = a + bX_1, b < 0$ or $b > 0$, respectively). See also Section 1.4.

As a consequence of (A.11.22) and (A.11.20), we see that if $X_1, \dots, X_n$ are independent with finite variances, then

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var } X_i. \qquad \text{(A.11.23)}$$

References

Gnedenko (1967) Chapter 5, Sections 27, 28, 30
Hoel, Port, and Stone (1971) Sections 4.2–4.5, 7.3
Parzen (1960) Chapter 5; Chapter 8, Sections 1–4
Pitman (1993) Section 6.4

A.12 MOMENT AND CUMULANT GENERATING FUNCTIONS

A.12.1 If $E(e^{s_0|X|}) < \infty$ for some $s_0 > 0$, $M_X(s) = E(e^{sX})$ is well defined for $|s| \le s_0$ and is called the *moment generating function of* X. By (A.10.5) and (A.10.11),

$$\begin{aligned} M_X(s) &= \sum_{i=1}^{\infty} e^{sx_i} p_X(x_i) && \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} e^{sx} p_X(x)dx && \text{if } X \text{ is continuous.} \end{aligned} \qquad \text{(A.12.2)}$$

If M_X is well defined in a neighborhood $\{s : |s| \le s_0\}$ of zero, all moments of X are finite and

$$M_X(s) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} s^k,\ |s| \le s_0. \qquad \text{(A.12.3)}$$

A.12.4 The moment generating function M_X has derivatives of all orders at $s = 0$ and

$$\frac{d^k}{ds^k} M_X(s) \bigg|_{s=0} = E(X^k).$$

A.12.5 If defined, M_X determines the distribution of X uniquely and is itself uniquely determined by the distribution of X.

If $X_1, \dots, X_n$ are independent random variables with moment generating functions $M_{X_1}, \dots, M_{X_n}$, then $X_1 + \cdots + X_n$ has moment generating function given by

$$M_{(X_1+\cdots+X_n)}(s) = \prod_{i=1}^{n} M_{X_i}(s). \tag{A.12.6}$$

This follows by induction from the definition and (A.11.21). For a generalization of the notion of moment generating function to random vectors, see Section B.5.

The function

$$K_X(s) = \log M_X(s) \tag{A.12.7}$$

is called the *cumulant generating function* of X. If M_X is well defined in some neighborhood of zero, K_X can be represented by the convergent Taylor expansion

$$K_X(s) = \sum_{j=0}^{\infty} \frac{c_j}{j!} s^j \tag{A.12.8}$$

where

$$c_j = c_j(X) = \frac{d^j}{ds^j} K_X(s)|_{s=0} \tag{A.12.9}$$

is called the jth *cumulant* of X, $j \geq 1$. For $j \geq 2$ and any constant a, $c_j(X + a) = c_j(X)$. If X and Y are independent, then $c_j(X + Y) = c_j(X) + c_j(Y)$. The first cumulant c_1 is the mean μ of X, c_2 and c_3 equal the second and third central moments μ_2 and μ_3 of X, and $c_4 = \mu_4 - 3\mu_2^2$. The coefficients of skewness and kurtosis (see (A.11.10)) can be written as $\gamma_1 = c_3/c_2^{\frac{3}{2}}$ and $\gamma_2 = c_4/c_2^2$. If X is normally distributed, $c_j = 0$ for $j \geq 3$. See Problem B.3.8.

References

Hoel, Port, and Stone (1971) Chapter 8, Section 8.1
Parzen (1960) Chapter 5, Section 3; Chapter 8, Sections 2–3
Rao (1973) Section 2b.4

A.13 SOME CLASSICAL DISCRETE AND CONTINUOUS DISTRIBUTIONS

By definition, the probability distribution of a random variable or vector is just a probability measure on a suitable Euclidean space. In this section we introduce certain families of distributions, which arise frequently in probability and statistics, and list some of their properties. Following the name of each distribution we give a shorthand notation that will sometimes be used as will obvious abbreviations such as "binomial (n, θ)" for "the binomial distribution with parameter (n, θ)". The symbol p as usual stands for a frequency or density function. *If anywhere below p is not specified explicitly for some value of x it shall be assumed that p vanishes at that point. Similarly, if the value of the distribution function F is not specified outside some set, it is assumed to be zero to the "left" of the set and one to the "right" of the set.*

I. Discrete Distributions

The binomial distribution with parameters n and θ : $\mathcal{B}(n, \theta)$.

$$p(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}, \ k = 0, 1, \ldots, n. \tag{A.13.1}$$

The parameter n can be any integer ≥ 0 whereas θ may be any number in $[0, 1]$.

A.13.2 If X is the total number of successes obtained in n Bernoulli trials with probability of success θ, then X has a $\mathcal{B}(n, \theta)$ distribution (see (A.6.3)).

If X has a $\mathcal{B}(n, \theta)$ distribution, then

$$E(X) = n\theta, \ \text{Var}\, X = n\theta(1-\theta). \tag{A.13.3}$$

Higher-order moments may be computed from the moment generating function

$$M_X(t) = [\theta e^t + (1-\theta)]^n. \tag{A.13.4}$$

A.13.5 If $X_1, X_2, \ldots, X_k$ are independent random variables distributed as $\mathcal{B}(n_1, \theta)$, $\mathcal{B}(n_2, \theta), \ldots, \mathcal{B}(n_k, \theta)$, respectively, then $X_1 + X_2 + \cdots + X_k$ has a $\mathcal{B}(n_1 + \cdots + n_k, \theta)$ distribution. This result may be derived by using (A.12.5) and (A.12.6) in conjunction with (A.13.4).

