Chapter 2
The Probability Background

2.1 Probability and Measure

The mathematical framework for statistical decision theory is provided by the theory of probability, which in turn has its foundations in the theory of measure and integration. The present chapter serves to define some of the basic concepts of these theories, to establish some notation, and to state without proof some of the principal results which will be used throughout Chapters 3–9. In the remainder of this chapter, certain special topics are treated in more detail. Basic notions of convergence in probability theory which will be needed for large-sample statistical theory are deferred to Section 11.2.

Probability theory is concerned with situations which may result in different outcomes. The totality of these possible outcomes is represented abstractly by the totality of points in a space $\mathcal{Z}$. Since the events to be studied are aggregates of such outcomes, they are represented by subsets of $\mathcal{Z}$. The union of two sets C_1, C_2 will be denoted by $C_1 \cup C_2$, their intersection by $C_1 \cap C_2$, the complement of C by $C^c = \mathcal{Z} - C$, and the empty set by 0. The probability $P(C)$ of an event C is a real number between 0 and 1, in particular,

$$P(0) = 0 \quad \text{and} \quad P(\mathcal{Z}) = 1. \tag{2.1}$$

Probabilities have the property of *countable additivity*,

$$P\left(\bigcup C_i\right) = \sum P(C_i) \qquad \text{if} \quad C_i \cap C_j = 0 \quad \text{for all} \quad i \neq j. \tag{2.2}$$

Unfortunately it turns out that the set functions with which we shall be concerned usually cannot be defined in a reasonable manner for all subsets of $\mathcal{Z}$ if they are to satisfy (2.2). It is, for example, not possible to give a reasonable definition of "area" for all subsets of a unit square in the plane.

The sets for which the probability function P will be defined are said to be "measurable". The domain of definition of P should include with any set C its complement C^c, and with any countable number of events their union. By (2.1), it should also

E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics, https://doi.org/10.1007/978-3-030-70578-7_2

include $\mathcal{Z}$. A class of sets that contains $\mathcal{Z}$ and is closed under complementation and countable unions is a σ-field. Such a class is automatically also closed under countable intersections.

The starting point of any probabilistic considerations is therefore a space $\mathcal{Z}$, representing the possible outcomes, and a σ-field $\mathcal{C}$ of subsets of $\mathcal{Z}$, representing the events whose probability is to be defined. Such a couple $(\mathcal{Z}, \mathcal{C})$ is called a *measurable space*, and the elements of $\mathcal{C}$ constitute the *measurable sets*. A countably additive nonnegative (not necessarily finite) set function μ defined over $\mathcal{C}$ and such that $\mu(0) = 0$ is called a *measure*. If it assigns the value 1 to $\mathcal{Z}$, it is a *probability measure*. More generally, μ is *finite* if $\mu(\mathcal{Z}) < \infty$ and σ-finite if there exist $C_1, C_2, \ldots$ in $\mathcal{C}$ (which may always be taken to be mutually exclusive) such that $\cup C_i = \mathcal{Z}$ and $\mu(C_i) < \infty$ for $i = 1, 2, \ldots$. Important special cases are provided by the following examples.

Example 2.1.1 (**Lebesgue measure**) Let $\mathcal{Z}$ be the n-dimensional Euclidean space E_n, and $\mathcal{C}$ the smallest σ-field containing all rectangles[1]

$$R = \{(z_1, \ldots, z_n) : a_i < z_i \leq b_i, i = 1, \ldots, n\}.$$

The elements of $\mathcal{C}$ are called the *Borel sets* of E_n. Over $\mathcal{C}$ a unique measure μ can be defined, which to any rectangle R assigns as its measure the volume of R,

$$\mu(R) = \prod_{i=1}^{n} (b_i - a_i).$$

The measure μ can be *completed* by adjoining to $\mathcal{C}$ all subsets of sets of measure zero. The domain of μ is thereby enlarged to a σ-field $\mathcal{C}'$, the class of *Lebesgue-measurable* sets. The term *Lebesgue measure* is used for μ both when it is defined over the Borel sets and when it is defined over the Lebesgue-measurable sets. ■

This example can be generalized to any nonnegative set function ν, which is defined and countably additive over the class of rectangles R. There exists then, as before, a unique measure μ over $(\mathcal{Z}, \mathcal{C})$ that agrees with ν for all R. This measure can again be completed; however, the resulting σ-field depends on μ and need not agree with the σ-field $\mathcal{C}'$ obtained above.

Example 2.1.2 (**Counting measure**) Suppose the $\mathcal{Z}$ is countable, and let $\mathcal{C}$ be the class of all subsets of $\mathcal{Z}$. For any set C, define $\mu(C)$ as the number of elements of C if that number is finite, and otherwise as $+\infty$. This measure is sometimes called *counting measure*. ■

[1] If $\pi(z)$ is a statement concerning certain objects z, then $\{z : \pi(z)\}$ denotes the set of all those z for which $\pi(z)$ is true.

In applications, the probabilities over $(\mathcal{Z}, \mathcal{C})$ refer to random experiments or observations, the possible outcomes of which are the points $z \in \mathcal{Z}$. When recording the results of an experiment, one is usually interested only in certain of its aspects, typically some counts or measurements. These may be represented by a function T taking values in some space $\mathcal{T}$.

Such a function generates in $\mathcal{T}$ the σ-field $\mathcal{B}'$ of sets B whose inverse image

$$C = T^{-1}(B) = \{z : z \in \mathcal{Z}, T(z) \in B\}$$

is in $\mathcal{C}$, and for any given probability measure P over $(\mathcal{Z}, \mathcal{C})$ a probability measure Q over $(\mathcal{T}, \mathcal{B}')$ defined by

$$Q(B) = P(T^{-1}(B)). \tag{2.3}$$

Frequently, there is given a σ-field $\mathcal{B}$ of sets in $\mathcal{T}$ such that the probability of B should be defined if and only if $B \in \mathcal{B}$. This requires that $T^{-1}(B) \in \mathcal{C}$ for all $B \in \mathcal{B}$, and the function (or transformation) T from $(\mathcal{Z}, \mathcal{C})$ into[2] $(\mathcal{T}, \mathcal{B})$ is then said to be $\mathcal{C}$-measurable. Another implication is the sometimes convenient restriction of probability statements to the sets $B \in \mathcal{B}$ even though there may exist sets $B \notin \mathcal{B}$ for which $T^{-1}(B) \in \mathcal{C}$ and whose probability therefore could be defined.

Of particular interest is the case of a single measurement in which the function of T is real-valued. Let us denote it by X, and let $\mathcal{A}$ be the class of Borel sets on the real line $\mathcal{X}$. Such a measurable real-valued X is called a *random variable*, and the probability measure it generates over $(\mathcal{X}, \mathcal{A})$ will be denoted by P^X and called the probability distribution of X. The value this measure assigns to a set $A \in \mathcal{A}$ will be denoted interchangeably by $P^X(A)$ and $P(X \in A)$. Since the intervals $\{x : x \leq a\}$ are in $\mathcal{A}$, the probabilities $F(a) = P(X \leq a)$ are defined for all a. The function F, the *cumulative distribution function* (cdf) of X, is nondecreasing and continuous on the right, and $F(-\infty) = 0$, $F(+\infty) = 1$. Conversely, if F is any function with these properties, a measure can be defined over the intervals by $P\{a < X \leq b\} = F(b) - F(a)$. It follows from Example 2.1.1 that this measure uniquely determines a probability distribution over the Borel sets. Thus the probability distribution P^X and the cumulative distribution function F uniquely determine each other. These remarks extend to probability distributions over n-dimensional Euclidean space, where the cumulative distribution function is defined by

$$F(a_1, \ldots, a_n) = P\{X_1 \leq a_1, \ldots, X_n \leq a_n\}.$$

In concrete problems, the space $(\mathcal{Z}, \mathcal{C})$, corresponding to the totality of possible outcomes, is usually not specified and remains in the background. The real starting point is the set X of observations (typically vector-valued) that are being recorded and which constitute the *data*, and the associated measurable space $(\mathcal{X}, \mathcal{A})$, the *sample space*. Random variables or vectors that are measurable transformations T

[2] The term *into* indicates that the range of T is in $\mathcal{T}$; if $T(\mathcal{Z}) = \mathcal{T}$, the transformation is said to be from $\mathcal{Z}$ *onto* $\mathcal{T}$.

from $(\mathcal{X}, \mathcal{A})$ into some $(\mathcal{T}, \mathcal{B})$ are called *statistics*. The distribution of T is then given by (2.3) applied to all $B \in \mathcal{B}$. With this definition, a statistic is specified by the function T and the σ-field $\mathcal{B}$. We shall, however, adopt the convention that when a function T takes on its values in a Euclidean space, unless otherwise stated the σ-field $\mathcal{B}$ of measurable sets will be taken to be the class of Borel sets. It then becomes unnecessary to mention it explicitly or to indicate it in the notation.

The distinction between statistics and random variables as defined here is slight. The term statistic is used to indicate that the quantity is a function of more basic observations; all statistics in a given problem are functions defined over the same sample space $(\mathcal{X}, \mathcal{A})$. On the other hand, any real-valued statistic T is a random variable, since it has a distribution over $(\mathcal{T}, \mathcal{B})$, and it will be referred to as a random variable when its origin is irrelevant. Which term is used therefore depends on the point of view and to some extent is arbitrary.

2.2 Integration

According to the convention of the preceding section, a real-valued function f defined over $(\mathcal{X}, \mathcal{A})$ is measurable if $f^{-1}(B) \in \mathcal{A}$ for every Borel set B on the real line. Such a function f is said to be *simple* if it takes on only a finite number of values. Let μ be a measure defined over $(\mathcal{X}, \mathcal{A})$, and let f be a simple function taking on the distinct values $a_1, \ldots, a_m$ on the sets $A_1, \ldots, A_m$, which are in $\mathcal{A}$, since f is measurable. If $\mu(A_i) < \infty$ when $a_i \neq 0$, the integral of f with respect to μ is defined by

$$\int f \, d\mu = \sum a_i \mu(A_i). \tag{2.4}$$

Given any nonnegative measurable function f, there exists a nondecreasing sequence of simple functions f_n converging to f. Then the integral of f is defined as

$$\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu, \tag{2.5}$$

which can be shown to be independent of the particular sequence of f_n's chosen. For any measurable function f its positive and negative parts

$$f^+(x) = \max[f(x), 0] \quad \text{and} \quad f^-(x) = \max[-f(x), 0] \tag{2.6}$$

are also measurable, and

$$f(x) = f^+(x) - f^-(x).$$

If the integrals of f^+ and f^- are both finite, then f is said to be *integrable*, and its integral is defined as

$$\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu.$$

If of the two integrals one is finite and one infinite, then the integral of f is defined to be the appropriate infinite value; if both are infinite, the integral is not defined.

