1

Statistical Experiments within the Measure Theoretical Framework

1.1 Introduction

This chapter is intended to serve several purposes. Firstly it contains tools which will be used throughout this work. Thus several of the basic concepts are introduced. Some of the definitions are repeated in later chapters, but others are only given here. It is not necessary to have acquired an understanding of everything here in order to proceed. Indeed some of the material in this chapter anticipates later developments and may not be fully comprehended before these developments have been studied. Thus the reader may just look over the chapter and later return to this material when the need arises.

Another purpose is to provide a self-contained introduction to some ideas which later will play a central role. Thus it is intended to be possible to go through this chapter before (or whilst) reading the later chapters.

A third purpose is to present some ‘classical’ material which is important as background for the theory of statistical experiments as it is presented here. In particular some of this material indicates the need of extending, as LeCam did, the measure theoretical framework of decision theory.

The notion of a statistical experiment (statistical model) is introduced in section 2. It is argued there that the usual notion of a statistic as a random variable is too narrow and that it should be extended to the more general concept of a consistent family. The optimistic goal is to make everything as simple (or difficult) as it is in the ‘finite’ case. Difficulties beyond the finite case may then be considered as being of technical nature.

Extending the notion of a bounded variable we arrive, following LeCam (1964), at the notion of the $M$-space of an experiment. The $M$-space is introduced as the dual of another space called the $L$-space. This $L$-space consists of those finite measures which are absolutely continuous with respect to
Statistical experiments within the measure theoretical framework

countably infinite convex combinations of the distributions which constitute
the experiment.

$L$-spaces of experiments are, as will be explained in chapter 4, particular
examples of abstract $L$-spaces as defined in linear analysis. Indeed any abstract
$L$-space is isomorphic to the $L$-space of a (totally informative) experiment as
shown by Kakutani (1941a). The abstract $M$-spaces which can be represented
as $M$-spaces of experiments are thus the abstract $M$-spaces which are duals
of abstract $L$-spaces.

Several methods for deriving new experiments from old ones are considered
in section 3. This may e.g. be done by reducing (or expanding) the parameter
set or the sample space. In particular we may consider experiments obtained
by observing certain functions of the basic observations. These functions may
involve random mechanisms and may thus not be strictly determined by the
observations.

We may also consider independent combinations, i.e. products, of experi-
ments. Another possibility is to observe a random experiment and then carry
out the observed experiment. This leads to the notion of a mixture of experi-
ments. Together products and mixtures provide an interesting algebraic struc-
ture for statistical experiments.

Using LeCam’s weak (or strong) experiment topology we may also derive
experiments as limits.

In section 4 we discuss functionals of experiments. In particular we consider
functionals which are determined by non-negatively homogeneous functions
on the likelihood space. Among them are: statistical distance, minimum Bayes
risk, affinity, Hellinger transform, the Kullback – Leibler information number
and the Fisher information matrix.

Section 5 is an exposition of some of the ‘classical’ theory of sufficiency. As
shown by Burkholder (1961) this theory exhibits peculiarities which indicate
the need for a revision of the distinction between essential statistical problems
on the one hand and technical problems on the other. These matters were to
a large extent clarified by LeCam’s 1964 paper.

It is useful to know that $\lambda$-systems of events generated by $\pi$-systems are
themselves $\pi$-systems. As we have not found any other natural place for this
we have included an exposition on these systems in an appendix.

Complements in chapter 10 which are of interest in connection with the
topics treated in this chapter are numbers 1, 2, 4–8, 10–12, 18, 27, 28, 36, 38,
40, 46 and 47.

We shall often employ the notion of a Euclidean measurable space for a
measurable space which is Borel isomorphic to a Borel subset of the real line.
Thus a measurable space $(\mathcal{X}, \sigma)$ is Euclidean if and only if there is a $1:1$
bimeasurable map from $\mathcal{X}$ onto a Borel subset of a complete separable metric
space. The isomorphism theorem for Borel subsets of complete separable
metric spaces (see e.g. Parthasarathy, 1967) implies that if \( X \) is uncountable then any uncountable Borel subset of any complete separable metric space may be chosen as the range of this bimeasurable map. Thus if \( X \) is uncountable then \((X, A)\) is Euclidean if and only if \((X, A)\) is Borel isomorphic to the real line.

