Assorted background material for Nonparametric Statistical Theory

Comments and corrections to r.samworth@statslab.cam.ac.uk

Basic measure theory definitions

Definition: A *σ-algebra* $\mathcal{E}$ on an arbitrary set E is a set of subsets of E such that for all $A \in \mathcal{E}$ and all sequences (A_n) in $\mathcal{E}$, we have

$$\emptyset \in \mathcal{E}, \quad A^c \in \mathcal{E}, \quad \text{and} \quad \bigcup_{n=1}^{\infty} A_n \in \mathcal{E}.$$

The pair $(E, \mathcal{E})$ is called a *measurable space*, and each $A \in \mathcal{E}$ is called a *measurable set.*

Definition: A *measure* is a function $\mu : \mathcal{E} \to [0, \infty]$ satisfying $\mu(\emptyset) = 0$, and such that for all disjoint sequences (A_n) in $\mathcal{E}$, we have

$$\mu\biggl(\bigcup_{n=1}^{\infty} A_n\biggr) = \sum_{n=1}^{\infty} \mu(A_n).$$

The triple $(E, \mathcal{E}, \mu)$ is then called a *measure space.* The measure is *finite* if $\mu(E) < \infty$, and *σ-finite* if we can find a countable collection of sets (E_n) with $\mu(E_n) < \infty$ and $\cup_{n=1}^{\infty} E_n = E$.

Example 1. *Let $E = \mathbb{Z}$, and let $\mathcal{E}$ denote the set of all subsets of E. The function $\mu : \mathcal{E} \to [0, \infty]$ defined by $\mu(A) = \mathrm{card}(A)$ (so $\mu(A)$ is the number of elements in A if this is finite, and is ∞ otherwise) is a measure, called* counting measure*. This is clearly a σ-finite measure, but not a finite measure. In a very similar way, we can define counting measure on $\mathbb{Z}^n$.*

Example 2. *Let $E = \mathbb{R}$, and let $\mathcal{E}$ denote the smallest σ-algebra containing all the open sets (called the* Borel *σ-algebra). Then $\mathcal{E}$ is a proper subset of the set of all subsets of $\mathbb{R}$, but contains all sets of the form*

$$(a, b), \quad [a, b), \quad (a, b], \quad [a, b],$$

for $a \leq b$, as well as countable unions of such intervals, their complements (e.g. the irrationals) and so on. In general, one cannot write down explicitly what a general Borel set looks like. Nevertheless, there exists a unique measure μ on $\mathcal{E}$, called Lebesgue measure*, satisfying*

$$\mu\bigl((a, b)\bigr) = \mu\bigl((a, b]\bigr) = \mu\bigl([a, b)\bigr) = \mu\bigl([a, b]\bigr) = b - a.$$

Again, this is a σ-finite measure, but not a finite measure. We can also define Lebesgue measure on $\mathbb{R}^n$; it satisfies

$$\mu\bigl((a_1, b_1) \times \ldots \times (a_n, b_n)\bigr) = \prod_{i=1}^{n} (b_i - a_i).$$

If $\mu(E) = 1$, then we say μ is a *probability measure*, and usually use the notation $(\Omega, \mathcal{F}, \mathbb{P})$ instead of $(E, \mathcal{E}, \mu)$.

Exercise: Show that $[a, b]$ for $a \leq b$ belongs to the Borel σ-algebra on $\mathbb{R}$.

Definition: Given two measurable spaces $(E, \mathcal{E})$ and $(G, \mathcal{G})$, a function $f : E \to G$ is *measurable* if for every $B \in \mathcal{G}$, we have $f^{-1}(B) \in \mathcal{E}$. Here, $f^{-1}(B) = \{x \in E : f(x) \in B\}$. In particular, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then a *random variable* is a measurable function $Y : \Omega \to \mathbb{R}$. A *random vector* is a measurable function $Y : \Omega \to \mathbb{R}^d$, and a random element of E is a measurable function $Y : \Omega \to E$. The *distribution* (or *law*) of a random variable Y is the measure P on the Borel subsets of $\mathbb{R}$ given by $P(A) = \mathbb{P}(Y \in A) \equiv \mathbb{P}(Y^{-1}(A))$.