Example 2.2.1 Let $\mathcal{X}$ be the closed interval $[a, b]$, $\mathcal{A}$ be the class of Borel sets or of Lebesgue measurable sets in $\mathcal{X}$, and μ be Lebesgue measure. Then the integral of f with respect to μ is written as $\int_a^b f(x)\,dx$, and is called the Lebesgue integral of f. This integral generalizes the Riemann integral in that it exists and agrees with the Riemann integral of f whenever the latter exists. ■

Example 2.2.2 Let $\mathcal{X}$ be countable and consist of the points $x_1, x_2, \ldots$; let $\mathcal{A}$ be the class of all subsets of $\mathcal{X}$, and let μ assign measure b_i to the point x_i. Then f is integrable provided $\sum f(x_i)b_i$ converges absolutely, and $\int f\,d\mu$ is given by this sum. ■

Let P^X be the probability distribution of a random variable X, and let T be a real-valued statistic. If the function $T(x)$ is integrable, its *expectation* is defined by

$$E(T) = \int T(x)\,dP^X(x). \tag{2.7}$$

It will be seen from Lemma 2.3.2 in Section 2.3 that the integration can be carried out alternatively in t-space with respect to the distribution of T defined by (2.3), so that also

$$E(T) = \int t\,dP^T(t). \tag{2.8}$$

Definition (2.5) of the integral permits the basic convergence theorems.

Theorem 2.2.1 Fatou's Lemma *Let f_n be a sequence of measurable functions such that $f_n(x) \geq 0$ and $f_n(x) \to f(x)$, except possibly on a set of x values having μ measure 0. Then,*

$$\int f\,d\mu \leq \liminf \int f_n d\mu\,.$$

Theorem 2.2.2 *Let f_n be a sequence of measurable functions, and let $f_n(x) \to f(x)$, except possibly on a set of x values having μ measure 0. Then*

$$\int f_n\,d\mu \to \int f\,d\mu$$

if any one of the following conditions holds:

(i) **Lebesgue Monotone Convergence Theorem:** *the f_n's are nonnegative and the sequence is nondecreasing;*

or

(ii) **Lebesgue Dominated Convergence Theorem:** *there exists an integrable function g such that $|f_n(x)| \leq g(x)$ for n and x.*

or

(iii) **General Form:** *there exist g_n and g with $|f_n| \leq g_n$, $g_n(x) \to g(x)$ except possibly on a μ null set, and $\int g_n d\mu \to \int g d\mu$.*

Corollary 2.2.1 Vitali's Theorem *Suppose f_n and f are real-valued measurable functions with $f_n(x) \to f(x)$, except possibly on a set having μ measure 0. Assume*

$$\limsup_n \int f_n^2(x) d\mu(x) \leq \int f^2(x) d\mu(x) < \infty .$$

Then,

$$\int |f_n(x) - f(x)|^2 d\mu(x) \to 0 \ .$$

For a proof of this result, see Theorem 6.1.3 of Hájek et al. (1999).

For any set $A \in \mathcal{A}$, let I_A be its *indicator function* defined by

$$I_A(x) = 1 \text{ or } 0 \quad as \quad x \in A \text{ or } x \in A^c, \tag{2.9}$$

and let

$$\int_A f \, d\mu = \int f I_A \, d\mu. \tag{2.10}$$

If μ is a measure and f a nonnegative measurable function over $(\mathcal{X}, \mathcal{A})$, then

$$\nu(A) = \int_A f \, d\mu \tag{2.11}$$

defines a new measure over $(\mathcal{X}, \mathcal{A})$. The fact that (2.11) holds for all $A \in \mathcal{A}$ is expressed by writing

$$d\nu = f \, d\mu \quad \text{or} \quad f = \frac{d\nu}{d\mu}. \tag{2.12}$$

Let μ and ν be two given σ-finite measures over $(\mathcal{X}, \mathcal{A})$. If there exists a function f satisfying (2.12), it is determined through this relation up to sets of measure zero, since

$$\int_A f \, d\mu = \int_A g \, d\mu \qquad \text{for all} \quad A \in \mathcal{A}$$

implies that $f = g$ a.e. μ.[3] Such an f is called the *Radon–Nikodym derivative* of ν with respect to μ, and in the particular case that ν is a probability measure, the *probability density* of ν with respect to μ.

The question of existence of a function f satisfying (2.12) for given measures μ and ν is answered in terms of the following definition. A measure ν is *absolutely continuous* with respect to μ if

$$\mu(A) = 0 \quad \text{implies} \quad \nu(A) = 0.$$

Theorem 2.2.3 (Radon–Nikodym) *If μ and ν are σ-finite measures over $(\mathcal{X}, \mathcal{A})$, then there exists a measurable function f satisfying (2.12) if and only if ν is absolutely continuous with respect to μ.*

The *direct* (or *Cartesian*) *product* $A \times B$ of two sets A and B is the set of all pairs (x, y) with $x \in A$, $y \in B$. Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be two measurable spaces, and let $\mathcal{A} \times \mathcal{B}$ be the smallest σ-field containing all sets $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. If μ and ν are two σ-finite measures over $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, respectively, then there exists a unique measure $\lambda = \mu \times \nu$ over $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B})$, the *product* of μ and ν, such that for any $A \in \mathcal{A}$, $B \in \mathcal{B}$,

$$\lambda(A \times B) = \mu(A)\nu(B). \tag{2.13}$$

Example 2.2.3 Let $\mathcal{X}, \mathcal{Y}$ be Euclidean spaces of m and n dimensions, and let $\mathcal{A}, \mathcal{B}$ be the σ-fields of Borel sets in these spaces. Then $\mathcal{X} \times \mathcal{Y}$ is an $(m + n)$-dimensional Euclidean space, and $\mathcal{A} \times \mathcal{B}$ the class of its Borel sets. ■

Example 2.2.4 Let $Z = (X, Y)$ be a random variable defined over $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B})$, and suppose that the random variables X and Y have distributions P^X, P^Y over $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$. Then X and Y are said to be *independent* if the probability distribution P^Z of Z is the product $P^X \times P^Y$. ■

In terms of these concepts the reduction of a double integral to a repeated one is given by the following theorem.

Theorem 2.2.4 (Fubini) *Let μ and ν be σ-finite measures over $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ respectively, and let $\lambda = \mu \times \nu$. If $f(x, y)$ is integrable with respect to λ, then*

(i) *for almost all (ν) fixed y, the function $f(x, y)$ is integrable with respect to μ,*
(ii) *the function $\int f(x, y)\, d\mu(x)$ is integrable with respect to ν, and*

$$\int f(x, y)\, d\lambda(x, y) = \int \left[\int f(x, y)\, d\mu(x) \right] d\nu(y). \tag{2.14}$$

[3] A statement that holds for all points x except possibly on a set of μ-measure zero is said to hold *almost everywhere* μ, abbreviated a.e. μ, or to hold a.e. $(\mathcal{A}, \mu)$ if it is desirable to indicate the σ-field over which μ is defined.

2.3 Statistics and Subfields

According to the definition of Section 2.1, a statistic is a measurable transformation T from the sample space $(\mathcal{X}, \mathcal{A})$ into a measurable space $(\mathcal{T}, \mathcal{B})$. Such a transformation induces in the original sample space the subfield[4]

$$\mathcal{A}_0 = T^{-1}(\mathcal{B}) = \left\{T^{-1}(B) : B \in \mathcal{B}\right\}. \tag{2.15}$$

Since the set $T^{-1}[T(A)]$ contains A but is not necessarily equal to A, the σ-field $\mathcal{A}_0$ need not coincide with $\mathcal{A}$ and hence can be a proper subfield of $\mathcal{A}$. On the other hand, suppose for a moment that $\mathcal{T} = T(\mathcal{X})$, that is, that the transformation T is onto rather than into $\mathcal{T}$. Then

$$T\left[T^{-1}(B)\right] = B \qquad \text{for all} \qquad B \in \mathcal{B}, \tag{2.16}$$

so that the relationship $A_0 = T^{-1}(B)$ establishes a 1:1 correspondence between the sets of $\mathcal{A}_0$ and $\mathcal{B}$, which is an isomorphism—that is, which preserves the set operations of intersection, union, and complementation. For most purposes it is therefore immaterial whether one works in the space $(\mathcal{X}, \mathcal{A}_0)$ or in $(\mathcal{T}, \mathcal{B})$. These generate two equivalent classes of events, and therefore of measurable functions, possible decision procedures, etc. If the transformation T is only into $\mathcal{T}$, the above 1:1 correspondence applies to the class $\mathcal{B}'$ of subsets of $\mathcal{T}' = T(\mathcal{X})$ which belong to $\mathcal{B}$, rather than to $\mathcal{B}$ itself. However, any set $B \in \mathcal{B}$ is equivalent to $B' = B \cap \mathcal{T}'$ in the sense that any measure over $(\mathcal{X}, \mathcal{A})$ assigns the same measure to B' as to B. Considered as classes of events, $\mathcal{A}_0$ and $\mathcal{B}$ therefore continue to be equivalent, with the only difference that $\mathcal{B}$ contains several (equivalent) representations of the same event.

As an example, let $\mathcal{X}$ be the real line and $\mathcal{A}$ the class of Borel sets, and let $T(x) = x^2$. Let $\mathcal{T}$ be either the positive real axis or the whole real axis, and let $\mathcal{B}$ be the class of Borel subsets of $\mathcal{T}$. Then $\mathcal{A}_0$ is the class of Borel sets that are symmetric with respect to the origin. When considering, for example, real-valued measurable functions, one would, when working in $\mathcal{T}$-space, restrict attention to measurable function of x^2. Instead, one could remain in the original space, where the restriction would be to the class of even measurable functions of x. The equivalence is clear. Which representation is more convenient depends on the situation.

That the correspondence between the sets $A_0 = T^{-1}(B) \in \mathcal{A}_0$ and $B \in \mathcal{B}$ establishes an analogous correspondence between measurable functions defined over $(\mathcal{X}, \mathcal{A}_0)$ and $(\mathcal{T}, \mathcal{B})$ is shown by the following lemma.

[4] We shall use this term in place of the more cumbersome "sub-σ-field".

Lemma 2.3.1 *Let the statistic T from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{T}, \mathcal{B})$ induce the subfield $\mathcal{A}_0$. Then a real-valued $\mathcal{A}$-measurable function f is $\mathcal{A}_0$-measurable if and only if there exists a $\mathcal{B}$-measurable function g such that*

$$f(x) = g[T(x)]$$

for all x.