The hypergeometric distribution with parameters D, N, and n : $\mathcal{H}(D, N, n)$.

$$p(k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} \tag{A.13.6}$$

for k a natural number with $\max(0, n-(N-D)) \leq k \leq \min(n, D)$. The parameters D and n may be any natural numbers that are less than or equal to the natural number N.

A.13.7 If X is the number of defectives (special objects) in a sample of size n taken without replacement from a population with D defectives and $N - D$ nondefectives, then X has an $\mathcal{H}(D, N, n)$ distribution (see (A.6.10)). If the sample is taken with replacement, X has a $\mathcal{B}(n, D/N)$ distribution.

If X has an $\mathcal{H}(D, N, n)$ distribution, then

$$E(X) = n\frac{D}{N}, \text{ Var } X = n\frac{D}{N}\left(1 - \frac{D}{N}\right)\frac{N-n}{N-1}. \qquad \text{(A.13.8)}$$

Formulae (A.13.8) may be obtained directly from the definition (A.13.6). An easier way is to use the interpretation (A.13.7) by writing $X = \sum_{j=1}^{n} I_j$ where $I_j = 1$ if the jth object sampled is defective and 0 otherwise, and then applying formulae (A.10.4), (A.10.7), and (A.11.20).

The Poisson distribution with parameter $\lambda : \mathcal{P}(\lambda)$**.**

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{(A.13.9)}$$

for $k = 0, 1, 2, \ldots$. The parameter λ can be any positive number.

If X has a $\mathcal{P}(\lambda)$ distribution, then

$$E(X) = \text{Var } X = \lambda. \qquad \text{(A.13.10)}$$

The moment generating function of X is given by

$$M_X(t) = e^{\lambda(e^t - 1)}. \qquad \text{(A.13.11)}$$

A.13.12 If $X_1, X_2, \ldots, X_n$ are independent random variables with $\mathcal{P}(\lambda_1), \mathcal{P}(\lambda_2), \ldots, \mathcal{P}(\lambda_n)$ distributions, respectively, then $X_1 + X_2 + \dot{+} X_n$ has the $\mathcal{P}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$ distribution. This result may be derived in th same manner as the corresponding fact for the binomial distribution.

The multinomial distribution with parameters $n, \theta_1, \ldots, \theta_q : \mathcal{M}(n, \theta_1, \ldots, \theta_q)$**.**

$$p(k_1, \ldots, k_q) = \frac{n!}{k_1! \ldots k_q!}\theta_1^{k_1} \ldots \theta_q^{k_q} \qquad \text{(A.13.13)}$$

whenever k_i are nonnegative integers such that $\sum_{i=1}^{q} k_i = n$. The parameter n is any natural number while $(\theta_1, \ldots, \theta_q)$ is any vector in

$$\Theta = \left\{(\theta_1, \ldots, \theta_q) : \theta_i \geq 0,\ 1 \leq i \leq q,\ \sum_{i=1}^{q} \theta_i = 1\right\}.$$

A.13.14 If $\mathbf{X} = (X_1, \ldots, X_q)'$, where X_i is the number of times outcome ω_i occurs in n multinomial trials with probabilities $(\theta_1, \ldots, \theta_q)$, then $\mathbf{X}$ has a $\mathcal{M}(n, \theta_1, \ldots, \theta_q)$ distribution (see (A.6.6)).

If $\mathbf{X}$ has a $\mathcal{M}(n, \theta_1, \ldots, \theta_q)$ distribution,

$$\begin{aligned} E(X_i) &= n\theta_i, \text{ Var } X_i = n\theta_i(1 - \theta_i) \\ \text{Cov}(X_i, X_j) &= -n\theta_i\theta_j,\ i \neq j,\ i, j = 1, \ldots, q. \end{aligned} \qquad \text{(A.13.15)}$$

These results may either be derived directly or by a representation such as that discussed in (A.13.8) and an application of formulas (A.10.4), (A.10.7), (A.13.13), and (A.11.20).

A.13.16 If $\mathbf{X}$ has a $\mathcal{M}(n, \theta_1, \dots, \theta_q)$ distribution, then $(X_{i_1}, \dots, X_{i_s},\ n - \sum_{j=1}^{s} X_{i_j})'$ has a $\mathcal{M}(n, \theta_{i_1}, \dots, \theta_{i_s},\ 1 - \sum_{j=1}^{s} \theta_{i_j})$ distribution for any set $\{i_1, \dots, i_s\} \subset \{1, \dots, q\}$. Therefore, X_j has $\mathcal{B}(n, \theta_j)$ distributions for each j and more generally $\sum_{j=1}^{s} X_{i_j}$ has a $\mathcal{B}(n, \sum_{j=1}^{s} \theta_{i_j})$ distribution if $s < q$. These remarks follow from the interpretation (A.13.14).

II. Continuous Distributions

Before beginning our listing we introduce some convenient notations: $X \sim F$ will mean that X is a random variable with d.f. F, and $X \sim p$ will similarly mean that X has density or frequency function p.

Let Y be a random variable with d.f. F. Let F_μ be the d.f. of $Y + \mu$. The family $\mathcal{F}_L = \{F_\mu : -\infty < \mu < \infty\}$ is called a *location parameter family*, μ is called a *location parameter*, and we say that Y *generates* $\mathcal{F}_L$. By definition, for any μ, $X \sim F_\mu \Leftrightarrow X - \mu \sim F$. Therefore, for any μ, γ,

$$F_\mu(x) = F(x - \mu) = F_0(x - \mu) = F_\gamma(x + (\gamma - \mu))$$

and all calculations involving F_μ can be referred back to F or any other member of the family. Similarly, if Y generates $\mathcal{F}_L$ so does $Y + \gamma$ for any fixed γ. If Y has a first moment, it follows that we may without loss of generality (as far as generating $\mathcal{F}_L$ goes) assume that $E(Y) = 0$. Then if $X \sim F_\mu$, $E(X) = \mu$.