Here are some other notions, notations and conventions which will be used throughout.

An ordering of a set \( D \) is a reflexive and transitive relation \( \geq \) on \( D \). Thus we do not require that an ordering is antisymmetric, i.e. that \( d_1 = d_2 \) when \( d_1 \geq d_2 \) and \( d_2 \geq d_1 \). The general terminology concerning ordered sets is described in section 2.4. We shall find it convenient to use the notations \( \wedge \) and \( \vee \) for inf and sup respectively.

A directed set \( D \) is an ordered set \( D \) such that for any two elements \( d_1, d_2 \) in \( D \), there is a third element \( d_3 \) in \( D \) such that \( d_3 \geq d_1 \) and \( d_3 \geq d_2 \). Thus \( D \) is directed if and only if finite subsets of \( D \) possess upper bounds.

A net (generalized sequence) is a family \((x_d : d \in D)\) where \( D \) is a directed set. We often use the simplified notation \((x_e)\) for a net \((x_d : d \in D)\). A net \((y_e : e \in E)\) is called a subnet of the net \((x_d : d \in D)\) if there is a function \( \phi \) such that \( y_e = x_{\phi(e)} \) for all elements \( e \) in \( E \) and also to each \( d \) in \( D \) there corresponds an element \( e_0 \) in \( E \) such that \( \phi(e) \geq_\phi d \) whenever \( e \geq_\phi e_0 \). A subnet of a net \((x_d)\) is often denoted by \((x_e)\).

A net \((x_d : d \in D)\) is in (or on) a set \( \mathcal{A} \) if \( x_d \in \mathcal{A} \) for all \( d \) in \( D \). If \( \mathcal{A} \) is a topological space and \( x \in \mathcal{A} \) then we say that a net \((x_d : d \in D)\) converges to \( x \), or that \( x \) is the limit of \((x_d)\), if any neighbourhood \( V \) of \( x \) contains some tail \( \{x_d : d \geq d_0\} \) of \((x_d)\). If \((x_d)\) converges to \( x \) then this may be expressed by writing \( \lim_d x_d = x \). Concerning a net \((x_d : d \in D)\) on the extended real line \([-\infty, \infty]\) we define the quantities \( \limsup_d x_d \) and \( \liminf_d x_d \) by:

\[
\liminf_d x_d = \sup_d \inf\{x_d : d \geq d_0\} \quad \text{and} \quad \limsup_d x_d = \inf_d \sup\{x_d : d \geq d_0\}.
\]

We refer the reader to, e.g. Kelley (1955) or Dunford & Schwartz (1958) (where nets are called generalized sequences) for the use of nets in point set topology.

We have found it convenient to use the notation \( \mathcal{L}(X) \) for the probability distribution of the random variable \( X \). Some additional notation is listed here:

\[
\# A = \text{the number of elements in } A.
\]

\[
o(h) = \text{function of the quantity } h \text{ having the property that } o(h)/h \to 0 \text{ as } h \to 0.
\]

\[
d\mu/d\nu = \text{the Radon–Nikodym density of the } \nu \text{-absolutely continuous part of } \mu \text{ with respect to } \nu.
\]

\[
\mu \gg \nu \text{ is short for ‘the measure } \mu \text{ dominates the measure } \nu', \text{i.e. } \nu \text{ is absolutely continuous with respect to } \mu.
\]
Statistical experiments within the measure theoretical framework

\(\mu \gg (\nu : \theta \in \Theta)\) is short for \(\mu \gg \nu\) when \(\theta \in \Theta\).
\(\langle a, b, \ldots \rangle = \) the convex hull of \(\{a, b, \ldots\}\).