PROOF. Suppose first that such a function g exists. Then the set

$$\{x : f(x) < r\} = T^{-1}(\{t : g(t) < r\})$$

is in $\mathcal{A}_0$, and f is $\mathcal{A}_0$-measurable. Conversely, if f is $\mathcal{A}_0$-measurable, then the sets

$$A_{in} = \left\{x : \frac{i}{2^n} < f(x) \leq \frac{i+1}{2^n}\right\}, \qquad i = 0, \pm 1, \pm 2, \ldots$$

are (for fixed n) disjoint sets in $\mathcal{A}_0$ whose union is $\mathcal{X}$, and there exist $B_{in} \in \mathcal{B}$ such that $A_{in} = T^{-1}(B_{in})$. Let

$$B_{in}^* = B_{in} \cap \{\bigcup_{j \neq i} B_{jn}\}^c .$$

Since A_{in} and A_{jn} are mutually exclusive for $i \neq j$, the set $T^{-1}(B_{in} \cap B_{jn})$ is empty and so is the set $T^{-1}(B_{in} \cap \{B_{in}^*\}^c)$. Hence, for fixed n, the sets B_{in}^* are disjoint, and still satisfy $A_{in} = T^{-1}(B_{in}^*)$. Defining

$$f_n(x) = \frac{i}{2^n} \quad \text{if} \quad x \in A_{in}, \qquad i = 0 \pm 1, \pm 2, \ldots,$$

one can write

$$f_n(x) = g_n[T(x)],$$

where

$$g_n(t) = \begin{cases} \frac{i}{2^n} & \text{for } t \in B_{in}^*, \quad i = 0 \pm 1, \pm 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Since the functions g_n are $\mathcal{B}$-measurable, the set B on which $g_n(t)$ converges to a finite limit is in $\mathcal{B}$. Let $R = T(\mathcal{X})$ be the range of T. Then for $t \in R$,

$$\lim g_n[T(x)] = \lim f_n(x) = f(x)$$

for all $x \in \mathcal{X}$ so that R is contained in B. Therefore, the function g defined by $g(t) = \lim g_n(t)$ for $t \in B$ and $g(t) = 0$ otherwise possesses the required properties. ■

The relationship between integrals of the functions f and g above is given by the following lemma.

Lemma 2.3.2 *Let T be a measurable transformation from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{T}, \mathcal{B})$, μ a σ-finite measure over $(\mathcal{X}, \mathcal{A})$, and g a real-valued measurable function of t. If μ^* is the measure defined over $(\mathcal{T}, \mathcal{B})$ by*

$$\mu^*(B) = \mu\left[T^{-1}(B)\right] \quad \text{for all} \quad B \in \mathcal{B}, \tag{2.17}$$

then for any $B \in \mathcal{B}$,

$$\int_{T^{-1}(B)} g[T(x)]\, d\mu(x) = \int_B g(t)\, d\mu^*(t) \tag{2.18}$$

in the sense that if either integral exists, so does the other and the two are equal.

PROOF. Without loss of generality let B be the whole space $\mathcal{T}$. If g is the indicator of a set $B_0 \in \mathcal{B}$, the lemma holds, since the left- and right-hand sides of (2.18) reduce respectively to $\mu[T^{-1}(B_0)]$ and $\mu^*(B_0)$, which are equal by the definition of μ^*. If follows that (2.18) holds successively for all simple functions, for all nonnegative measurable functions, and hence finally for all integrable functions. ■

2.4 Conditional Expectation and Probability

If two statistics induce the same subfield $\mathcal{A}_0$, they are equivalent in the sense of leading to equivalent classes of measurable events. This equivalence is particularly relevant to considerations of conditional probability. Thus if X is normally distributed with zero mean, the information carried by the statistics $|X|$, X^2, e^{-X^2}, and so on, is the same. Given that $|X| = t$, $X^2 = t^2$, $e^{-X^2} = e^{-t^2}$, it follows that X is $\pm t$, and any reasonable definition of conditional probability will assign probability $\frac{1}{2}$ to each of these values. The general definition of conditional probability to be given below will in fact involve essentially only $\mathcal{A}_0$ and not the range space $\mathcal{T}$ of T. However, when referred to $\mathcal{A}_0$ alone the concept loses much of its intuitive meaning, and the gap between the elementary definition and that of the general case becomes unnecessarily wide. For these reasons it is frequently more convenient to work with a particular representation of a statistic, involving a definite range space $(\mathcal{T}, \mathcal{B})$.

Let P be a probability measure over $(\mathcal{X}, \mathcal{A})$, T a statistic with range space $(\mathcal{T}, \mathcal{B})$, and $\mathcal{A}_0$ the subfield it induces. Consider a nonnegative function f which is integrable $(\mathcal{A}, P)$, that is, $\mathcal{A}$-measurable and P-integrable. Then $\int_A f\, dP$ is defined for all $A \in \mathcal{A}$ and therefore for all $A_0 \in \mathcal{A}_0$. If follows from the Radon–Nikodym Theorem (Theorem 2.2.3) that there exists a function f_0 which is integrable $(\mathcal{A}_0, P)$ and such that

$$\int_{A_0} f\, dP = \int_{A_0} f_0\, dP \qquad \text{for all} \quad A_0 \in \mathcal{A}_0, \tag{2.19}$$

and that f_0 is unique $(\mathcal{A}_0, P)$. By Lemma 2.3.1, f_0 depends on x only through $T(x)$. In the example of a normally distributed variable X with zero mean, and $T = X^2$, the function f_0 is determined by (2.19) holding for all sets A_0 that are symmetric with respect to the origin, so that $f_0(x) = \frac{1}{2}[f(x) + f(-x)]$.

The function f_0 defined through (2.19) is determined by two properties:

(i) Its average value over any set A_0 with respect to P is the same as that of f;
(ii) It depends on x only through $T(x)$ and hence is constant on the sets D_x over which T is constant.

Intuitively, what one attempts to do in order to construct such a function is to define $f_0(x)$ as the conditional P-average of f over the set D_x. One would thereby replace the single averaging process of integrating f represented by the left-hand side with a two-stage averaging process such as an iterated integral. Such a construction can actually be carried out when X is a discrete variable and in the regular case considered in Section 1.9; $f_0(x)$ is then just the conditional expectation of $f(X)$ given $T(x)$. In general, it is not clear how to define this conditional expectation directly. Since it should, however, possess properties (i) and (ii), and since these through (2.19) determine f_0 uniquely $(\mathcal{A}_0, P)$, we shall take $f_0(x)$ of (2.19) as the general definition of the *conditional expectation* $E[f(X) \mid T(x)]$. Equivalently, if $f_0(x) = g[T(x)]$, one can write

$$E[f(X) \mid t] = E[f(X) \mid T = t] = g(t),$$

so that $E[f(X) \mid t]$ is a $\mathcal{B}$-measurable function defined up to equivalence $(\mathcal{B}, P^T)$. In the relationship of integrals given in Lemma 2.3.2, if $\mu = P^X$, then $\mu^* = P^T$, and it is seen that the function g can be defined directly in terms of f through

$$\int_{T^{-1}(B)} f(x)\, dP^X(x) = \int_B g(t)\, dP^T(t) \qquad \text{for all} \quad B \in \mathcal{B}, \tag{2.20}$$

which is equivalent to (2.19).

So far, f has been assumed to be nonnegative. In the general case, the conditional expectation of f is defined as

$$E[f(X) \mid t] = E[f^+(X) \mid t] - E[f^-(X) \mid t].$$

Example 2.4.1 (**Order statistics**) Let $X_1, \ldots, X_n$ be identically and independently distributed random variables with continuous distribution function, and let

$$T(x_1, \ldots, x_n) = (x_{(1)}, \ldots, x_{(n)}),$$

where $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the ordered x's. Without loss of generality one can restrict attention to the points with $x_{(1)} < \cdots < x_{(n)}$, since the probability of two coordinates being equal is 0. Then $\mathcal{X}$ is the set of all n-tuples with distinct coordinates, $\mathcal{T}$ the set of all ordered n-tuples, and $\mathcal{A}$ and $\mathcal{B}$ are the classes of Borel subsets of $\mathcal{X}$ and

$\mathcal{T}$. Under T^{-1} the set consisting of the single point $a = (a_1, \ldots, a_n)$ is transformed into the set consisting of the $n!$ points $(a_{i_1}, \ldots, a_{i_n})$ that are obtained from a by permuting the coordinates in all possible ways. It follows that $\mathcal{A}_0$ is the class of all sets that are symmetric in the sense that if A_0 contains a point $x = (x_1, \ldots, x_n)$, then it also contains all points $(x_{i_1}, \ldots, x_{i_n})$.

For any integrable function f, let

$$f_0(x) = \frac{1}{n!} \sum f(x_{i_1}, \ldots, x_{i_n}),$$

where the summation extends over the $n!$ permutations of $(x_1, \ldots, x_n)$. Then f_0 is $\mathcal{A}_0$-measurable, since it is symmetric in its n arguments. Also

$$\int_{A_0} f(x_1, \ldots, x_n)\, dP(x_1) \ldots dP(x_n) = \int_{A_0} f(x_{i_1}, \ldots, x_{i_n})\, dP(x_1) \ldots dP(x_n),$$

so that f_0 satisfies (2.19). It follows that $f_0(x)$ is the conditional expectation of $f(X)$ given $T(x)$.

The conditional expectation of $f(X)$ given the above statistic $T(x)$ can also be found without assuming the X's to be identically and independently distributed. Suppose that X has a density $h(x)$ with respect to a measure μ (such as Lebesgue measure), which is symmetric in the variables $x_1, \ldots, x_n$ in the sense that for any $A \in \mathcal{A}$ it assigns to the set $\{x : (x_{i_1}, \ldots, x_{i_n}) \in A\}$ the same measure for all permutations $(i_1, \ldots, i_n)$. Let

$$f_0(x_1, \ldots, x_n) = \frac{\sum f(x_{i_1}, \ldots, x_{i_n}) h(x_{i_1}, \ldots, x_{i_n})}{\sum h(x_{i_1}, \ldots, x_{i_n})};$$

here and in the sums below the summation extends over the $n!$ permutations of $(x_1, \ldots, x_n)$. The function f_0 is symmetric in its n arguments and hence $\mathcal{A}_0$-measurable. For any symmetric set A_0, the integral

$$\int_{A_0} f_0(x_1, \ldots, x_n) h(x_{j_1}, \ldots, x_{j_n})\, d\mu(x_1, \ldots, x_n)$$

has the same value for each permutation $(x_{j_1}, \ldots, x_{j_n})$, and therefore

$$\begin{aligned} &\int_{A_0} f_0(x_1, \ldots, x_n) h(x_1, \ldots, x_n)\, d\mu(x_1, \ldots, x_n) \\ &\quad = \int_{A_0} f_0(x_1, \ldots, x_n) \frac{1}{n!} \sum h(x_{i_1}, \ldots, x_{i_n})\, d\mu(x_1, \ldots, x_n) \\ &\quad = \int_{A_0} f(x_1, \ldots, x_n) h(x_1, \ldots, x_n)\, d\mu(x_1, \ldots, x_n). \end{aligned}$$

It follows that $f_0(x) = E[f(X) \mid T(x)]$.

Equivalent to the statistic $T(x) = (x_{(1)}, \ldots, x_{(n)})$, the set of *order statistics* is $U(x) = \left(\sum x_i, \sum x_i^2, \ldots, \sum x_i^n\right)$. This is an immediate consequence of the fact, to be shown below, that if $T(x^0) = t^0$ and $U(x^0) = u^0$, then

$$T^{-1}\left(\{t^0\}\right) = U^{-1}\left(\{u^0\}\right) = S,$$

where $\{t^0\}$ and $\{u^0\}$ denote the sets consisting of the single point t^0 and u^0, respectively, and where S consists of the totality of points $x = (x_1, \ldots, x_n)$ obtained by permuting the coordinates of $x^0 = (x_1^0, \ldots, x_n^0)$ in all possible ways.