Similarly let F_σ^* be the d.f. of σY, $\sigma > 0$. The family $\mathcal{F}_S = \{F_\sigma^* : \sigma > 0\}$ is called a *scale parameter family*, σ is a *scale parameter*, and Y is said to *generate* $\mathcal{F}_S$. By definition, for any $\sigma > 0$, $X \sim F_\sigma^* \Leftrightarrow X/\sigma \sim F$. Again all calculations involving one member of the family can be referred back to any other because for any $\sigma, \tau > 0$,

$$F_\sigma^*(x) = F_\tau^*\left(\frac{\tau x}{\sigma}\right).$$

If Y generates $\mathcal{F}_S$ and Y has a first moment different from 0, we may without loss of generality take $E(Y) = 1$ and, hence, if $X \sim F_\sigma^*$, then $E(X) = \sigma$. Alternatively, if Y has a second moment, we may select F as being the unique member of the family $\mathcal{F}_S$ having $\text{Var}\, Y = 1$ and then $X \sim F_\sigma^* \Rightarrow \text{Var}\, X = \sigma^2$. Finally, define $F_{\mu,\sigma}$ as the d.f. of $\sigma Y + \mu$. The family $\mathcal{F}_{L,S} = \{F_{\mu,\sigma} : -\infty < \mu < \infty, \sigma > 0\}$ is called a *location-scale parameter family*, μ is called a *location parameter*, and σ a *scale parameter*, and Y is said to generate $\mathcal{F}_{L,S}$. From

$$F_{\mu,\sigma}(x) = F\left(\frac{x - \mu}{\sigma}\right) = F_{\gamma,\tau}\left(\frac{\tau(x - \mu)}{\sigma} + \gamma\right),$$

we see as before how to refer calculations involving one member of the family back to any other. Without loss of generality, if Y has a second moment, we may take

$$E(Y) = 0,\ \text{Var}\, Y = 1.$$

Then if $X \sim F_{\mu,\sigma}$, we obtain

$$E(X) = \mu, \text{ Var } X = \sigma^2.$$

Clearly $F_\mu = F_{\mu,1}$, $F_\sigma^* = F_{0,\sigma}$.

The relation between the density of $F_{\mu,\sigma}$ and that of F is given by (A.8.10). All the families of densities we now define are location-scale or scale families.

The normal (Gaussian) distribution with parameters μ and σ^2 : $\mathcal{N}(\mu, \sigma^2)$.

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \tag{A.13.17}$$

The parameter μ can be any real number while σ is positive. The normal distribution with $\mu = 0$ and $\sigma = 1$ is known as the *standard normal distribution*. Its density will be denoted by $\varphi(z)$ and its d.f. by $\Phi(z)$.

A.13.18 The family of $\mathcal{N}(\mu, \sigma^2)$ distributions is a location-scale family. If Z has a $\mathcal{N}(0, 1)$ distribution, then $\sigma Z + \mu$ has a $\mathcal{N}(\mu, \sigma^2)$ distribution, and conversely if X has a $\mathcal{N}(\mu, \sigma^2)$ distribution, then $(X - \mu)/\sigma$ has a standard normal distribution.

If X has a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

$$E(X) = \mu, \text{ Var } X = \sigma^2. \tag{A.13.19}$$

More generally, all moments may be obtained from

$$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \tag{A.13.20}$$

for $-\infty < t < \infty$. In particular if $\mu = 0$, $\sigma^2 = 1$, then

$$M_X(t) = \sum_{k=0}^{\infty} \left[\frac{(2k)!}{2^k k!}\right] \frac{t^{2k}}{(2k)!} \tag{A.13.21}$$

and, hence, in this case we can conclude from (A.12.4) that

$$\begin{aligned} E(X^k) &= 0 && \text{if } k \geq 0 \text{ is odd} \\ E(X^k) &= \frac{k!}{2^{k/2}(k/2)!} && \text{if } k \geq 0 \text{ is even.} \end{aligned} \tag{A.13.22}$$

A.13.23 If $X_1, \ldots, X_n$ are independent normal random variables such that $E(X_i) = \mu_i$, Var $X_i = \sigma_i^2$, and $c_1, \ldots, c_n$ are any constants that are not all 0, then $\sum_{i=1}^n c_i X_i$ has a $\mathcal{N}(c_1\mu_1 + \cdots + c_n\mu_n, c_1^2\sigma_1^2 + \cdots + c_n^2\sigma_n^2)$ distribution. This follows from (A.13.20), (A.12.5), and (A.12.6).

Further information about the normal distribution may be found in Section A.15 and Appendix B.

The exponential distribution with parameter λ : $\mathcal{E}(\lambda)$.

$$p(x) = \lambda e^{-\lambda x},\ x > 0. \tag{A.13.24}$$

The range of λ is $(0, \infty)$. The distribution function corresponding to this p is given by

$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0. \tag{A.13.25}$$

A.13.26 If $\sigma = 1/\lambda$, then σ is a scale parameter. $\mathcal{E}(1)$ is called the *standard exponential distribution*.

If X has an $\mathcal{E}(\lambda)$ distribution,

$$E(X) = \frac{1}{\lambda},\ \text{Var}\, X = \frac{1}{\lambda^2}. \tag{A.13.27}$$

More generally, all moments may be obtained from

$$M_X(t) = \frac{1}{1 - (t/\lambda)} = \sum_{k=0}^{\infty} \left[\frac{k!}{\lambda^k}\right] \frac{t^k}{k!} \tag{A.13.28}$$

which is well defined for $t < \lambda$.

Further information about the exponential distribution may be found in Appendix B.