Notation for intervals is:

\[
\begin{align*}
[a, b[ &= \{x : a < x < b\} \\
]a, b] &= \{x : a < x \leq b\} \\
[a, b[ &= \{x : a \leq x < b\} \\
[a, b] &= \{x : a \leq x \leq b\}.
\end{align*}
\]

1.2 The measure theoretical framework

Leaving mathematical niceties aside for the moment, a statistical experiment may be described as consisting of three parts:

(i) the list \(\mathcal{X}\) of possible outcomes. This list is usually called the sample space.
(ii) the list \(\Theta\) of explaining theories. This list is usually called the parameter set.
(iii) the rule which to each explaining theory \(\theta\) in \(\Theta\) associates the chance mechanism \(P_\theta\) explained by the theory \(\theta\). The chance mechanism \(P_\theta\) is the probability law which governs the random outcome in \(\mathcal{X}\) when the explaining theory \(\theta\) prevails.

Mathematically a statistical experiment \(\mathcal{E}\) may conveniently be described as a triple \(\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)\) where \((\mathcal{X}, \mathcal{A})\) is a measurable space, \(\Theta\) is a set and \(P_\theta\), for each \(\theta\) in \(\Theta\), is a probability measure on \(\mathcal{A}\). The measurable space \((\mathcal{X}, \mathcal{A})\) is called the sample space of the experiment \(\mathcal{E}\), while the set \(\Theta\) is called the parameter set of \(\mathcal{E}\). You may think of \(\theta\) as a possible law of nature (see (ii) above). If nature obeys \(\theta\) then the outcome of the experiment is assumed to be distributed according to \(P_\theta\).

The parameter set will, unless otherwise stated, be assumed to be fixed, although arbitrary, throughout. An experiment \(\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_{\theta} : \theta \in \Theta)\) with parameter set \(\Theta\) is simply a family of probability measures on a common measurable space. The sample space will often be suppressed in the notation of an experiment i.e. we may write \(\mathcal{E} = (P_{\theta} : \theta \in \Theta)\) instead of \(\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_{\theta} : \theta \in \Theta)\).

We shall find it convenient to say that a (\(\mathcal{X}\)-valued) random variable \(X\) realizes the experiment \(\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_{\theta} : \theta \in \Theta)\) if the distribution of \(X\) is \(P_{\theta}\) when \(\theta\) is the true value of the unknown parameter, i.e. when \(\theta\) prevails. Making this rigorous we may think of an experiment \((\mathcal{Z}, \mathcal{G}; R_{\theta} : \theta \in \Theta)\) and a measurable map \(X\) from \((\mathcal{Z}, \mathcal{G})\) to \((\mathcal{X}, \mathcal{A})\) so that \(R_{\theta} X^{-1} \equiv_{\theta} P_{\theta}\).

The reader should be aware that an experiment \(\mathcal{E} = (P_{\theta} : \theta \in \Theta)\) is a family of probability measures and is therefore quite different from the set \(\{P_{\theta} : \theta \in \Theta\}\) of probability measures. Thus if \(a_1\) and \(a_2\) are different numbers
then the pairs \((N(a_1, 1), N(a_2, 1))\) and \((N(a_2, 1), N(a_1, 1))\) of normal distributions define different experiments. Now \(a_1 + a_2 - X\) is distributed as, respectively, \(N(a_2, 1)\) and \(N(a_1, 1)\) as \(X\) is distributed as \(N(a_1, 1)\) or as \(N(a_2, 1)\). It follows that these two pairs, although different, ought to be statistically equivalent for any reasonable concept of statistical equivalence. This, however, is due to a particular property of symmetry which is satisfied by these pairs. We shall see in chapter 6 that ordered pairs \((P, Q)\) and \((Q, P)\) of probability measures on the same measurable space define ‘statistically’ equivalent experiments if and only if \(\mathcal{L}(dP/d(P + Q))^{1/2}(P + Q)\) is symmetric about \(\frac{1}{2}\).

Before proceeding let us make some general comments on the structure of an experiment \(\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_0 : \theta \in \Theta)\).