That $T^{-1}\left(\{t^0\}\right) = S$ is obvious. To see the corresponding fact for U^{-1}, let

$$V(x) = \left(\sum_i x_i, \sum_{i<j} x_i x_j, \sum_{i<j<k} x_i x_j x_k, \ldots, x_1 x_2 \cdots x_n\right),$$

so that the components of $V(x)$ are the elementary symmetric functions $v_1 = \sum x_i, \ldots, v_n = x_1 \ldots x_n$ of the n arguments $x_1, \ldots, x_n$. Then

$$(x - x_1) \ldots (x - x_n) = x^n - v_1 x^{n-1} + v_2 x^{n-2} - \cdots + (-1)^n v_n.$$

Hence $V(x^0) = v^0 = (v_1^0, \ldots, v_n^0)$ implies that $V^{-1}(\{v^0\}) = S$. That then also $U^{-1}(\{u^0\}) = S$ follows from the 1:1 correspondence between u and v established by the relations (known as Newton's identities)[5]:

$$u_k - v_1 u_{k-1} + v_2 u_{k-2} - \cdots + (-1)^{k-1} v_{k-1} u_1 + (-1)^k k v_k = 0$$

for $1 \leq k \leq n$. ■

It is easily verified from the above definition that conditional expectation possesses most of the usual properties of expectation. It follows of course from the nonuniqueness of the definition that these properties can hold only $(\mathcal{B}, P^T)$. We state this formally in the following lemma.

Lemma 2.4.1 *If T is a statistic and the functions f, g, ... are integrable $(\mathcal{A}, P)$, then a.e. $(\mathcal{B}, P^T)$*

(i) $E[af(X) + bg(X) \mid t] = aE[f(X) \mid t] + bE[g(X) \mid t]$;
(ii) $E[h(T)f(X) \mid t] = h(t)E[f(X) \mid t]$;
(iii) $a \leq f(x) \leq b$ $(\mathcal{A}, P)$ *implies* $a \leq E[f(X) \mid t] \leq b$;
(iv) $|f_n| \leq g$, $f_n(x) \to f(x)$ $(\mathcal{A}, P)$ *implies* $E[f_n(X) \mid t] \to E[f(X) \mid t]$.

A further useful result is obtained by specializing (2.20) to the case that B is the whole space $\mathcal{T}$. One then has

[5] For a proof of these relations, see for example Turnbull (1952), Section 32.

Lemma 2.4.2 *If* $E[|f(X)|] < \infty$, *and if* $g(t) = E[f(X) \mid t]$, *then*

$$E[f(X)] = E[g(T)]\,, \tag{2.21}$$

that is, the expectation can be obtained as the expected value of the conditional expectation.

Since $P\{X \in A\} = E[I_A(X)]$, where I_A denotes the indicator of the set A, it is natural to define the *conditional probability* of A given $T = t$ by

$$P(A \mid t) = E[I_A(X) \mid t]. \tag{2.22}$$

In view of (2.20) the defining equation for $P(A \mid t)$ can therefore be written as

$$\begin{aligned} P^X\left(A \cap T^{-1}(B)\right) &= \int_{A \cap T^{-1}(B)} dP^X(x) \\ &= \int_B P(A \mid t)\, dP^T(t) \qquad \text{for all} \quad B \in \mathcal{B}. \end{aligned} \tag{2.23}$$

It is an immediate consequence of Lemma 2.4.1 that subject to the appropriate null-set[6] qualifications, $P(A \mid t)$ possesses the usual properties of probabilities, as summarized in the following lemma.

Lemma 2.4.3 *If* T *is a statistic with range space* $(\mathcal{T}, \mathcal{B})$, *and* $A, B, A_1, A_2, \ldots$ *are sets belonging to* $\mathcal{A}$, *then a.e.* $(\mathcal{B}, P^T)$

(i) $0 \le P(A \mid t) \le 1$;
(ii) *if the sets* $A_1, A_2, \ldots$ *are mutually exclusive,*

$$P\left(\bigcup A_i \mid t\right) = \sum P(A_i \mid t);$$

(iii) $A \subset B$ *implies* $P(A \mid t) \le P(B \mid t)$.

According to definition (2.22), the conditional probability $P(A \mid t)$ must be considered for fixed A as a $\mathcal{B}$-measurable function of t. This is in contrast to the elementary definition in which one takes t as fixed and considers $P(A \mid t)$ for varying A as a set function over $\mathcal{A}$. Lemma 2.4.3 suggests the possibility that the interpretation of $P(A \mid t)$ for fixed t as a probability distribution over $\mathcal{A}$ may be valid also in the general case. However, the equality $P(A_1 \cup A_2 \mid t) = P(A_1 \mid t) + P(A_2 \mid t)$, for example, can break down on a null set that may vary with A_1 and A_2, and the union of all these null sets need no longer have measure zero.

For an important class of cases, this difficulty can be overcome through the nonuniqueness of the functions $P(A \mid t)$, which for each fixed A are determined only up to sets of measure zero in t. Since all determinations of these functions are

[6] This term is used as an alternative to the more cumbersome "set of measure zero".

equivalent, it is enough to find a specific determination for each A so that for each fixed t these determinations jointly constitute a probability distribution over $\mathcal{A}$. This possibility is illustrated by Example 2.4.1, in which the conditional probability distribution given $T(x) = t$ can be taken to assign probability $1/n!$ to each of the $n!$ points satisfying $T(x) = t$. Sufficient conditions for the existence of such conditional distributions will be given in the next section. For counterexamples see Blackwell and Dubins (1975).

2.5 Conditional Probability Distributions

We shall now investigate the existence[7] of conditional probability distributions under the assumption, satisfied in most statistical applications, that $\mathcal{X}$ is a Borel set in a Euclidean space. We shall then say for short that $\mathcal{X}$ is Euclidean and assume that, unless otherwise stated, $\mathcal{A}$ is the class of Borel subsets of $\mathcal{X}$.

Theorem 2.5.1 *If $\mathcal{X}$ is Euclidean, there exist determinations of the functions $P(A \mid t)$ such that for each t, $P(A \mid t)$ is a probability measure over $\mathcal{A}$.*

PROOF. By setting equal to 0 the probability of any Borel set in the complement of $\mathcal{X}$, one can extend the given probability measure to the class of all Borel sets and can therefore assume without loss of generality that $\mathcal{X}$ is the full Euclidean space. For simplicity we shall give the proof only in the one-dimensional case. For each real x put $F(x, t) = P((-\infty, x] \mid t)$ for some version of this conditional probability function, and let $r_1, r_2, \ldots$ denote the set of all rational numbers in some order. Then $r_i < r_j$ implies that $F(r_i, t) \leq F(r_j, t)$ for all t except those in a null set N_{ij}, and hence that $F(x, t)$ is nondecreasing in x over the rationals for all t outside of the null set $N' = \bigcup N_{ij}$. Similarly, it follows from Lemma 2.4.1(iv) that for all t not in a null set N'', as n tends to infinity $\lim F(r_i + 1/n, t) = F(r_i, t)$ for $i = 1, 2, \ldots$, $\lim F(n, t) = 1$, and $\lim F(-n, t) = 0$. Therefore, for all t outside of the null set $N' \cup N''$, $F(x, t)$ considered as a function of x is properly normalized, monotone, and continuous on the right over the rationals. For t not in $N' \cup N''$ let $F^*(x, t)$ be the unique function that is continuous on the right in x and agrees with $F(x, t)$ for all rational x. Then $F^*(x, t)$ is a cumulative distribution function and therefore determines a probability measure $P^*(A \mid t)$ over $\mathcal{A}$. We shall now show that $P^*(A \mid t)$ is a conditional probability of A given t, by showing that for each fixed A it is a $\mathcal{B}$-measurable function of t satisfying (2.23). This will be accomplished by proving that for each fixed $A \in \mathcal{A}$

$$P^*(A \mid t) = P(A \mid t) \qquad (\mathcal{B}, P^T).$$

[7] This section may be omitted at first reading. Its principal application is in the proof of Lemma 2.7.2(ii) in Section 2.7, which in turn is used only in the proof of Theorem 4.4.1.

By definition of P^* this is true whenever A is one of the sets $(-\infty, x]$ with x rational. It holds next when A is an interval $(a, b] = (-\infty, b] - (-\infty, a]$ with a, b rational, since P^* is a measure and P satisfies Lemma 2.4.3(ii). Therefore, the desired equation holds for the field $\mathcal{F}$ of all sets A which are finite unions of intervals $(a_i, b_i]$ with rational end points. Finally, the class of sets for which the equation holds is a monotone class (see Problem 2.1) and hence contains the smallest σ-field containing $\mathcal{F}$, which is $\mathcal{A}$. The measure $P^*(A \mid t)$ over $\mathcal{A}$ was defined above for all t not in $N' \cup N''$. However, since neither the measurability of a function nor the values of its integrals are affected by its values on a null set, one can take arbitrary probability measures over $\mathcal{A}$ for t in $N' \cup N''$ and thereby complete the determination.

If X is a vector-valued random variable with probability distribution P^X and T is a statistic defined over $(\mathcal{X}, \mathcal{A})$, let $P^{X|t}$ denote any version of the family of conditional distributions $P(A \mid t)$ over $\mathcal{A}$ guaranteed by Theorem 2.5.1. The connection with conditional expectation is given by the following theorem. ■

Theorem 2.5.2 *If X is a vector-valued random variable and $E|f(X)| < \infty$, then*

$$E[f(X) \mid t] = \int f(x)\, dP^{X|t}(x) \qquad (\mathcal{B}, P^T). \tag{2.24}$$

PROOF. Equation (2.24) holds if f is the indicator of any set $A \in \mathcal{A}$. It then follows from Lemma 2.4.1 that it also holds for any simple function and hence for any integrable function. ■

The determination of the conditional expectation $E[f(X) \mid t]$ given by the right-hand side of (2.24) possesses for each t the usual properties of an expectation, (i), (iii), and (iv) of Lemma 2.4.1, which previously could be asserted only up to sets of measure zero depending on the functions $f, g, \ldots$ involved. Under the assumptions of Theorem 2.5.1 a similar strengthening is possible with respect to (ii) of Lemma 2.4.1, which can be shown to hold except possibly on a null set N not depending on the function h. It will be sufficient for the present purpose to prove this under the additional assumption that the range space of the statistic T is also Euclidean. For a proof without this restriction, see for example Billingsley (1995).