The uniform distribution on $(a, b) : \mathcal{U}(a, b)$**.**

$$p(x) = \frac{1}{(b-a)},\ a < x < b \tag{A.13.29}$$

where (a, b) is any pair of real numbers such that $a < b$. The corresponding distribution function is given by

$$F(x) = \frac{(x-a)}{(b-a)} \text{ for } a < x < b. \tag{A.13.30}$$

If X has a $\mathcal{U}(a, b)$ distribution, then

$$E(X) = \frac{a+b}{2},\ \text{Var}\, X = \frac{(b-a)^2}{12}. \tag{A.13.31}$$

A.13.32 If we set $\mu = a$, $\sigma = (b - a)$, then we can check that the $\mathcal{U}(a, b)$ family is a location-scale family generated by Y, where $Y \sim \mathcal{U}(0, 1)$.

References

Gnedenko (1967) Chapter 4, Sections 21–24; Chapter 5, Sections 26–28, 30
Hoel, Port, and Stone (1971) Sections 3.4.1, 5.3.1, 5.3.2
Parzen (1962) Chapter 4, Sections 4–6; Chapter 5; Chapter 6
Pitman (1993) pages 475–487

A.14 MODES OF CONVERGENCE OF RANDOM VARIABLES AND LIMIT THEOREMS

Much of probability theory can be viewed as variations, extensions, and generalizations of two basic results, the central limit theorem and the law of large numbers. Both of these theorems deal with the limiting behavior of sequences of random variables. The notions of limit that are involved are the subject of this section. All limits in the section are as $n \to \infty$.

A.14.1 We say that the sequence of random variables $\{Z_n\}$ *converges* to the random variable Z in *probability* and write $Z_n \overset{P}{\to} Z$ if $P[|Z_n - Z| \geq \epsilon] \to 0$ as $n \to \infty$ for every $\epsilon > 0$. That is, $Z_n \overset{P}{\to} Z$ if the chance that Z_n and Z differ by any given amount is negligible for n large enough.

A.14.2 We say that the sequence $\{Z_n\}$ *converges in law (in distribution)* to Z and write $Z_n \overset{\mathcal{L}}{\to} Z$ if $F_{Z_n}(t) \to F_Z(t)$ *for every point* t *such that* F_Z *is continuous at* t. (Recall that F_Z is continuous at t if and only if $P[Z = t] = 0$ (A.7.17).) This is the mode of convergence needed for approximation of one distribution by another.

$$\text{If } Z_n \overset{P}{\to} Z, \text{ then } Z_n \overset{\mathcal{L}}{\to} Z. \tag{A.14.3}$$

Because convergence in law requires nothing of the joint distribution of the Z_n and Z whereas convergence in probability does, it is not surprising and easy to show that, in general, convergence in law does not imply convergence in probability (e.g., Chung, 1974), but consider the following.

A.14.4 If $Z = z_0$ (a constant), convergence in law of $\{Z_n\}$ to Z implies convergence in probability.

Proof. Note that $z_0 \pm \epsilon$ are points of continuity of F_Z for every $\epsilon > 0$. Then

$$\begin{aligned} P[|Z_n - z_0| \geq \epsilon] &= 1 - P(Z_n < z_0 + \epsilon) + P(Z_n \leq z_0 - \epsilon) \\ &\leq 1 - F_{Z_n}\left(z_0 + \frac{\epsilon}{2}\right) + F_{Z_n}(z_0 - \epsilon). \end{aligned} \tag{A.14.5}$$

By assumption the right-hand side of (A.14.5) converges to $(1 - F_Z(z_0 + \epsilon/2)) + F_Z(z_0 - \epsilon) = 0$. □

A.14.6 If $Z_n \overset{P}{\to} z_0$ (a constant) and g is continuous at z_0, then $g(Z_n) \overset{P}{\to} g(z_0)$.

Proof. If ϵ is positive, there exists a δ such that $|z - z_0| < \delta$ implies $|g(z) - g(z_0)| < \epsilon$. Therefore,

$$P[|g(Z_n) - g(z_0)| < \epsilon] \geq P[|Z_n - z_0| < \delta] = 1 - P[|Z_n - z_0| \geq \delta]. \tag{A.14.7}$$

Because the right-hand side of (A.14.7) converges to 1, by the definition (A.14.1) the result follows. □

A more general result is given by the following.

A.14.8 If $Z_n \xrightarrow{\mathcal{L}} Z$ and g is continuous, then $g(Z_n) \xrightarrow{\mathcal{L}} g(Z)$.

The following theorem due to Slutsky will be used repeatedly.

Theorem A.14.9 *If* $Z_n \xrightarrow{\mathcal{L}} Z$ *and* $U_n \xrightarrow{P} u_0$ *(a constant), then*

(a) $Z_n + U_n \xrightarrow{\mathcal{L}} Z + u_0$,

(b) $U_n Z_n \xrightarrow{\mathcal{L}} u_0 Z$.