You should think of measurability as a very weak requirement. In particular, if (f_n) is a sequence of measurable, real-valued functions, then so are the sum $f_1 + f_2$, the product $f_1 f_2$, as well as:

$$\inf_{n \in \mathbb{N}} f_n, \quad \sup_{n \in \mathbb{N}} f_n, \quad \liminf_{n \to \infty} f_n, \quad \limsup_{n \to \infty} f_n.$$

Integration

Let $(E, \mathcal{E}, \mu)$ be a measure space. We want to define the integral of a measurable function $f : E \to [-\infty, \infty]$, to be denoted

$$\mu(f) = \int_E f(x)\, d\mu(x) = \int_E f\, d\mu = \int_E f(x) \mu(dx).$$

In the special case where μ is Lebesgue measure, we often write $\int_E f(x)\, dx$. Note that we use the same $\mu(\cdot)$ to denote the measure of a set and the integral of a function – we will see below that $\mu(A) = \mu(\mathbb{1}_A)$. However, on a probability space by convention we usually write $\mathbb{E}(Y)$, not $\mathbb{P}(Y)$ for the integral of a random variable Y.

Definition: A non-negative *simple* function is one of the form $f = \sum_{k=1}^m a_k \mathbb{1}_{A_k}$, where $a_k \in [0, \infty]$ and $A_k \in \mathcal{E}$, and we define the integral of such a function by

$$\mu(f) = \sum_{k=1}^m a_k \mu(A_k).$$

For a general, non-negative measurable function f, we define

$$\mu(f) = \sup\{\mu(g) : g \leq f,\ g \text{ simple}\}.$$

For measurable functions f that are not necessarily non-negative, we write $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. If $\mu(|f|) < \infty$, we say f is *integrable*, and define $\mu(f) = \mu(f^+) - \mu(f^-)$.

In practice, this definition is rarely used to compute integrals directly. In the measure space of Example 5, note that $\mu(f) = \sum_{n=-\infty}^{\infty} f(n)$, while in the measure space of Example 6, we can evaluate integrals using the Fundamental Theorem of Calculus.

Theorem 3 (Monotone convergence). *Let (f_n) be a sequence of non-negative measurable functions on E, with $f_n(x) \nearrow f(x)$ for all $x \in E$. Then $\mu(f_n) \nearrow \mu(f)$ as $n \to \infty$.*

Theorem 4 (Fatou's lemma). *Let (f_n) be a sequence of non-negative measurable functions on E. Then*

$$\mu\big(\liminf_{n\to\infty} f_n\big) \leq \liminf_{n\to\infty} \mu(f_n).$$

Theorem 5 (Dominated convergence). *Let (f_n) be a sequence of measurable functions with $f_n \to f$ pointwise as $n \to \infty$. If there exists an integrable function g such that $|f_n(x)| \leq g(x)$ for all $x \in E$, then $\mu(f_n) \to \mu(f)$ as $n \to \infty$.*

Remark: In the statements of both the monotone and dominated convergence theorems, it is enough that the convergence of f_n to f occurs for *almost all* $x \in E$ (i.e., except on a set of μ-measure zero).

Exercise: For each of the following sequences of measurable functions $f_n : \mathbb{R} \to \mathbb{R}$, find the pointwise limit f if it exists, and state with justification whether or not $\int f_n(x)\,dx \to \int f(x)\,dx$ as $n \to \infty$:

$$(a) f_n(x) = \mathbb{1}_{[n,n+1]}(x), \quad (b) f_n(x) = (-1)^n, \quad (c) f_n(x) = x^{-1/2}\mathbb{1}_{[1/n,1]}(x),$$
$$(d) f_n(x) = x^{-1}\mathbb{1}_{[1/n,1]}(x), \quad (e) f_n(x) = \frac{e^{-|x|}}{n}\sin(nx).$$

Exercise: Let (Y_n) be a sequence of random variables with $\mathbb{E}(|Y_n|) \leq 1$ and $Y_n \to Y$ pointwise as $n \to \infty$. Is it necessarily the case that $\mathbb{E}(|Y|) \leq 1$? (*Hint: Fatou*).