Theorem 2.5.3 *If T is a statistic with Euclidean domain and range spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{T}, \mathcal{B})$, there exists a determination $P^{X|t}$ of the conditional probability distribution and a null set N such that the conditional expectation computed by*

$$E[f(X) \mid t] = \int f(x)\, dP^{X|t}(x)$$

satisfies for all $t \notin N$.

$$E[h(T)f(X) \mid t] = h(t)E[f(X) \mid t]. \tag{2.25}$$

PROOF. For the sake of simplicity and without essential loss of generality suppose that T is real-valued. Let $P^{X|t}(A)$ be a probability distribution over $\mathcal{A}$ for each t,

the existence of which is guaranteed by Theorem 2.5.1. For $B \in \mathcal{B}$, the indicator function $I_B(t)$ is $\mathcal{B}$-measurable and

$$\int_{B'} I_B(t)\, dP^T(t) = P^T(B' \cap B) = P^X(T^{-1}B' \cap T^{-1}B) \qquad \text{for all} \quad B' \in \mathcal{B}.$$

Thus by (2.20)

$$I_B(t) = P^{X|t}\left(T^{-1}B\right) \qquad \text{a.e. } P^T.$$

Let B_n, $n = 1, 2, \ldots$, be the intervals of $\mathcal{T}$ with rational end points. Then there exists a P-null set $N = \cup N_n$ such that for $t \notin N$

$$I_{B_n}(t) = P^{X|t}\left(T^{-1}B_n\right)$$

for all n. For fixed $t \notin N$, the two set functions $P^{X|t}\left(T^{-1}B\right)$ and $I_B(t)$ are probability distributions over $\mathcal{B}$, the latter assigning probability 1 or 0 to a set as it does or does not contain the point t. Since these distributions agree over the rational intervals B_n, they agree for all $B \in \mathcal{B}$. In particular, for $t \notin N$, the set consisting of the single point t is in $\mathcal{B}$, and if

$$A^{(t)} = \{x : T(x) = t\},$$

it follows that for all $t \notin N$

$$P^{X|t}\left(A^{(t)}\right) = 1. \tag{2.26}$$

Thus

$$\begin{aligned}\int h[T(x)]f(x)\, dP^{X|t}(x) &= \int_{A^{(t)}} h[T(x)]f(x)\, dP^{X|t}(x) \\ &= h(t)\int f(x)\, dP^{X|t}(x)\end{aligned}$$

for $t \notin N$, as was to be proved. ■

It is a consequence of Theorem 2.5.3 that for all $t \notin N$, $E[h(T) \mid t] = h(t)$ and hence in particular $P(T \in B \mid t) = 1$ or 0 as $t \in B$ or $t \notin B$.

The conditional distributions $P^{X|t}$ still differ from those of the elementary case considered in Section 1.9, in being defined over $(\mathcal{X}, \mathcal{A})$ rather than over the set $A^{(t)}$ and the σ-field $\mathcal{A}^{(t)}$ of its Borel subsets. However, (2.26) implies that for $t \notin N$

$$P^{X|t}(A) = P^{X|t}(A \cap A^{(t)}).$$

The calculations of conditional probabilities and expectations are therefore unchanged if for $t \notin N$, $P^{X|t}$ is replaced by the distribution $\bar{P}^{X|t}$, which is defined over $(A^{(t)}, \mathcal{A}^{(t)})$ and which assigns to any subset of $A^{(t)}$ the same probability as $P^{X|t}$.

Theorem 2.5.3 establishes for all $t \notin N$ the existence of conditional probability distributions $\bar{P}^{X|t}$, which are defined over $(A^{(t)}, \mathcal{A}^{(t)})$ and which by Lemma 2.4.2 satisfy

$$E[f(X)] = \int_{\mathcal{T}-N} \left[\int_{A^{(t)}} f(x)\, dP^{(X|t)}(x) \right] dP^{T}(t) \tag{2.27}$$

for all integrable functions f. Conversely, consider any family of distributions satisfying (2.27), and the experiment of observing first T, and then, if $T = t$, a random quantity with distribution $\bar{P}^{X|t}$. The result of this two-stage procedure is a point distributed over $(\mathcal{X}, \mathcal{A})$ with the same distribution as the original X. Thus $\bar{P}^{X|t}$ satisfies this "functional" definition of conditional probability.

If $(\mathcal{X}, \mathcal{A})$ is a product space $(\mathcal{T} \times \mathcal{Y}, \mathcal{B} \times \mathcal{C})$, then $A^{(t)}$ is the product of $\mathcal{Y}$ with the set consisting of the single point t. For $t \notin N$, the conditional distribution $\bar{P}^{X|t}$ then induces a distribution over $(\mathcal{Y}, \mathcal{C})$, which in analogy with the elementary case will be denoted by $P^{Y|t}$. In this case, the definition can be extended to all of $\mathcal{T}$ by letting $P^{Y|t}$ assign probability 1 to a common specified point y_0 for all $t \in N$. With this definition, (2.27) becomes

$$Ef(T, Y) = \int_{\mathcal{T}} \left[\int_{\mathcal{Y}} f(t, y)\, dP^{Y|t}(y) \right] dP^{T}(t). \tag{2.28}$$

As an application, we shall prove the following lemma, which will be used in Section 2.7.

Lemma 2.5.1 *Let $(\mathcal{T}, \mathcal{B})$ and $(\mathcal{Y}, \mathcal{C})$ be Euclidean spaces, and let $P_0^{T,Y}$ be a distribution over the product space $(\mathcal{X}, \mathcal{A}) = (\mathcal{T} \times \mathcal{Y}, \mathcal{B} \times \mathcal{C})$. Suppose that another distribution P_1 over $(\mathcal{X}, \mathcal{A})$ is such that*

$$dP_1(t, y) = a(y)b(t)\, dP_0(t, y),$$

with $a(y) > 0$ for all y. Then under P_1 the marginal distribution of T and a version of the conditional distribution of Y given t are given by

$$dP_1^{T}(t) = b(t) \left[\int a(y)\, dP_0^{Y|t}(y) \right] dP_0^{T}(t)$$

and

$$dP_1^{Y|t}(y) = \frac{a(y)\, dP_0^{Y|t}(y)}{\int_{\mathcal{Y}} a(y')\, dP_0^{Y|t}(y')}.$$

PROOF. The first statement of the lemma follows from the equation

$$\begin{aligned} P_1\{T \in B\} = E_1[I_B(T)] &= E_0[I_B(T)a(Y)b(T)] \\ &= \int_B b(T) \left[\int_{\mathcal{Y}} a(y)\, dP_0^{Y|t}(y) \right] dP_0^{T}(t). \end{aligned}$$

To check the second statement, one need only to show that for any integrable f the expectation $E_1 f(Y, T)$ satisfies (2.28), which is immediate. The denominator of $dP_1^{Y|t}$ is positive, since $a(y) > 0$ for all y. ■

2.6 Characterization of Sufficiency

We can now generalize the definition of sufficiency given in Section 1.9. If $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ is any family of distributions defined over a common sample space $(\mathcal{X}, \mathcal{A})$, a statistic T is *sufficient* for $\mathcal{P}$ (or for θ) if for each A in $\mathcal{A}$ there exists a determination of the conditional probability function $P_\theta(A \mid t)$ that is independent of θ. As an example suppose that $X_1, \ldots, X_n$ are identically and independently distributed with continuous distribution function $F_\theta, \theta \in \Omega$. Then it follows from Example 2.4.1 that the set of order statistics $T(X) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for θ.

Theorem 2.6.1 *If $\mathcal{X}$ is Euclidean, and if the statistic T is sufficient for $\mathcal{P}$, then there exist determinations of the conditional probability distributions $P_\theta(A \mid t)$ which are independent of θ and such that for each fixed t, $P_\theta(A \mid t)$ is a probability measure over $\mathcal{A}$.*

PROOF. This is seen from the proof of Theorem 2.5.1. By the definition of sufficiency one can, for each rational number r, take the functions $F(r, t)$ to be independent of θ, and the resulting conditional distributions will then also not depend on θ. ■

In Chapter 1, the definition of sufficiency was justified by showing that in a certain sense a sufficient statistic contains all the available information. In view of Theorem 2.6.1, the same justification applies quite generally when the sample space is Euclidean. With the help of a random mechanism one can then construct from a sufficient statistic T a random vector X' having the same distribution as the original sample vector X. Another generalization of the earlier result, not involving the restriction to a Euclidean sample space, is given in Problem 2.13.

The factorization criterion of sufficiency, derived in Chapter 1, can be extended to any *dominated* family of distributions, that is, any family $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ possessing probability densities p_θ with respect to some σ-finite measure μ over $(\mathcal{X}, \mathcal{A})$. The proof of this statement is based on the existence of a probability distribution $\lambda = \sum c_i P_{\theta_i}$ (Theorem 2.2.3 of the Appendix), which is *equivalent* to $\mathcal{P}$ in the sense that for any $A \in \mathcal{A}$

$$\lambda(A) = 0 \quad \text{if and only if} \quad P_\theta = 0 \quad \text{for all } \theta \in \Omega. \tag{2.29}$$

Theorem 2.6.2 *Let $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ be a dominated family of probability distributions over $(\mathcal{X}, \mathcal{A})$, and let $\lambda = \sum c_i P_{\theta_i}$ satisfy (2.29). Then a statistic T with range*

space $(\mathcal{T}, \mathcal{B})$ *is sufficient for* $\mathcal{P}$ *if and only if there exist nonnegative* $\mathcal{B}$*-measurable functions* $g_\theta(t)$ *such that*

$$dP_\theta(x) = g_\theta[T(x)]\, d\lambda(x) \tag{2.30}$$

for all $\theta \in \Omega$.

PROOF. Let $\mathcal{A}_0$ be the subfield induced by T, and suppose that T is sufficient for θ. Then for all $\theta \in \Omega$, $A_0 \in \mathcal{A}_0$, and $A \in \mathcal{A}$

$$\int_{A_0} P(A \mid T(x))\, dP_\theta(x) = P_\theta(A \cap A_0),$$

and since $\lambda = \sum c_i P_{\theta_i}$,

$$\int_{A_0} P(A \mid T(x))\, d\lambda(x) = \lambda(A \cap A_0),$$

so that $P(A \mid T(x))$ serves as conditional probability function also for λ. Let $g_\theta(T(x))$ be the Radon–Nikodym derivative $dP_\theta(x)/d\lambda(x)$ for $(\mathcal{A}_0, \lambda)$. To prove (2.30) it is necessary to show that $g_\theta(T(x))$ is also the derivative of P_θ for $(\mathcal{A}, \lambda)$. If A_0 is put equal to $\mathcal{X}$ in the first displayed equation, this follows from the relation

$$\begin{aligned} P_\theta(A) &= \int P(A \mid T(x))\, dP_\theta(x) = \int E_\lambda[I_A(x) \mid T(x)]\, dP_\theta(x) \\ &= \int E_\lambda[I_A(x) \mid T(x)]\, g_\theta(T(x))\, d\lambda(x) \\ &= \int E_\lambda[g_\theta(T(x)) I_A(x) \mid T(x)]\, d\lambda(x) \\ &= \int g_\theta(T(x)) I_A(x)\, d\lambda(x) = \int_A g_\theta(T(x))\, d\lambda(x). \end{aligned}$$

Here the second equality uses the fact, established at the beginning of the proof, that $P(A \mid T(x))$ is also the conditional probability for λ; the third equality holds because the function being integrated is $\mathcal{A}_0$-measurable and because $dP_\theta = g_\theta\, d\lambda$ for $(\mathcal{A}_0, \lambda)$; the fourth is an application of Lemma 2.4.1(ii); and the fifth employs the defining property of conditional expectation.