Proof. We prove (a). The other claim follows similarly. Begin by writing

$$\begin{aligned} F_{(Z_n+U_n)}(t) &= P[Z_n + U_n \le t,\ U_n \ge u_0 - \epsilon] \\ &\quad + P[Z_n + U_n \le t,\ U_n < u_0 - \epsilon]. \end{aligned} \tag{A.14.10}$$

Let t be a point of continuity of $F_{(Z+u_0)}$. Because a distribution function has at most countably many points of discontinuity, we may for any t choose ϵ positive and arbitrarily small such that $t \pm \epsilon$ are both points of continuity of $F_{(Z+u_0)}$. Now by (A.14.10),

$$F_{(Z_n+U_n)}(t) \le P[Z_n \le t - u_0 + \epsilon] + P[|U_n - u_0| > \epsilon]. \tag{A.14.11}$$

Moreover,

$$P[Z_n \le t - u_0 + \epsilon] = F_{(Z_n+u_0)}(t + \epsilon). \tag{A.14.12}$$

Because $F_{Z_n+u_0}(t) = P[Z_n \le t - u_0] = F_{Z_n}(t - u_0)$, we must have $Z_n + u_0 \xrightarrow{\mathcal{L}} Z + u_0$. Thus,

$$\limsup_n F_{(Z_n+U_n)}(t) \le \lim_n F_{(Z_n+u_0)}(t+\epsilon) + \lim_n P[|U_n - u_0| \ge \epsilon] = F_{(Z+u_0)}(t+\epsilon). \tag{A.14.13}$$

Similarly,

$$1 - F_{(Z_n+U_n)}(t) = P[Z_n + U_n > t] \le P[Z_n > t - u_0 - \epsilon] + P[|U_n - u_0| > \epsilon]. \tag{A.14.14}$$

and, hence,

$$\liminf_n F_{(Z_n+U_n)}(t) \ge \lim_n F_{(Z_n+u_0)}(t - \epsilon) = F_{(Z+u_0)}(t - \epsilon). \tag{A.14.15}$$

Therefore,

$$F_{(Z+u_0)}(t-\epsilon) \le \liminf_n F_{(Z_n+U_n)}(t) \le \limsup_n F_{(Z_n+U_n)}(t) \le F_{(Z+u_0)}(t+\epsilon). \tag{A.14.16}$$

Because ϵ may be permitted to tend to 0 and $F_{(Z+u_0)}$ is continuous at t, the result follows. □

A.14.17 Corollary. *Suppose that* a_n *is a sequence of constants tending to* ∞, b *is a fixed number, and* $a_n(Z_n - b) \xrightarrow{\mathcal{L}} X$. *Let* g *be a function of a real variable that is differentiable and whose derivative* g' *is continuous at* b.[1] *Then*

$$a_n[g(Z_n) - g(b)] \xrightarrow{\mathcal{L}} g'(b)X. \tag{A.14.18}$$

Proof. By Slutsky's theorem

$$Z_n - b = \frac{1}{a_n}[a_n(Z_n - b)] \xrightarrow{\mathcal{L}} 0 \cdot X = 0. \tag{A.14.19}$$

By (A.14.4), $|Z_n - b| \xrightarrow{P} 0$. Now apply the mean value theorem to $g(Z_n) - g(b)$ getting

$$a_n[g(Z_n) - g(b)] = a_n[g'(Z_n^*)(Z_n - b)]$$

where $|Z_n^* - b| \leq |Z_n - b|$. Because $|Z_n - b| \xrightarrow{P} 0$, so does $|Z_n^* - b|$ and, hence, $Z_n^* \xrightarrow{P} b$. By the continuity of g' and (A.14.6), $g'(Z_n^*) \xrightarrow{P} g'(b)$. Therefore, applying (A.14.9) again, $g'(Z_n^*)[a_n(Z_n - b)] \xrightarrow{\mathcal{L}} g'(b)X$. □

A.14.20 Suppose that $\{Z_n\}$ takes on only natural number values and $p_{Z_n}(z) \to p_Z(z)$ for all z. Then $Z_n \xrightarrow{\mathcal{L}} Z$.

This is immediate because whatever be z, $F_{Z_n}(z) = \sum_{k=0}^{[z]} p_{Z_n}(k) \to \sum_{k=0}^{[z]} p_Z(k) = F_Z(z)$ where $[z]$ = greatest integer $\leq z$. The converse is also true and easy to establish.

A.14.21 (Scheffé) Suppose that $\{Z_n\}$, Z are continuous and $p_{Z_n}(z) \to p_Z(z)$ for (almost) all z. Then $Z_n \xrightarrow{\mathcal{L}} Z$ (Hajek and Sidak, 1967, p. 64).

References

Grimmett and Stirzaker (1992) Sections 7.1–7.4

A.15 FURTHER LIMIT THEOREMS AND INEQUALITIES

The Bernoulli law of large numbers, which we give next, brings us back to the motivation of our definition of probability. There we noted that in practice the relative frequency of occurrence of an event A in many repetitions of an experiment tends to stabilize and that this "limit" corresponds to what we mean by the probability of A. Now having defined probability and independence of experiments abstractly we can prove that, in fact, the relative frequency of occurrence of an event A approaches its probability as the number of repetitions of the experiment increases. Such results and their generalizations are known as laws of large numbers. The first was discovered by Bernoulli in 1713. Its statement is as follows.

Bernoulli's (Weak) Law of Large Numbers

If $\{S_n\}$ is a sequence of random variables such that S_n has a $\mathcal{B}(n, p)$ distribution for $n \geq 1$, then

$$\frac{S_n}{n} \xrightarrow{P} p. \tag{A.15.1}$$

As in (A.13.2), think of S_n as the number of successes in n binomial trials in which we identify success with occurrence of A and failure with occurrence of A^c. Then S_n/n can be interpreted as the relative frequency of occurrence of A in n independent repetitions of the experiment in which A is an event and the Bernoulli law is now evidently a statement of the type we wanted.

Bernoulli's proof of this result was rather complicated and it remained for the Russian mathematician Chebychev to give a two-line argument. His generalization of Bernoulli's result is based on an inequality that has proved to be of the greatest importance in probability and statistics.