Firstly the class \(\mathcal{A}\) of subsets of the set \(\mathcal{X}\) is an algebra and each probability \(P_0\) is an additive non-negative measure on \(\mathcal{A}\). This implies essential assumptions on permissible operations of events as well as on behaviour of probabilities. The stronger assumptions, which are inherent in the definition, of the closure of \(\mathcal{A}\) under countably infinite unions (intersections) and of the \(\sigma\)-additivity of the probabilities \(P_0\), are of a more technical nature.

It is important to realize the flexibility of this notion of a sample space. The algebra \(\mathcal{A}\) decides what can be observed and what cannot be observed. If it should turn out, for example, that some observations are not performed or not utilized, then this may be described mathematically by decreasing the algebra \(\mathcal{A}\), keeping the set \(\mathcal{X}\) fixed.

The statistical reasons for using algebras \(\mathcal{A}\) which in general are smaller than the class of all subsets of \(\mathcal{X}\) should not be confused with the fact that well known measures may not be extended to all subsets. Indeed as the Hahn–Banach theorem implies that any additive probability set function may be extended to all subsets, it is doubtful if this has any statistical relevance.

If we try to classify experiments according to the cardinality of the parameter set \(\Theta\) then the simplest case is that of experiments having one-point parameter sets. This is the situation where the underlying theory is known. Within the context of comparison of statistical experiments this is essentially a trivial case. However, you lose more than you gain by excluding this case.

Experiments with two-point parameter sets are called dichotomies. Although the study of dichotomies is far from trivial, the general structure of dichotomies may be described in relatively simple terms. The richness of the general theory is already apparent in this case. Thus, as we shall see in chapter 9, the basic ideas of majorization and Schur convexity may be phrased within the framework of the theory of dichotomies. The results obtained for dichotomies may, of course, be applied to general experiments by, for example, restricting attention to two-point subsets of the parameter set.

As the next level of generality we may consider the case of experiments possessing a general finite parameter set. Now the theory of comparison of
experiments, as it has been developed by LeCam, adheres to the principle that
general structures should be approximable by finite ones. General statements
on experiments may often be obtained by first establishing statements for
experiments with finite parameter sets, and then using some approximation.
The difficulties encountered in the last step may appear technical, although
far from trivial. For the scope of the theory as well as for many applications,
it is important to know how such problems may be handled. We shall therefore
elaborate somewhat on this. The nicest approach, as shown by LeCam, is to
utilize the theory of vector lattices. We shall not assume any knowledge of
this theory here. Later on, in chapter 5, we shall present some results from the
theory of vector lattices which will simplify life greatly.

Let us proceed by describing some further collections of experiments which
all contain any experiment having a countable parameter set.

In order to describe the first collection we shall need the statistical distance
between probability distributions $P$ and $Q$ on a common measurable space
$(\mathcal{X}, \mathcal{A})$. This distance is the total variation $\|P - Q\| = 2 \sup \{P(A) - Q(A) : A \in \mathcal{A}\}$
of the difference measure $P - Q$. This important quantity will enter into
several of our evaluations. For the time being just note that $\|P - Q\|/4 + \frac{1}{2}$ is
the maximum probability of guessing the true distribution in the Bayes
estimation problem where the prior distribution assigns probability $\frac{1}{2}$ to $P$
and probability $\frac{1}{2}$ to $Q$.

Consider now an experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$. Then the map $(\theta_1, \theta_2) \sim \rightarrow \|P_{\theta_1} - P_{\theta_2}\|$
defines a pseudo-metric on $\Theta$ (‘pseudo’ because the map $\theta \sim \rightarrow$
$P_\theta$ may not be $1 - 1$). The parameter set $\Theta$ together with this pseudo-metric
becomes a pseudo-metric space. We may then classify experiments with
respect to the properties of these spaces. Thus we may call an experiment
totally bounded, compact or separable according to whether this pseudo-metric
space is, respectively, totally bounded, compact or separable.