Suppose conversely that (2.30) holds. We shall then prove that the conditional probability function $P_\lambda(A \mid t)$ serves as a conditional probability function for all $P \in \mathcal{P}$. Let $g_\theta(T(x)) = dP_\theta(x)/d\lambda(x)$ on $\mathcal{A}$ and for fixed A and θ define a measure ν over $\mathcal{A}$ by the equation $d\nu = I_A\, dP_\theta$. Then over $\mathcal{A}_0$, $d\nu(x)/dP_\theta(x) = E_\theta[I_A(X) \mid T(x)]$, and therefore

$$\frac{d\nu(x)}{d\lambda(x)} = P_\theta[A \mid T(x)] g_\theta(T(x)) \qquad \text{over } \mathcal{A}_0.$$

On the other hand, $d\nu(x)/d\lambda(x) = I_A(x)g_\theta(T(x))$ over $\mathcal{A}$, and hence

$$\begin{aligned}\frac{d\nu(x)}{d\lambda(x)} &= E_\lambda[I_A(X)g_\theta(T(X)) \mid T(x)] \\ &= P_\lambda[A \mid T(x)]g_\theta(T(x)) \qquad \text{over } \mathcal{A}_0.\end{aligned}$$

It follows that $P_\lambda(A \mid T(x))g_\theta(T(x)) = P_\theta(A \mid T(x))g_\theta(T(x))$ $(\mathcal{A}_0, \lambda)$ and hence $(\mathcal{A}_0, P_\theta)$. Since $g_\theta(T(x)) \neq 0$ $(\mathcal{A}_0, P_\theta)$, this shows that $P_\theta(A \mid T(x)) = P_\lambda(A \mid T(x))$ $(\mathcal{A}_0, P_\theta)$, and hence that $P_\lambda(A \mid T(x))$ is a determination of $P_\theta(A \mid T(x))$. ■

Instead of the above formulation, which explicitly involves the distribution λ, it is sometimes more convenient to state the result with respect to a given dominating measure μ.

Corollary 2.6.1 (Factorization Theorem) *If the distributions P_θ of $\mathcal{P}$ have probability densities $p_\theta = dP_\theta/d\mu$ with respect to a σ-finite measure μ, then T is sufficient for $\mathcal{P}$ if and only if there exist nonnegative $\mathcal{B}$-measurable functions g_θ on T and a nonnegative $\mathcal{A}$-measurable function h on $\mathcal{X}$ such that*

$$p_\theta(x) = g_\theta[T(x)]h(x) \qquad (\mathcal{A}, \mu). \tag{2.31}$$

PROOF. Let $\lambda = \sum c_i P_{\theta_i}$ satisfy (2.29). Then if T is sufficient, (2.31) follows from (2.30) with $h = d\lambda/d\mu$. Conversely, if (2.31) holds

$$d\lambda(x) = \sum c_i g_{\theta_i}[T(x)]h(x)\,d\mu(x) = k[T(x)]h(x)\,d\mu(x)$$

and therefore $dP_\theta(x) = g^*_\theta(T(x))\,d\lambda(x)$ where $g^*_\theta(t) = g_\theta(t)/k(t)$ when $k(t) > 0$ and may be defined arbitrarily when $k(t) = 0$. ■

For extensions of the factorizations theorem to undominated families, see Ghosh et al. (1981) and the literature cited there.

2.7 Exponential Families

An important family of distributions which admits a reduction by means of sufficient statistics is the *exponential family*, defined by probability densities of the form

$$p_\theta(x) = C(\theta)\exp\left[\sum_{j=1}^{k} Q_j(\theta)T_j(x)\right]h(x) \tag{2.32}$$

with respect to a σ-finite measure μ over a Euclidean sample space $(\mathcal{X}, \mathcal{A})$. Particular cases are the distributions of a sample $X = (X_1, \ldots, X_n)$ from a binomial, Poisson,

or normal distribution. In the binomial case, for example, the density (with respect to counting measure) is

$$\binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \exp\left[x \log\left(\frac{p}{1-p}\right)\right] \binom{n}{x}.$$

Example 2.7.1 If $Y_1, \ldots, Y_n$ are independently distributed, each with density (with respect to Lebesgue measure)

$$p_\sigma(y) = \frac{y^{[(f/2)-1]} \exp\left[-y/\left(2\sigma^2\right)\right]}{\left(2\sigma^2\right)^{f/2} \Gamma(f/2)}, \qquad y > 0, \tag{2.33}$$

then the joint distribution of the Y's constitutes an exponential family. For $\sigma = 1$, (2.33) is the density of the χ^2-distribution with f degrees of freedom, in particular, for f an integer this is the density of $\sum_{j=1}^{f} X_j^2$, where the X's are a sample from the normal distribution $N(0, 1)$. ■

Example 2.7.2 Consider n independent trials, each of them resulting in one of the s outcomes $E_1, \ldots, E_s$ with probabilities $p_1, \ldots, p_s$, respectively. If X_{ij} is 1 when the outcome of the ith trial is E_j and 0 otherwise, the joint distribution of the X's is

$$P\{X_{11} = x_{11}, \ldots, X_{ns}\} = p_1^{\sum x_{i1}} p_2^{\sum x_{i2}} \cdots p_s^{\sum x_{is}},$$

where all $x_{ij} = 0$ or 1 and $\sum_j x_{ij} = 1$. This forms an exponential family with $T_j(x) = \sum_{i=1}^{n} x_{ij}$ $(j = 1, \ldots, s-1)$. The joint distribution of the T's is the multinomial distribution $M(n; p_1, \ldots, p_s)$ given by

$$P\{T_1 = t_1, \ldots, T_{s-1} = t_{s-1}\} \tag{2.34}$$

$$= \frac{n!}{t_1! \ldots t_{s-1}!(n - t_1 - \cdots - t_{s-1})!}$$

$$\times p_1^{t_1} \ldots p_{s-1}^{t_{s-1}} (1 - p_1 - \cdots - p_{s-1})^{n - t_1 - \cdots - t_{s-1}}. \blacksquare$$

If $X_1, \ldots, X_n$ is a sample from a distribution with density (2.32), the joint distribution of the X's constitutes an exponential family with the sufficient statistics $\sum_{i=1}^{n} T_j(X_i)$, $j = 1, \ldots, k$. Thus there exists a k-dimensional sufficient statistic for $(X_1, \ldots, X_n)$ regardless of the sample size. Suppose conversely that $X_1, \ldots, X_n$ is a sample from a distribution with some density $p_\theta(x)$ and that the set over which this density is positive is independent of θ. Then under regularity assumptions which make the concept of dimensionality meaningful, if there exists a k-dimensional sufficient statistic with $k < n$, the densities $p_\theta(x)$ constitute an exponential family. For a proof of this result, see Darmois (1935), Koopman (1936), and Pitman (1937,1938a). Regularity conditions of the result are discussed in Barankin and Maitra (1963), Brown (1964), Barndorff–Nielsen and Pedersen (1968), and Hipp (1974).

Employing a more natural parametrization and absorbing the factor $h(x)$ into μ, we shall write an exponential family in the form $dP_\theta(x) = p_\theta(x)\,d\mu(x)$ with

$$p_\theta(x) = C(\theta)\exp\left[\sum_{j=1}^{k}\theta_j T_j(x)\right]. \tag{2.35}$$

For suitable choice of the constant $C(\theta)$, the right-hand side of (2.35) is a probability density provided its integral is finite. The set Ω of parameter points $\theta = (\theta_1, \ldots, \theta_k)$ for which this is the case is the *natural parameter space* of the exponential family (2.35).

Optimum tests of certain hypotheses concerning any θ_j are obtained in Chapter 4. We shall now consider some properties of exponential families required for this purpose.

Lemma 2.7.1 *The natural parameter space of an exponential family is convex.*

PROOF. Let $(\theta_1, \ldots, \theta_k)$ and $(\theta'_1, \ldots, \theta'_k)$ be two parameter points for which the integral of (2.35) is finite. Then by Hölder's inequality,

$$\int \exp\left[\sum\left[\alpha\theta_j + (1-\alpha)\theta'_j\right]T_j(x)\right]d\mu(x)$$
$$\leq \left[\int \exp\left[\sum\theta_j T_j(x)\right]d\mu(x)\right]^{\alpha}\left[\int \exp\left[\sum\theta'_j T_j(x)\right]d\mu(x)\right]^{1-\alpha} < \infty$$

for any $0 < \alpha < 1$. ■

If the convex set Ω lies in a linear space of dimension $< k$, then (2.35) can be rewritten in a form involving fewer than k components of T. We shall therefore, without loss of generality, assume Ω to be k-dimensional.

It follows from the factorization theorem that $T(x) = (T_1(x), \ldots, T_k(x))$ is sufficient for $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$.

Lemma 2.7.2 *Let X be distributed according to the exponential family*

$$dP^T_{\theta,\vartheta}(x) = C(\theta,\vartheta)\exp\left[\sum_{i=1}^{r}\theta_i U_i(x) + \sum_{j=1}^{s}\vartheta_j T_j(x)\right]d\mu(x).$$

Then there exist measures λ_θ and ν_t over s- and r-dimensional Euclidean space respectively such that

(i) *the distribution of $T = (T_1, \ldots, T_s)$ is an exponential family of the form*

$$dP^T_{\theta,\vartheta}(t) = C(\theta,\vartheta)\exp\left(\sum_{j=1}^{s}\vartheta_j t_j\right)d\lambda_\theta(t), \tag{2.36}$$

(ii) *the conditional distribution of* $U = (U_1, \ldots, U_r)$ *given* $T = t$ *is an exponential family of the form*

$$dP_{\theta\cdot}^{U|t}(u) = C(\theta) \exp\left(\sum_{i=1}^{r} \theta_i u_i\right) d\nu_t(u), \tag{2.37}$$

and hence in particular is independent of ϑ.