Chebychev's Inequality

If X is any random variable, then

$$P[|X| \geq a] \leq \frac{E(X^2)}{a^2}. \tag{A.15.2}$$

The Bernoulli law follows readily from (A.15.2) and (A.13.3) via the calculation

$$p\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] \leq \frac{E(S_n/n - p)^2}{\epsilon^2} = \frac{\text{Var } S_n}{n^2\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \to 0 \text{ as } n \to \infty. \tag{A.15.3}$$

A generalization of (A.15.2), which contains various important and useful inequalities, is the following. Let g be a nonnegative function on R such that g is nondecreasing on the range of a random variable Z. Then

$$P[Z \geq a] \leq \frac{E(g(Z))}{g(a)}. \tag{A.15.4}$$

If we put $Z = |X|$, $g(t) = t^2$ if $t \geq 0$ and 0 otherwise, we get (A.15.2). Other important cases are obtained by taking $Z = |X|$ and $g(t) = t$ if $t \geq 0$ and 0 otherwise (*Markov's inequality*), and $Z = X$ and $g(t) = e^{st}$ for $s > 0$ and all real t (*Bernstein's inequality*, see B.9.5 for the binomial case Bernstein's inequality).

Proof of (A.15.4). Note that by the properties of g,

$$g(a)1[Z \geq a] \leq g(Z)1[Z \geq a] \leq g(Z). \tag{A.15.5}$$

Therefore, by (A.10.8)

$$g(a)P[Z \geq a] = E(g(a)1[Z \geq a]) \leq E(g(Z)), \tag{A.15.6}$$

which is equivalent to (A.15.4). □

The following result, which follows from Chebychev's inequality, is a useful generalization of Bernoulli's law.

Khintchin's (Weak) Law of Large Numbers

Let $\{X_i\}$, $i \geq 1$, be a sequence of independent identically distributed random variables with finite mean μ and define $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{P} \mu. \tag{A.15.7}$$

Upon taking the X_i to be indicators of binomial trials, we obtain (A.15.1).

De Moivre–Laplace Theorem

Suppose that $\{S_n\}$ is a sequence of random variables such that for each n, S_n has a $\mathcal{B}(n, p)$ distribution where $0 < p < 1$. Then

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} Z, \tag{A.15.8}$$

where Z has a standard normal distribution. That is, the standardized versions of S_n converge in law to a standard normal random variable. If we write

$$\frac{S_n - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\frac{S_n}{n} - p\right)$$

and use (A.14.9), it is easy to see that (A.15.8) implies (A.15.1).

The De Moivre–Laplace theorem is generalized by the following.

Central Limit Theorem

Let $\{X_i\}$ be a sequence of independent identically distributed random variables with (common) expectation μ and variance σ^2 such that $0 < \sigma^2 < \infty$. Then, if $S_n = \sum_{i=1}^n X_i$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} Z, \tag{A.15.9}$$

where Z has the standard normal distribution.

The last two results are most commonly used in statistics as approximation theorems. Let k and l be nonnegative integers. The De Moivre–Laplace theorem is used as

$$\begin{aligned} P[k \le S_n \le l] &= P\left[k - \frac{1}{2} \le S_n \le l + \frac{1}{2}\right] \\ &= P\left[\frac{k - np - \frac{1}{2}}{\sqrt{npq}} \le \frac{S_n - np}{\sqrt{npq}} \le \frac{l - np + \frac{1}{2}}{\sqrt{npq}}\right] \\ &\approx \Phi\left(\frac{l - np + \frac{1}{2}}{\sqrt{npq}}\right) - \Phi\left(\frac{k - np - \frac{1}{2}}{\sqrt{npq}}\right) \end{aligned} \tag{A.15.10}$$

where $q = 1 - p$. The $\frac{1}{2}$ appearing in $k - \frac{1}{2}$ and $l + \frac{1}{2}$ is called the *continuity correction*. We have an excellent idea of how good this approximation is. An illustrative discussion is given in Feller (1968, pp. 187–188). A rule of thumb is that for most purposes the approximation can be used when np and $n(1-p)$ are both larger than 5.

Only when the X_i are integer-valued is the first step of (A.15.10) followed. Otherwise (A.15.9) is applied in the form

$$P[a \le S_n \le b] \approx \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{a - n\mu}{\sqrt{n}\sigma}\right). \tag{A.15.11}$$

The central limit theorem (and some of its generalizations) are also used to justify the assumption that “most” random variables that are measures of numerical characteristics of real populations, such as intelligence, height, weight, and blood pressure, are approximately normally distributed. The argument is that the observed numbers are sums of a large number of small (unobserved) independent factors. That is, each of the characteristic variables is expressible as a sum of a large number of small variables such as influences of particular genes, elements in the diet, and so on. For example, height is a sum of factors corresponding to heredity and environment.

If a bound for $E|X_i - \mu|^3$ is known, it is possible to give a theoretical estimate of the error involved in replacing $P(S_n \leq b)$ by its normal approximation:

Berry–Esséen Theorem

Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $\sigma^2 > 0$. Then, for all n,

$$\sup_t \left| P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq t \right) - \Phi(t) \right| \leq \frac{33}{4} \frac{E|X_1 - \mu|^3}{\sqrt{n}\sigma^3}. \tag{A.15.12}$$

For a proof, see Chung (1974, p. 224).

In practice, if we need the distribution of S_n we try to calculate it exactly for small values of n and then observe empirically when the approximation can be used with safety. This process of combining a limit theorem with empirical investigations is applicable in many statistical situations where the distributions of transformations $g(\mathbf{x})$ (see A.8.6) of interest become progressively more difficult to compute as the sample size increases and yet tend to stabilize. Examples of this process may be found in Chapter 5.

We conclude this section with two simple limit theorems that lead to approximations of one classical distribution by another. The very simple proofs of these results may, for instance, be found in Gnedenko (1967, p. 53 and p. 105).