Suppose $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ is separable and that $\{\theta_1, \theta_2, \ldots\}$ is a dense
countable subset of $\Theta$. Put $\pi = \sum_{i=1}^{\infty} 2^{-i} P_{\theta_i}$. Then $\pi$ is a probability distribution on $\mathcal{A}$
such that each $P_\theta$ is absolutely continuous with respect to $\pi$. It follows then, by the Radon–Nikodym theorem, that each distribution $P_\theta$
has a density with respect to $\pi$. Experiments $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ such that each
probability measure $P_\theta$ is absolutely continuous with respect to some common
$\sigma$-finite measure $\mu$ are called dominated. If $\mathcal{E}$ is dominated then, as we shall see
later, the dominating measure may always be chosen as a probability measure
$\pi = \sum_{\theta = 1}^{\infty} 2^{-i} P_{\theta_i}$, where $\theta_1, \theta_2, \ldots \in \Theta$. More generally a family $(\mu_\theta : \theta \in \Theta)$ of
finite measures on a common measurable space is called dominated if there is
a non-negative and $\sigma$-finite measure $\mu$ such that each $\mu_\theta$ is absolutely con-
tinuous with respect to $\mu$.

A much stronger condition than domination is the requirement that $\mathcal{E} =$
$(P_\theta : \theta \in \Theta)$ is homogeneous i.e. that the distributions $P_\theta : \theta \in \Theta$ are mutually
absolutely continuous.
Suppose now that $\mathcal{E}$ is dominated by, say, a probability measure $\pi$. If the $\sigma$-algebra $\mathcal{A}$ has a countable basis, then the space of $\pi$-integrable functions is separable for the $\pi$-mean of absolute differences as the distance. Identifying each probability measure $P_\theta$ with its density with respect to $\pi$ we find, since
\[
\|P_\theta - P_\theta\| = \int|dP_\theta/d\pi - dP_\theta/d\pi|d\pi,
\]
that $\mathcal{E}$ is separable for the total variation distance. It follows that an experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ where $\mathcal{A}$ has a countable basis is dominated if and only if it is separable.

An interesting characterization of separability within the collection of dominated experiments due to Pfanzagl (1969) is as follows. Suppose $\mathcal{E}$ is dominated by the $\sigma$-finite measure $\mu$. Let $\Theta$ have the measurability structure given by the $\sigma$-algebra consisting of all subsets. Then $\mathcal{E}$ is separable if and only if $[dP_\theta/d\mu]_x$ may be specified to be jointly measurable in $(\theta, x)$.

Piecing ‘disjoint’ dominated experiments together we obtain a still more general collection of experiments. The experiments $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ obtained in this way should have the following two properties:

(i) there is a measurable partition $\{\mathcal{X}_t : t \in T\}$ of $\mathcal{X}$ such that a subset $A$ of $\mathcal{X}$ is measurable whenever $A \cap \mathcal{X}_t$ is measurable for each $t$.

(ii) for each $\theta \in \Theta$ and each $t \in T$, let $P_\theta t$ be the restriction of $P_\theta$ to measurable subsets of $\mathcal{X}_t$. Then, for each $t \in T$, the family $(P_\theta : \theta \in \Theta)$ of measures on the $\sigma$-algebra of measurable subsets of $\mathcal{X}_t$ is dominated.

Experiments satisfying (i) and (ii) are called $\Sigma$-finite by LeCam. Since it is not usual to call dominated experiments finite we have found it more natural to use the term $\Sigma$-dominated for experiments satisfying (i) and (ii).

Any dominated experiment is clearly $\Sigma$-dominated.

If in the description above we assume that each $\mathcal{X}_t$ is a one-point set and if it is assumed that each $P_\theta$ has a countable support, then we obtain another subcollection of the collection of $\Sigma$-dominated experiments. These experiments are called discrete. Thus a discrete experiment is an experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ where $\mathcal{A}$ is the class of all subsets of $\mathcal{X}$ and where each $P_\theta$ has a countable support.