PROOF. Let (θ^0, ϑ^0) be a point of the natural parameter space, and let $\mu^* = P_{\theta^0,\vartheta^0}^X$. Then

$$dP_{\theta^0,\vartheta^0}^X(x) = \frac{C(\theta, \vartheta)}{C(\theta^0, \vartheta^0)} \times \exp\left[\sum_{i=1}^{r}(\theta_i - \theta_i^0)U_i(x) + \sum_{j=1}^{s}(\vartheta_j - \vartheta_j^0)T_j(x)\right] d\mu^*(x),$$

and the result follows from Lemma 2.5.1, with

$$d\lambda_\theta(t) = \exp\left(-\sum \vartheta_i^0 t_i\right)\left[\int \exp\left[\sum_{i=1}^{r}(\theta_i - \theta_i^0)u_i\right] dP_{\theta^0,\vartheta^0}^{U|t}(u)\right] dP_{\theta^0,\vartheta^0}^{T}(t)$$

and

$$d\nu_t(u) = \exp\left(-\sum \theta_i^0 u_i\right) dP_{\theta^0,\vartheta^0}^{U|t}(u). \blacksquare$$

Theorem 2.7.1 *Let* ϕ *be any function on* $(\mathcal{X}, \mathcal{A})$ *for which the integral*

$$\int \phi(x) \exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] d\mu(x) \tag{2.38}$$

considered as a function of the complex variables $\theta_j = \xi_j + i\eta_j$ $(j = 1, \ldots, k)$ *exists for all* $(\xi_1, \ldots, \xi_k) \in \Omega$ *and is finite. Then*

(i) *the integral is an analytic function of each of the* θ*'s in the region* R *of parameter points for which* $(\xi_1, \ldots, \xi_k)$ *is an interior point of the natural parameter space* Ω*;*

(ii) *the derivatives of all orders with respect to the* θ*'s of the integral (2.38) can be computed under the integral sign.*

PROOF. Let $(\xi_1, \ldots, \xi_k)$ be any fixed point in the interior of Ω, and consider one of the variables in question, say θ_1. Breaking up the factor

$$\phi(x) \exp\left[\left(\xi_2^0 + i\eta_2^0\right) T_2(x) + \cdots + \left(\xi_k^0 + i\eta_k^0\right) T_k(x)\right]$$

into its real and complex parts and each of these into its positive and negative parts, and absorbing this factor in each of the four terms thus obtained into the measure μ, one sees that as a function of θ_1 the integral (2.38) can be written as

$$\int \exp\left[\theta_1 T_1(x)\right] d\mu_1(x) - \int \exp\left[\theta_1 T_1(x)\right] d\mu_2(x) + i \int \exp\left[\theta_1 T_1(x)\right] d\mu_3(x) - i \int \exp\left[\theta_1 T_1(x)\right] d\mu_4(x).$$

It is therefore sufficient to prove the result for integrals of the form

$$\psi(\theta_1) = \int \exp\left[\theta_1 T_1(x)\right] d\mu(x).$$

Since $(\xi_1^0, \ldots, \xi_k^0)$ is in the interior of Ω, there exists $\delta > 0$ such that $\psi(\theta_1)$ exists and is finite for all θ_1 with $|\xi_1 - \xi_1^0| \leq \delta$. Consider the difference

$$\frac{\psi(\theta_1) - \psi(\theta_1^0)}{\theta_1 - \theta_1^0} = \int \frac{\exp\left[\theta_1 T_1(x)\right] - \exp\left[\theta_1^0 T_1(x)\right]}{\theta_1 - \theta_1^0} d\mu(x).$$

The integrand can be written as

$$\exp\left[\theta_1^0 T_1(x)\right] \left[\frac{\exp\left[(\theta_1 - \theta_1^0) T_1(x)\right] - 1}{\theta_1 - \theta_1^0}\right].$$

Applying to the second factor the inequality

$$\left|\frac{\exp(az) - 1}{z}\right| \leq \frac{\exp(\delta|a|)}{\delta} \qquad \text{for} \quad |z| \leq \delta,$$

the integrand is seen to be bounded above in absolute value by

$$\frac{1}{\delta}\left|\exp\left(\theta_1^0 T_1 + \delta|T_1|\right)\right| \leq \frac{1}{\delta}\left|\exp\left[\left(\theta_1^0 + \delta\right) T_1\right] + \exp\left[\left(\theta_1^0 - \delta\right) T_1\right]\right|$$

for $|\theta_1 - \theta_1^0| \leq \delta$. Since the right-hand side integrable, it follows from the Lebesgue Dominated Convergence Theorem [Theorem 2.2.2(ii)] that for any sequence of points $\theta_1^{(n)}$ tending to θ_1^0, the difference quotient of ψ tends to

$$\int T_1(x) \exp\left[\theta_1^0 T_1(x)\right] d\mu(x).$$

This completes the proof of (i), and proves (ii) for the first derivative. The proof for the higher derivatives is by induction and is completely analogous. ■

2.8 Problems

Section 2.1

Problem 2.1 *Monotone class.* A class $\mathcal{F}$ of subsets of a space is a *field* if it contains the whole space and is closed under complementation and under finite unions; a class $\mathcal{M}$ is *monotone* if the union and intersection of every increasing and decreasing sequence of sets of $\mathcal{M}$ is again in $\mathcal{M}$. The smallest monotone class $\mathcal{M}_0$ containing a given field $\mathcal{F}$ coincides with the smallest σ-field $\mathcal{A}$ containing $\mathcal{F}$. [One proves first that $\mathcal{M}_0$ is a field. To show, for example, that $A \cap B \in \mathcal{M}_0$ when A and B are in $\mathcal{M}_0$, consider, for a fixed set $A \in \mathcal{F}$, the class $\mathcal{M}_A$ of all B in $\mathcal{M}_0$ for which $A \cap B \in \mathcal{M}_0$. Then $\mathcal{M}_A$ is a monotone class containing $\mathcal{F}$, and hence $\mathcal{M}_A = \mathcal{M}_0$. Thus $A \cap B \in \mathcal{M}_A$ for all B. The argument can now be repeated with a fixed set $B \in \mathcal{M}_0$ and the class $\mathcal{M}_B$ of sets A in $\mathcal{M}_0$ for which $A \cap B \in \mathcal{M}_0$. Since $\mathcal{M}_0$ is a field and monotone, it is a σ-field containing $\mathcal{F}$ and hence contains $\mathcal{A}$. But any σ-field is a monotone class so that also $\mathcal{M}_0$ is contained in $\mathcal{A}$.]

Section 2.2

Problem 2.2 Prove Corollary 2.2.1 using Theorems 2.2.1 and 2.2.2.

Problem 2.3 *Radon–Nikodym derivatives.*
(i) If λ and μ are σ-finite measures over $(\mathcal{X}, \mathcal{A})$ and μ is absolutely continuous with respect to λ, then

$$\int f \, d\mu = \int f \frac{d\mu}{d\lambda} \, d\lambda$$

for any μ-integrable function f.
(ii) If λ, μ, and ν are σ-finite measures over $(\mathcal{X}, \mathcal{A})$ such that ν is absolutely continuous with respect to μ and μ with respect to λ, then

$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu} \frac{d\mu}{d\lambda} \qquad \text{a.e. } \lambda.$$

(iii) If μ and ν are σ-finite measures,, which are *equivalent* in the sense that each is absolutely continuous with respect to the other, then

$$\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu} \right)^{-1} \qquad \text{a.e. } \mu, \nu.$$

(iv) If $\mu_k, k = 1, 2, \ldots,$ and μ are finite measures over $(\mathcal{X}, \mathcal{A})$ such that $\sum_{k=1}^{\infty} \mu_k(A) = \mu(A)$ for all $A \in \mathcal{A}$, and if the μ_k are absolutely continuous with respect to a σ-finite measure λ, then μ is absolutely continuous with respect to λ, and

$$\frac{d \sum_{k=1}^{n} \mu_k}{d\lambda} = \sum_{k=1}^{n} \frac{d\mu_k}{d\lambda}, \quad \lim_{n\to\infty} \frac{d \sum_{k=1}^{n} \mu_k}{d\lambda} = \frac{d\mu}{d\lambda} \qquad \text{a.e. } \lambda.$$

[(i): The equation in question holds when f is the indicator of a set, hence when f is simple, and therefore for all integrable f.
(ii): Apply (i) with $f = d\nu/d\mu$.]

Problem 2.4 If $f(x) > 0$ for all $x \in S$ and μ is σ-finite, then $\int_S f\, d\mu = 0$ implies $\mu(S) = 0$.

[Let S_n be the subset of S on which $f(x) \geq 1/n$. Then $\mu(S) \leq \sum \mu(S_n)$ and $\mu(S_n) \leq n \int_{S_n} f\, d\mu \leq n \int_S f\, d\mu = 0$.]

Section 2.3

Problem 2.5 Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, and $\mathcal{A}_0$ a σ-field contained in $\mathcal{A}$. Suppose that for any function T, the σ-field $\mathcal{B}$ is taken as the totality of sets B such that $T^{-1}(B) \in \mathcal{A}$. Then it is not necessarily true that there exists a function T such that $T^{-1}(\mathcal{B}) \in \mathcal{A}_0$. [An example is furnished by any $\mathcal{A}_0$ such that for all x the set consisting of the single point x is in $\mathcal{A}_0$.]

Section 2.4

Problem 2.6 (i) Let $\mathcal{P}$ be any family of distributions $X = (X_1, \ldots, X_n)$ such that

$$P\{(X_i, X_{i+1}, \ldots, X_n, X_1, \ldots, X_{i-1}) \in A\} = P\{(X_1, \ldots, X_n) \in A\}$$

for all Borel sets A and all $i = 1, \ldots, n$. For any sample point $(x_1, \ldots, x_n)$ define $(y_1, \ldots, y_n) = (x_i, x_{i+1}, \ldots, x_n, x_1, \ldots, x_{i-1})$, where $x_i = x_{(1)} = \min(x_1, \ldots, x_n)$. Then the conditional expectation of $f(X)$ given $Y = y$ is

$$f_0(y_1, \ldots, y_n) = \frac{1}{n}[f(y_1, \ldots, y_n) + f(y_2, \ldots, y_n, y_1) + \cdots + f(y_n, y_1, \ldots, y_{n-1})].$$

(ii) Let $G = \{g_1, \ldots, g_r\}$ be any group of permutations of the coordinates $x_1, \ldots, x_n$ of a point x in n-space, and denote by gx the point obtained by applying g to the coordinates of x. Let $\mathcal{P}$ be any family of distributions P of $X = (X_1, \ldots, X_n)$ such that

$$P\{gX \in A\} = P\{X \in A\} \qquad \text{for all} \quad g \in G. \tag{2.39}$$

For any point x let $t = T(x)$ be any rule that selects a unique point from the r points $g_k x, k = 1, \ldots, r$ (for example the smallest first coordinate if this defines it uniquely, otherwise also the smallest second coordinate, etc.). Then

$$E[f(X) \mid t] = \frac{1}{r} \sum_{k=1}^{r} f(g_k t).$$

(iii) Suppose that in (ii) the distributions P do not satisfy the invariance condition (2.39) but are given by

$$dP(x) = h(x)\, d\mu(x),$$

where μ is invariant in the sense that $\mu\{x : gx \in A\} = \mu(A)$. Then

$$E[f(X) \mid t] = \frac{\sum_{k=1}^{r} f(g_k t) h(g_k t)}{\sum_{k=1}^{r} h(g_k t)}.$$

Section 2.5

Problem 2.7 Prove Theorem 2.5.1 for the case of an n-dimensional sample space. [The condition that the cumulative distribution function is nondecreasing is replaced by $P\{x_1 < X_1 \le x_1', \ldots, x_n < X_n \le x_n'\} \ge 0$; the condition that it is continuous on the right can be stated as $\lim_{m\to\infty} F(x_1 + 1/m, \ldots, x_n + 1/m) = F(x_1, \ldots, x_n)$.]