A.15.12 The first of these results reflects the intuitively obvious fact that if the populations sampled are large and the samples are comparatively small, sampling with and without replacement leads to approximately the same probability distribution. Specifically, suppose that $\{X_N\}$ is a sequence of random variables such that X_N has a hypergeometric $\mathcal{H}(D_N, N, n)$, distribution where $D_N/N \to p$ as $N \to \infty$ and n is fixed. Then

$$p_{X_N}(k) \to \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{A.15.13}$$

as $N \to \infty$ for $k = 0, 1, \ldots, n$. By (A.14.20) we conclude that

$$X_n \xrightarrow{\mathcal{L}} X, \tag{A.15.14}$$

where X has a $\mathcal{B}(n, p)$ distribution. The approximation of the hypergeometric distribution by the binomial distribution indicated by this theorem is rather good. For instance, if $N = 50$, $n = 5$, and $D = 20$, the approximating binomial distribution to $\mathcal{H}(D, N, n)$ is $\mathcal{B}(5, 0.4)$. If $\mathcal{H}$ holds, $P[X \leq 2] = 0.690$ while under the approximation,

$P[X \le 2] = 0.683$. As indicated in this example, the approximation is reasonable when $(n/N) \le 0.1$.

The next elementary result, due to Poisson, plays an important role in advanced probability theory.

Poisson's Theorem

Suppose that $\{X_n\}$ is a sequence of random variables such that X_n has a $\mathcal{B}(n, p_n)$ distribution and $np_n \to \lambda$ as $n \to \infty$, where $0 \le \lambda < \infty$. Then

$$p_{X_n}(k) \to \frac{e^{-\lambda}\lambda^k}{k!} \tag{A.15.15}$$

for $k = 0, 1, 2, \ldots$ as $n \to \infty$. By (A.14.20) it follows that $X_n \xrightarrow{\mathcal{L}} X$ where X has a $\mathcal{P}(\lambda)$ distribution. This theorem suggests that we approximate the $\mathcal{B}(n, p)$ distribution by the $\mathcal{P}(np)$ distribution. Tables 3 on p. 108 and 2 on p. 154 of Feller (1968) indicate the excellence of the approximation when p is small and np is moderate. It may be shown that the error committed is always bounded by np^2.

References

Gnedenko (1967) Chapter 2, Section 13; Chapter 6, Section 32; Chapter 8, Section 42
Hoel, Port, and Stone (1971) Chapter 3, Section 3.4.2
Parzen (1960) Chapter 5, Sections 4, 5; Chapter 6, Section 2; Chapter 10, Section 2

A.16 POISSON PROCESS

A.16.1 A *Poisson process with parameter* λ is a collection of random variables $\{N(t)\}$, $t > 0$, such that

(i) $N(t)$ has a $\mathcal{P}(\lambda t)$ distribution for each t.

(ii) $N(t + h) - N(t)$ is independent of $N(s)$ for all $s \le t$, $h > 0$, and has a $\mathcal{P}(\lambda h)$ distribution.

Poisson processes are frequently applicable when we study phenomena involving events that occur "rarely" in small time intervals. For example, if $N(t)$is the number of disintegrations of a fixed amount of some radioactive substance in the period from time 0 to time t, then $\{N(t)\}$ is a Poisson process. The numbers $N(t)$ of "customers" (people, machines, etc.) arriving at a service counter from time 0 to time t are sometimes well approximated by a Poisson process as is the number of people who visit a WEB site from time 0 to t. Many interesting examples are discussed in the books of Feller (1968), Parzen (1962), Karlin (1969). In each of the preceding examples of a Poisson process $N(t)$ represents the number of times an "event" (radioactive disintegration, arrival of a customer) has occurred in the time from 0 to t. We use the word *event* here for lack of a better one because these are not events in terms of the probability model on which the $N(t)$ are defined. If we keep temporarily to this notion of event as a recurrent phenomenon that is randomly determined in some fashion and define $N(t)$ as the number of events occurring between time 0 and time t, we can ask under what circumstances $\{N(t)\}$ will form a Poisson process.

A.16.2 Formally, let $\{N(t)\}$, $t > 0$ be a collection of natural number valued random variables. It turns out that, $\{N(t)\}$ is a Poisson process with parameter λ if and only if the following conditions hold:

(a) $N(t+h) - N(t)$ is independent of $N(s)$, $s \le t$, for $h > 0$,

(b) $N(t+h) - N(t)$ has the same distribution as $N(h)$ for $h > 0$,

(c) $P[N(h) = 1] = \lambda h + o(h)$, and

(d) $P[N(h) > 1] = o(h)$.

(The quantity $o(h)$ is such that $o(h)/h \to 0$ as $h \to 0$.) Physically, these assumptions may be interpreted as follows.

(i) The time of recurrence of the "event" is unaffected by past occurrences.

(ii) The distribution of the number of occurrences of the "event" depends only on the length of the time for which we observe the process.

(iii) and (iv) The chance of any occurrence in a given time period goes to 0 as the period shrinks and having only one occurrence becomes far more likely than multiple occurrences.