Order notation

Let $u(x)$ be a (possibly vector-valued) function and let $v(x) > 0$. We write $u(x) = O(v(x))$ as $x \to x_0$ if there exists $C > 0$ such that

$$\frac{\|u(x)\|}{v(x)} \leq C$$

for all x sufficiently close to x_0. We write $u(x) = o(v(x))$ as $x \to x_0$ if

$$\frac{u(x)}{v(x)} \to 0$$

as $x \to x_0$. Usually, $u(x)$ and $v(x)$ are sequences indexed by n, say, and the limit is as $n \to \infty$.

If (Y_n) is a sequence of random vectors and (a_n) is a sequence of positive constants, we write $Y_n = O_p(a_n)$ as $n \to \infty$ if, given $\epsilon > 0$, there exists $C > 0$ such that

$$\mathbb{P}\bigg(\frac{\|Y_n\|}{a_n} > C\bigg) < \epsilon$$

for all sufficiently large n. We write $Y_n = o_p(a_n)$ if $Y_n/a_n \overset{p}{\to} 0$.

Example. If (Y_n) is a sequence of independent and identically distributed random variables with mean and finite variance, then $\sum_{i=1}^{n}(Y_i - \mu) = O_p(n^{1/2})$, by the central limit theorem.

Convergence

Definition: We say a sequence of random vectors (Y_n) converges almost surely to Y, and write $Y_n \overset{a.s.}{\to} Y$, if $\mathbb{P}(Y_n \to Y) = 1$; equivalently, if for every $\epsilon > 0$,

$$\mathbb{P}\biggl(\sup_{m \geq n} \|Y_m - Y\| > \epsilon\biggr) \to 0$$

as $n \to \infty$.

Definition: We say (Y_n) converges in probability to Y, and write $Y_n \overset{p}{\to} Y$, if for every $\epsilon > 0$,

$$\mathbb{P}(\|Y_n - Y\| > \epsilon) \to 0$$

as $n \to \infty$.

Definition: We say (Y_n) converges in distribution to Y, and write $Y_n \overset{d}{\to} Y$, if $\mathbb{E}\{f(Y_n)\} \to \mathbb{E}\{f(Y)\}$ for all bounded, continuous, real-valued functions f. In fact, it is enough that the convergence occurs when f is bounded and Lipschitz i.e. there exists $L > 0$ such that $|f(x) - f(y)| \leq L\|x - y\|$. Equivalently, $Y_n \overset{d}{\to} Y$ if and only if

$$\mathbb{P}(Y_n \leq y) \to \mathbb{P}(Y \leq y)$$

at all points where the distribution function of Y is continuous.

Theorem 6. $Y_n \overset{a.s.}{\to} Y \implies Y_n \overset{p}{\to} Y \implies Y_n \overset{d}{\to} Y$.

Theorem 7 (Strong law of large numbers)**.** *If (Y_n) are independent and identically distributed with finite mean μ, then $n^{-1}\sum_{i=1}^{n} Y_i \overset{a.s.}{\to} \mu$.*

Theorem 8 (Multidimensional central limit theorem)**.** *If (Y_n) are independent and identically distributed in $\mathbb{R}^d$ with mean vector μ and covariance matrix Σ, then*

$$n^{1/2}(\bar{Y}_n - \mu) \overset{d}{\to} N_d(0, \Sigma),$$

where $\bar{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$.

Exercise: Prove this theorem using the univariate central limit theorem and the Cramér–Wold device, which says that $Y_n \overset{d}{\to} Y$ if and only if $t^T Y_n \overset{d}{\to} t^T Y$ for all $t \in \mathbb{R}^d$.