**Example 1.2.1** (Experiments associated with sampling plans). Consider a population $\Pi$ which is an enumerable set and which may be any such set. Suppose also that there is a characteristic which, with varying amount (value, degree, …) is possessed by all individuals in $\Pi$. Let $\theta(i)$ be the amount of this characteristic for individual $i \in \Pi$. The function $\theta$ on $\Pi$ defined this way is our parameter of interest. We shall assume that it is a priori known that $\theta$ belongs to, and may be any element of, a set $\Theta$ of functions on $\Pi$.

In order to find out about $\theta$ we may take a sample from $\Pi$ and measure the characteristic for each of the individuals in the sample. An essential assumption is now that the sampling is carried out according to a known sequence sampling plan $\alpha$, i.e. a probability distribution on the space $\Pi$, of finite
Statistical experiments within the measure theoretical framework

sequences of elements from \( \Pi \). Before proceeding let us agree that a probability measure on an enumerable set is defined for all subsets. In order to retain the possibility of making no observations at all, we may include the ‘empty’ sequence \( \emptyset \) in \( \Pi \). If the sequence sampling plan \( \pi \) is used and if the characteristics of the sampled individuals are measured without errors, then the outcome \((i_1, \theta(i_1)), \ldots, (i_n, \theta(i_n))\) is obtained with probability \(\pi(i_1, \ldots, i_n)\). Thus we may let our sample space consist of all sequences \((i_1, f_1), (i_2, f_2), \ldots, (i_n, f_n)\) where \((i_1, \ldots, i_n) \in \Pi\), \(f_1, \ldots, f_n \in \bigcup_\theta \theta[\Pi]\) and where \(f_i = f_i'\) whenever \(i_i = i_i'\).

Let \(P_{\theta, \pi}\) denote the probability distribution of the outcome when \(\theta\) prevails and \(\alpha\) is used. Thus the sequence sampling plan \(\pi\) determines a discrete statistical experiment \(\sigma_\pi = (P_{\theta, \pi} : \theta \in \Theta)\). We may call \(\sigma_\pi\) the sequence sampling plan experiment determined by \(\alpha\).

Let \((I_1, F_1), \ldots, (I_n, F_n)\) be the random outcome and consider the statistics \(U\) and \(G\) where \(U = \{I_1, \ldots, I_n\}\) and \(G\) is the function on the set \(U\) determined by \(F\). Now \(P_{\theta, \pi}(i_1, f_1), \ldots, (i_n, f_n) = \alpha(i_1, \ldots, i_n)\) or \(= 0\) as \((f_1, \ldots, f_n) = (\theta(i_1), \ldots, \theta(i_n))\) or not. It is easily checked that conditional probabilities given \((U, G)\) may be specified independently of \(\theta\). Thus \((U, G)\) is sufficient. The important thing is that the reduction by sufficiency leads to another, and equivalent, discrete experiment \(\tilde{\sigma}_\pi = (\tilde{P}_{\theta, \pi} : \theta \in \Theta)\) which may be described as follows.

Let \(\mathcal{U}\) be the class of all finite subsets of \(\Pi\). If \(u \in \mathcal{U}\) and \(\alpha\) is a sequence sampling plan on \(\Pi\) then \(\bar{\alpha}\) is the probability distribution on \(\mathcal{U}\) induced from \(\alpha\) by the set-valued map \((i_1, \ldots, i_n) \mapsto \{i_1, \ldots, i_n\}\). Thus \(\bar{\alpha}\) is the probability distribution of the sampled subset of \(\Pi\).

We may then let the sample space, \(\mathcal{F}\), of \(\tilde{\sigma}_\pi\) consist of all pairs \((u, g)\) where \(u \in \mathcal{U}\) and \(g = \theta|u\) for some \(\theta \in \Theta\). If \(\alpha\) is used then the probability \(\tilde{P}_{\theta, \pi}((u, g))\) of the outcome \((u, g)\) is \(\bar{\alpha}(u)\) or \(= 0\) as \(g = \theta|u\) or not.