This assertion may be proved as follows. Fix t and divide $[0, t]$ into n intervals $[0, t/n]$, $(t/n, 2t/n], \ldots, ((n-1)t/n, t]$. Let I_{jn} be the indicator of the event $[N(jt/n) - N((j-1)t/n) \ge 1]$ and define $N_n(t) = \sum_{j=1}^{n} I_{jn}$. Then $N_n(t)$ differs from $N(t)$ only insofar as multiple occurrences in one of the small subintervals are only counted as one occurrence. By (a) and (b), $N_n(t)$ has a $\mathcal{B}(n, P[N(t/n) \ge 1])$ distribution. From (c) and (d) and Theorem (A.15.15) we see that $N_n(t) \xrightarrow{\mathcal{L}} Z$, where Z has a $\mathcal{P}(\lambda t)$ distribution. On the other hand,

$$\begin{aligned} P[|N_n(t) - N(t)| \ge \epsilon] &\le P[N_n(t) \ne N(t)] \\ &\le P\left[\bigcup_{j=1}^{n}\left[\left(N\left(\frac{jt}{n}\right) - N\left(\frac{(j-1)t}{n}\right)\right) > 1\right]\right] \\ &\le \sum_{j=1}^{n} P\left[\left(N\left(\frac{jt}{n}\right) - N\left(\frac{(j-1)t}{n}\right)\right) > 1\right] \\ &= nP\left[N\left(\frac{t}{n}\right) > 1\right] \\ &= no\left(\frac{t}{n}\right) \to 0 \text{ as } n \to \infty. \end{aligned} \tag{A.16.3}$$

The first of the inequalities in (A.16.3) is obvious, the second says that if $N_n(t) \neq N(t)$ there must have been a multiple occurrence in a small subinterval, the third is just (A.2.5), and the remaining identities follow from (b) and (d). The claim (A.16.3) now follows from Slutsky's theorem (A.14.9) upon writing $N(t) = N_n(t) + (N(t) - N_n(t))$.

A.16.4 Let T_1 be the time at which the "event" first occurs in a Poisson process (the first t such that $N(t) = 1$), T_2 be the time at which the "event" occurs for the second time, and so on. Then $T_1, T_2 - T_1, \dots, T_n - T_{n-1}, \dots$ are independent, identically distributed $\mathcal{E}(\lambda)$ random variables.

References

Gnedenko (1967) Chapter 10, Section 51
Grimmett and Stirzaker (1992) Section 6.8
Hoel, Port, and Stone (1971) Section 9.3
Parzen (1962) Chapter 6, Section 5
Pitman (1993) Sections 3.5, 4.2

A.17 NOTES

Notes for Section A.5

(1) We define $\mathcal{A}$ to be the smallest sigma field that has every set of the form $A_1 \times \cdots \times A_n$ with $A_i \in \mathcal{A}_i$, $1 \leq i \leq n$, as a member.

Notes for Section A.7

(1) Strictly speaking, the density is only defined up to a set of Lebesgue measure 0.

(2) We shall use the notation $g(x+0)$ for $\lim_{x_n \downarrow x} g(x_n)$ and $g(x-0)$ for $\lim_{x_n \uparrow x} g(x_n)$ for a function g of a real variable that possesses such limits.

Notes for Section A.8

(1) The requirement on the sets $X^{-1}(B)$ is purely technical. It is no restriction in the discrete case and is satisfied by any function of interest when Ω is R^k or a subset of R^k. Sets B that are members of $\mathcal{B}^k$ are called *measurable*. When considering subsets of R^k, we will assume automatically that they are measurable.

(2) Such functions $\mathbf{g}$ are called *measurable*. This condition ensures that $\mathbf{g}(\mathbf{X})$ satisfies definitions (A.8.1) and (A.8.2). For convenience, when we refer to functions we shall assume automatically that this condition is satisfied.

(3) A function $\mathbf{g}$ is said to be one to one if $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{y})$ implies $\mathbf{x} = \mathbf{y}$.

(4) Strictly speaking, (X, Y) and (x, y) in (A.8.11) and (A.8.12) should be transposed. However, we avoid this awkward notation when the meaning is clear.

(5) The integral in (A.8.12) may only be finite for "almost all" x. In the regular cases we study this will not be a problem.

Notes for Section A.14

(1) It may be shown that one only needs the existence of the derivative g' at b for (A.14.17) to hold. See Theorem 5.3.3.

A.18 REFERENCES

BERGER, J. O., *Statistical Decision Theory and Bayesian Analysis* New York: Springer, 1985.

BILLINGSLEY, P., *Probability and Measure*, 3rd ed. New York: J. Wiley & Sons, 1995.

CHUNG, K. L., *A Course in Probability Theory* New York: Academic Press, 1974.

DEGROOT, M. H., *Optimal Statistical Decisions* New York: McGraw Hill, 1970.

FELLER, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed. New York: J. Wiley & Sons, 1968.

GNEDENKO, B. V., *The Theory of Probability*, 4th ed. New York: Chelsea, 1967.

GRIMMETT, G. R., AND D. R. STIRZAKER, *Probability and Random Processes* Oxford: Clarendon Press, 1992.

HÁJEK, J. AND Z. SIDÁK, *Theory of Rank Tests* New York: Academic Press, 1967.

HOEL, P. G., S. C. PORT, AND C. J. STONE, *Introduction to Probability Theory* Boston: Houghton Mifflin, 1971.

KARLIN, S., *A First Course in Stochastic Processes* New York: Academic Press, 1969.

LINDLEY, D. V., *Introduction to Probability and Statistics from a Bayesian Point of View*, Part I: *Probability*; Part II: *Inference* London: Cambridge University Press, 1965.

LOÉVE, M., *Probability Theory*, Vol. I, 4th ed. Berlin: Springer, 1977.

PARZEN, E., *Modern Probability Theory and Its Application* New York: J. Wiley & Sons, 1960.

PARZEN, E., *Stochastic Processes* San Francisco: Holden–Day, 1962.

PITMAN, J., *Probability* New York: Springer, 1993.

RAIFFA, H., AND R. SCHLAIFFER, *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Boston: Harvard University, 1961.

RAO, C. R., *Linear Statistical Inference and Its Applications*, 2nd ed. New York: J. Wiley & Sons, 1973.

SAVAGE, L. J., *The Foundations of Statistics* New York: J. Wiley & Sons, 1954.

SAVAGE, L. J., *The Foundation of Statistical Inference* London: Methuen & Co., 1962.

This page intentionally left blank