Problem 2.8 Let $\mathcal{X} = \mathcal{Y} \times \mathcal{T}$, and suppose that P_0, P_1 are two probability distributions given by

$$dP_0(y, t) = f(y)g(t)\, d\mu(y)\, d\nu(t),$$
$$dP_1(y, t) = h(y, t)\, d\mu(y)\, d\nu(t),$$

where $h(y, t)/f(y)g(t) < \infty$. Then under P_1 the probability density of Y with respect to μ is

$$p_1^Y(y) = f(y) E_0\left[\frac{h(y, T)}{f(y)g(T)} \,\middle|\, Y = y\right].$$

[We have

$$p_1^Y(y) = \int_{\mathcal{T}} h(y,t)\,d\nu(t) = f(y)\int_{\mathcal{T}} \frac{h(y,t)}{f(y)g(t)}g(t)\,d\nu(t).]$$

Section 2.6

Problem 2.9 *Symmetric distributions.*

(i) Let $\mathcal{P}$ be any family of distributions of $X = (X_1, \ldots, X_n)$ which are symmetric in the sense that

$$P\left\{\left(X_{i_1}, \ldots, X_{i_n}\right) \in A\right\} = P\left\{(X_1, \ldots, X_n) \in A\right\}$$

for all Borel sets A and all permutations $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$. Then the statistic T of Example 2.4.1 is sufficient for $\mathcal{P}$, and the formula given in the first part of the example for the conditional expectation $E[f(X) \mid T(x)]$ is valid.

(ii) The statistic Y of Problem 2.6 is sufficient.

(iii) Let $X_1, \ldots, X_n$ be identically and independently distributed according to a continuous distribution $P \in \mathcal{P}$, and suppose that the distributions of $\mathcal{P}$ are symmetric with respect to the origin. Let $V_i = |X_i|$ and $W_i = V_{(i)}$. Then $(W_1, \ldots, W_n)$ is sufficient for $\mathcal{P}$.

Problem 2.10 *Sufficiency of likelihood ratios.* Let P_0, P_1 be two distributions with densities p_0, p_1. Then $T(x) = p_1(x)/p_0(x)$ is sufficient for $\mathcal{P} = \{P_0, P_1\}$. [This follows from the factorization criterion by writing $p_1 = T \cdot p_0$, $p_0 = 1 \cdot p_0$.]

Problem 2.11 *Pairwise sufficiency.* A statistic T is pairwise sufficient for $\mathcal{P}$ if it is sufficient for every pair of distributions in $\mathcal{P}$.

(i) If $\mathcal{P}$ is countable and T is pairwise sufficient for $\mathcal{P}$, then T is sufficient for $\mathcal{P}$.

(ii) If $\mathcal{P}$ is a dominated family and T is pairwise sufficient for $\mathcal{P}$, then T is sufficient for $\mathcal{P}$.

[(i): Let $\mathcal{P} = \{P_0, P_1, \ldots\}$, and let $\mathcal{A}_0$ be the sufficient subfield induced by T. Let $\lambda = \sum c_i P_i$ $(c_i > 0)$ be equivalent to $\mathcal{P}$. For each $j = 1, 2, \ldots$ the probability measure λ_j that is proportional to $(c_0/n)P_0 + c_j P_j$ is equivalent to $\{P_0, P_j\}$. Thus by pairwise sufficiency, the derivative $f_j = dP_0/[(c_0/n)\,dP_0 + c_j\,dP_j]$ is $\mathcal{A}_0$-measurable. Let $S_j = \{x : f_j(x) = 0\}$ and $S = \bigcup_{j=1}^n S_j$. Then $S \in \mathcal{A}_0$, $P_0(S) = 0$, and on $\mathcal{X} - S$ the derivative $dP_0/d\sum_{j=1}^n c_j P_j$ equals $(\sum_{j=1}^n 1/f_j)^{-1}$ which is $\mathcal{A}_0$-measurable. It then follows from Problem 2.3 that

$$\frac{dP_0}{d\lambda} = \frac{dP_0}{d\sum_{j=0}^n c_j P_j}\,\frac{d\sum_{j=0}^n c_j P_j}{d\lambda}$$

is also $\mathcal{A}_0$-measurable. (ii): Let $\lambda = \sum_{j=1}^{\infty} c_j P_{\theta_j}$ be equivalent to $\mathcal{P}$. Then pairwise sufficiency of T implies for any θ_0 that $dP_{\theta_0}/(dP_{\theta_0} + d\lambda)$ and hence $dP_{\theta_0}/d\lambda$ is a measurable function of T.]

Problem 2.12 If a statistic T is sufficient for $\mathcal{P}$, then for every function f which is $(\mathcal{A}, P_\theta)$-integrable for all $\theta \in \Omega$ there exists a determination of the conditional expectation function $E_\theta[f(X) \mid t]$ that is independent of θ. [If $\mathcal{X}$ is Euclidean, this follows from Theorems 2.5.2 and 2.6.1. In general, if f is nonnegative there exists a nondecreasing sequence of simple nonnegative functions f_n tending to f. Since the conditional expectation of a simple function can be taken to be independent of θ by Lemma 2.4.1(i), the desired result follows from Lemma 2.4.1(iv).]

Problem 2.13 For a decision problem with a finite number of decisions, the class of procedures depending on a sufficient statistic T only is essentially complete. [For Euclidean sample spaces this follows from Theorem 2.5.1 without any restriction on the decision space. For the present case, let a decision procedure be given by $\delta(x) = (\delta^{(1)}(x), \ldots, \delta^{(m)}(x))$ where $\delta^{(i)}(x)$ is the probability with which decision d_i is taken when x is observed. If T is sufficient and $\eta^{(i)}(t) = E[\delta^{(i)}(X) \mid t]$, the procedures δ and η have identical risk functions.] [More general versions of this result are discussed, for example, by Elfving (1952), Bahadur (1955), Burkholder (1961), LeCam (1964), and Roy and Ramamoorthi (1979).]

Section 2.7

Problem 2.14 Let X_i $(i = 1, \ldots, s)$ be independently distributed with Poisson distribution $P(\lambda_i)$, and let $T_0 = \sum X_j$, $T_i = X_i$, $\lambda = \sum \lambda_j$. Then T_0 has the Poisson distribution $P(\lambda)$, and the conditional distribution of $T_1, \ldots, T_{s-1}$ given $T_0 = t_0$ is the multinomial distribution (2.34) with $n = t_0$ and $p_i = \lambda_i/\lambda$.

Problem 2.15 *Life testing.* Let $X_1, \ldots, X_n$ be independently distributed with exponential density $(2\theta)^{-1}e^{-x/2\theta}$ for $x \geq 0$, and let the ordered X's be denoted by $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. It is assumed that Y_1 becomes available first, then Y_2, and so on, and that observation is continued until Y_r has been observed. This might arise, for example, in life testing where each X measures the length of life of, say, an electron tube, and n tubes are being tested simultaneously. Another application is to the disintegration of radioactive material, where n is the number of atoms, and observation is continued until r α-particles have been emitted.

(i) The joint distribution of $Y_1, \ldots, Y_r$ is an exponential family with density

$$\frac{1}{(2\theta)^r}\frac{n!}{(n-r)!}\exp\left[-\frac{\sum_{i=1}^{r} y_i + (n-r)y_r}{2\theta}\right], \qquad 0 \leq y_1 \leq \cdots \leq y_r.$$

(ii) The distribution of $[\sum_{i=1}^{r} Y_i + (n-r)Y_r]/\theta$ is χ^2 with $2r$ degrees of freedom.

(iii) Let $Y_1, Y_2, \ldots$ denote the time required until the first, second, ... event occurs in a Poisson process with parameter $1/2\theta'$ (see Problem 1.1). Then $Z_1 = Y_1/\theta'$, $Z_2 = (Y_2 - Y_1)/\theta'$, $Z_3 = (Y_3 - Y_2)/\theta', \ldots$ are independently distributed as χ^2 with 2 degrees of freedom, and the joint density $Y_1, \ldots, Y_r$ is an exponential family with density

$$\frac{1}{(2\theta')^r} \exp\left(-\frac{y_r}{2\theta'}\right), \qquad 0 \le y_1 \le \cdots \le y_r.$$

The distribution of Y_r/θ' is again χ^2 with $2r$ degrees of freedom.

(iv) The same model arises in the application to life testing if the number n of tubes is held constant by replacing each burned-out tube with a new one, and if Y_1 denotes the time at which the first tube burns out, Y_2 the time at which the second tube burns out, and so on, measured from some fixed time.

[(ii): The random variables $Z_i = (n-i+1)(Y_i - Y_{i-1})/\theta \quad (i = 1, 2, \ldots, r)$ are independently distributed as χ^2 with 2 degrees of freedom, and $[\sum_{i=1}^{r} Y_i + (n-r)Y_r/\theta = \sum_{i=1}^{r} Z_i$.]

Problem 2.16 For any θ which is an interior point of the natural parameter space, the expectations and covariances of the statistics T_j in the exponential family (2.35) are given by

$$E\left[T_j(X)\right] = -\frac{\partial \log C(\theta)}{\partial \theta_j} \qquad (j = 1, \ldots, k),$$

$$E\left[T_i(X)T_j(X)\right] - \left[ET_i(X)ET_j(X)\right] = -\frac{\partial^2 \log C(\theta)}{\partial \theta_i \partial \theta_j} \qquad (i, j = 1, \ldots, k).$$

Problem 2.17 Let Ω be the natural parameter space of the exponential family (2.35), and for any fixed $t_{r+1}, \ldots, t_k$ $(r < k)$ let $\Omega'_{\theta_1 \ldots \theta_r}$ be the natural parameter space of the family of conditional distributions given $T_{r+1} = t_{r+1}, \ldots, T_k = t_k$.

(i) Then $\Omega'_{\theta_1, \ldots, \theta_r}$ contains the projection $\Omega_{\theta_1, \ldots, \theta_r}$ of Ω onto $\theta_1, \ldots, \theta_r$.

(ii) An example in which $\Omega_{\theta_1, \ldots, \theta_r}$ is a proper subset of $\Omega'_{\theta_1, \ldots, \theta_r}$ is the family of densities

$$p_{\theta_1 \theta_2}(x, y) = C(\theta_1, \theta_2) \exp(\theta_1 x + \theta_2 y - xy), \qquad x, y > 0.$$

2.9 Notes

The theory of measure and integration in abstract spaces and its application to probability theory, including in particular conditional probability and expectation, is treated in a number of books, among them Dudley (1989), Williams (1991), and Billingsley (1995). The material on sufficient statistics and exponential families is complemented by the corresponding sections in Lehmann and Casella (1998). Much fuller treatments of exponential families (as well as sufficiency) are provided by Barndorff–Nielsen (1978) and Brown (1986).