It follows that the structure of experiments \(\sigma_\pi\) may be identified with a structure of probability measures on the set of finite subsets of the population \(\Pi\).

In general a probability distribution \(\bar{\alpha}\) on the enumerable set of finite subsets of \(\Pi\) may be called a set sampling plan. (All classes of finite subsets are then considered measurable.)

Any set sampling plan \(\bar{\alpha}\) determines a set sampling plan experiment \(\tilde{\sigma}_\pi\) according to the recipe just given.

The problem of comparing experiments determined by sampling plans is treated in Torgersen (1982) and Milbrodt (1985).

Note that if \(\sigma\) is discrete then each \(P_\theta\) has a density with respect to the counting measure on \(\mathcal{A}\). In general, if \(\sigma = (P_\theta : \theta \in \Theta)\) is \(\Sigma\)-dominated then, as noted above, there are always non-negative measures \(\mu\) on \(\mathcal{A}\) such that each \(P_\theta\) has a density with respect to \(\mu\). This may be seen as follows. Suppose \(\sigma\) satisfies
1.2 The measure theoretical framework

(i) and (ii) in the definition of $\Sigma$-dominatedness. We may assume that for each $t$ there is a point $\theta \in \Theta$ such that $P_\theta(\mathcal{X}_t) > 0$. Then for each $t$ there is a probability measure $\mu_t$ on $\mathcal{A}$ such that $P_\theta(A \cap \mathcal{X}_t) \equiv_{t} 0$ when $\mu_t(A) = 0$. We may assume that $\mu_t(\mathcal{X} - \mathcal{X}_t) \equiv_{t} 0$. Then each distribution $P_\theta$ has a density with respect to the measure $\mu = \sum_t \mu_t$. Note that $\mu$ is $\sigma$-finite if and only if $T$ is countable.

A non-negative measure $\mu$ is majorizing an experiment $\mathcal{E} = (P_\theta : \theta \in \Theta)$ if each distribution $P_\theta$ is given by a density with respect to $\mu$. If so then we shall say that $\mathcal{E}$ is majorized. Thus, as we have just seen, $\Sigma$-dominated experiments are majorized.

A very useful tool for proving the existence of decision procedures with specified properties is the so-called weak compactness lemma. Let us consider an example where this tool is not available.

**Example 1.2.2 (The set of power functions is not necessarily closed).** Let $\mathcal{E} = ([0, 1], \text{Borel class}, P_0 : \theta \in [0, 1])$ where $P_0$ is the uniform distribution on $[0, 1]$ and where $P_\theta$ is the one-point distribution in $\theta$ when $\theta > 0$. Consider the problem of testing the null hypothesis ‘$\theta > 0$’ against the alternative ‘$\theta = 0$’. If we had the additional information that $\theta$ belonged to some specified countable subset $C$ of $\Theta$ containing both 0 and positive points $\theta$ then the non-randomized test $\delta_C$ with rejection region $[0, 1] - C$ has power 1 and level of significance 0. Consider now the class $\mathcal{C}$ of countable subsets of $\Theta$ which contain 0. Then $\mathcal{C}$ is directed by inclusion so that $\delta_C = 1 - I_C = C \in \mathcal{C}$ is a net (= generalized sequence) of test functions. Clearly

$$\lim_{C} E_\theta \delta_C = I_{\{0\}}(\theta); \quad \theta \in \Theta.$$  

Thus the pointwise limit of the power functions looks like the ‘ideal’ power function. However it is not the power function of any test. On the contrary $\sup_{\theta} |\beta(\theta)| - \lim_{C} E_\theta \delta_C | \geq \frac{1}{2}$ for any power function $\beta$.

An experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ such that $P_{\theta_1}$ and $P_{\theta_2}$ are mutually singular (i.e. $\|P_{\theta_1} - P_{\theta_2}\| = 2$) when $\theta_1 \neq \theta_2$ is called totally informative. The experiments considered in the last example were clearly totally informative. In a sense, which will be made precise later, any totally informative experiment is as informative as observing the unknown parameter $\theta$ itself. ‘Pathologies’ as those in the last example will then be ‘defined away’ by the principle that whatever can be achieved by restricting attention to an arbitrarily large finite subset of $\Theta$ should also be achievable on all of $\Theta$. If, however, we want to discuss fine points concerning the measurability structure of an experiment, then the totally informative experiments exhibit by and large the same variability as there is within the totality of all experiments.

The example shows that if we want the class of power functions to be closed
for pointwise convergence, then we may either generalize our notion of statistical decision procedure or we may limit our attention to experiments satisfying a compactness condition.

An experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ will be called coherent if to each uniformly bounded net $(\delta_j)$ of real valued variables there corresponds a subnet $(\delta_j)$ and a real variable $\delta$ such that $\int \delta_j h \, dP_\theta \to \int \delta h \, dP_\theta$ when $\int |h| \, dP_\theta < \infty$.

Thus the experiment $\mathcal{E}$ considered in example 1.2.1 is not coherent.

Before proceeding to a study of coherent experiments let us consider some basic results on weak compactness and uniform integrability.

Let $(\mathcal{X}, \mathcal{A}, \pi)$ be a probability space and let $X_t : t \in T$ be a family of real valued variables. Then we shall say that the family $\{X_t : t \in T\}$ is uniformly integrable if $\int_{|X_t| > c} |X_t| \, d\pi \to 0$ as the constant $c \to \infty$ uniformly in $t$. If $\{X_t : t \in T\}$ is uniformly integrable then the $L_1$-norms, $\int |X_t| \, d\pi$, are uniformly bounded by any of the finite constants $c + \sup_t \int_{|X_t| > c} |X_t| \, d\pi$. A family $\{X_t : t \in T\}$ of integrable real variables is called uniformly absolutely continuous if $\int_A X_t \, d\pi \to 0$ as $\pi(A) \to 0$ uniformly in $t$.

It follows from the inequality $\int_A X_t \, d\pi \leq c \pi(A) + \sup_t \int_{|X_t| > c} |X_t| \, d\pi$ that any uniformly integrable family of real variables is also uniformly absolutely continuous. Suppose conversely that the family $\{X_t : t \in T\}$ is uniformly absolutely continuous and that $\sup \int |X_t| \, d\pi < \infty$. Then, since $\pi(|X_t| > c) \leq c^{-1} \sup_t \int_{|X_t| > c} |X_t| \, d\pi$, the integrals $\int_{|X_t| > c} |X_t| \, d\pi \to 0$ as $c \to \infty$ uniformly in $t$. Altogether we have proved the following proposition.

**Proposition 1.2.3.** A family $(X_t : t \in T)$ of real valued random variables is uniformly integrable if and only if it is uniformly absolutely continuous and $\sup_t E|X_t| < \infty$.

Some basic results on weak compactness are collected in the following.

**Theorem 1.2.4 (Weak compactness theorem).** Let $(\mathcal{X}, \mathcal{A}, \pi)$ be a probability space and, as usual, let $L_1 = L_1(\pi)$ and $L_\infty = L_\infty(\pi)$ denote the equivalence classes of, respectively, integrable functions and bounded measurable functions. Equip $L_1$ with its $L_\infty$-topology, i.e. the topology where convergence of a net $(X_t : t \in T)$ of integrable variables to an integrable variable $X$ takes place if and only if $\int X_t Y \, d\pi \to \int XY \, d\pi$ whenever $y \in L_\infty$.

Then the following three conditions on a set $V$ of integrable variables are equivalent:

(i) $V$ is uniformly integrable.

(ii) the closure of $V$ is compact.

(iii) each sequence $X_1, X_2, \ldots$ in $V$ has a convergent sub-sequence.

We shall not prove this theorem here. We refer to, e.g. section IV.2 in Neveu (1965), for a complete proof. See also section V.6 in Dunford & Schwarz (1958).