Theorem 9 (Mapping theorems)**.** *Let g be a continuous, real-valued function.*

(i) If $Y_n \overset{a.s.}{\to} Y$, then $g(Y_n) \overset{a.s.}{\to} g(Y)$

(ii) If $Y_n \xrightarrow{p} Y$, then $g(Y_n) \xrightarrow{p} g(Y)$

(iii) If $Y_n \xrightarrow{d} Y$, then $g(Y_n) \xrightarrow{d} g(Y)$

In fact, g may have a set of discontinuities D_g, provided that $\mathbb{P}(Y \in D_g) = 0$.

Theorem 10 (Slutsky's theorem)**.** *Let (Y_n) and (Z_n) be sequences of random vectors with $Y_n \xrightarrow{d} Y$ and $Z_n \xrightarrow{p} c$, where c is constant. If g is a continuous real-valued function, then $g(Y_n, Z_n) \xrightarrow{d} g(Y, c)$.*

Theorem 11 (Lindeberg–Feller central limit theorem)**.** *For each $n \in \mathbb{N}$, let $\{Y_{ni} : i = 1, \ldots, n\}$ be independent and identically distributed random variables with $\mathrm{Var}(Y_{n1}) \to \sigma^2$ as $n \to \infty$. If, for every $\epsilon > 0$, we have*

$$\mathbb{E}(Y_{n1}^2 \mathbb{1}_{\{|Y_{n1}| \geq \epsilon n^{1/2}\}}) \to 0$$

as $n \to \infty$, then

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} \{Y_{ni} - \mathbb{E}(Y_{ni})\} \xrightarrow{d} N(0, \sigma^2).$$

The delta method

Theorem 12 (The delta method)**.** *Let (Y_n) be a sequence of random vectors in $\mathbb{R}^d$ such that for some $\mu \in \mathbb{R}^d$ and a random vector Z, we have $n^{1/2}(Y_n - \mu) \xrightarrow{d} Z$. If $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable at μ, then*

$$n^{1/2}\{g(Y_n) - g(\mu)\} \xrightarrow{d} \nabla g(\mu)^T Z.$$

Proof. We give the proof first in the (simpler) case $d = 1$. Write $g'(\mu) = \nabla g(\mu)$ and define $h : \mathbb{R} \to \mathbb{R}$ by

$$h(y) = \begin{cases} \frac{g(y)-g(\mu)}{y-\mu} & \text{if } y \neq \mu \\ g'(\mu) & \text{if } y = \mu. \end{cases}$$

Note that h is continuous at μ. Thus, by the continuous mapping theorem and Slutsky's theorem,

$$n^{1/2}\{g(Y_n) - g(\mu)\} = h(Y_n) n^{1/2}(Y_n - \mu) \xrightarrow{d} g'(\mu) Z.$$

For the multidimensional case, given $\epsilon > 0$, choose $K > 0$ such that $\mathbb{P}(n^{1/2}\|Y_n - \mu\| > K) \leq \epsilon$ for all $n \in \mathbb{N}$. Now choose $\delta > 0$ such that

$$|g(y) - g(\mu) - \nabla g(\mu)^T (y - \mu)| \leq \frac{\epsilon}{K} \|y - \mu\|$$

for $\|y - \mu\| \leq \delta$. Then, for every $n \in \mathbb{N}$ large enough that $\mathbb{P}(\|Y_n - \mu\| > \delta) \leq \epsilon$, we have

$$\begin{aligned} &\mathbb{P}\big(|n^{1/2}\{g(Y_n) - g(\mu)\} - \nabla g(\mu)^T n^{1/2}(Y_n - \mu)| > \epsilon\big) \\ &\qquad \leq \mathbb{P}(n^{1/2}\|Y_n - \mu\| > K) + \mathbb{P}(\|Y_n - \mu\| > \delta) \leq \epsilon + \epsilon = 2\epsilon. \end{aligned}$$

We conclude using Slutsky's theorem that

$$n^{1/2}\{g(Y_n) - g(\mu)\} = \nabla g(\mu)^T n^{1/2}(Y_n - \mu) + o_p(1) \xrightarrow{d} \nabla g(\mu)^T Z,$$

as required. □