APPENDIX D - BASIC PROBABILITY THEORY

Probability theory has an axiomatization, as follows. There is a set, say X, called the *sample space*. An observation will be a point x of X. Probabilities will be defined for some subsets of X, called *events*. The collection $\mathcal{B}$ of all events will be assumed to be a *σ-algebra*, in other words, to satisfy the following conditions:

(a) The empty set $\emptyset$ and X are in $\mathcal{B}$;

(b) The complement $A^c := X \setminus A$ is the set of all x in X not in A. If A is in $\mathcal{B}$, so is A^c.

(c) For any sequence $A_1, A_2, ...$ of sets in $\mathcal{B}$, the union $\bigcup_{n=1}^{\infty} A_n$, in other words the set of all x such that $x \in A_n$ for some n, is also in $\mathcal{B}$.

It follows easily from the definition that any intersection of σ-algebras of subsets of X is a σ-algebra of subsets of X. The collection 2^X of all subsets of X is a σ-algebra. Thus for any collection $\mathcal{A}$ of subsets of X, there is a smallest σ-algebra including $\mathcal{A}$, called the σ-algebra *generated* by $\mathcal{A}$, namely, the intersection of all σ-algebras including $\mathcal{A}$, one of which is 2^X. If X is the real line $\mathbb{R}$ then an important σ-algebra of subsets of X is the *Borel* σ-algebra generated by the collection of all open intervals (a, b) for $a < b$ in $\mathbb{R}$.

If X is a set and $\mathcal{B}$ is a σ-algebra of subsets of X then $(X, \mathcal{B})$ is called a *measurable space*.

If $(X, \mathcal{B})$ is a measurable space then a function μ on $\mathcal{B}$ is called a *measure* if:

(d) $0 \leq \mu(A) \leq +\infty$ for all $A \in \mathcal{B}$;

(e) $\mu(\emptyset) = 0$;

(f) For any sequence $A_1, A_2, ...$, of sets in $\mathcal{B}$ which are disjoint, in other words $A_i \cap A_j = \emptyset$ whenever $i \neq j$, we have $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.

Then $(X, \mathcal{B}, \mu)$ is called a *measure space*.

A main example of a measure space is given by $X = \mathbb{R}$, with $\mathcal{B}$ as the Borel σ-algebra, and μ as *Lebesgue measure* λ, which equals the length for intervals.

A measure space $(X, \mathcal{B}, \mu)$ is called a *probability space* if and only if $\mu(X) = 1$. Then μ is often written as P, or as Q if two probability measures are considered, or sometimes as Pr.

If $(X, \mathcal{B})$ is any measurable space, then a real-valued function f on X is called *measurable* if for any Borel set $A \subset \mathbb{R}$, $f^{-1}(A) := \{x : f(x) \in A\} \in \mathcal{B}$. It turns out to be equivalent that for any $t \in \mathbb{R}$, $f^{-1}((t, \infty)) := \{x : f(x) > t\} \in \mathcal{B}$.

If X is a countable set such as the set $\mathbb{N}$ of nonnegative integers, then the usual σ-algebra on X will be the collection 2^X of all its subsets. A measure μ on such a set will be called *discrete*. Then μ of any set is given by a sum, $\mu(A) = \sum_{x \in A} \mu(\{x\})$, where $\{x\}$ is the set whose only member is x. Lebesgue measure λ, on the other hand, is not given by such sums, since $\lambda(\{x\}) = 0$ for all x. On X, we have the *counting measure* c where $c(A)$ is the number of elements of A if A is finite and $c(A) = \infty$ if A is infinite. Any probability measure P on the countable set X has a density f with respect to counting measure, which in this case is called a *probability mass function*. Thus $P(A) = \sum_{x \in A} f(x)$ for any set $A \subset X$.

For Lebesgue measure one can define an integral for suitable functions, $\int f d\lambda = \int f(x) dx$. In a very analogous way, one can define the integral $\int f d\mu$ for suitable functions

f on a measure space. See RAP, Chapters 3 and 4. Specifically, if f is any nonnegative, measurable function, then $\int f d\mu$ is always defined but may be $+\infty$. If f is real-valued let $f^+ := \max(f, 0)$ and $f^- := -\min(f, 0)$, so $f^+ \geq 0$, $f^- \geq 0$, and $f \equiv f^+ - f^-$. A measurable function f is said to be *integrable* if and only if both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite, and then $\int f d\mu$ is defined as $\int f^+ d\mu - \int f^- d\mu$.

In probability theory, if $(\Omega, \mathcal{B}, P)$ is a probability space then a *random variable* on Ω will be a real-valued, measurable function on Ω, often called X, Y, etc. The *expectation* or *mean* of X is defined by $EX = \int X dP$ if X is either nonnegative or integrable.

Suppose f is a nonnegative, measurable function on $\mathbb{R}$ with $\int_{-\infty}^{\infty} f(x)dx = 1$. Then f is called a *probability density* on $\mathbb{R}$. A probability measure P with density f is defined by setting $P(A) := \int_A f(x)dx := \int 1_A f \, d\lambda$ for any Borel set A, where $1_A(x) := 1$ for $x \in A$ and 0 for $x \notin A$.

If P has density f and g is a random variable whose expectation is defined, then it can be shown that $Eg = \int g(x)f(x)dx$ ("The law of the unconscious statistician"). This follows directly from the definition if $g = 1_A$ for some Borel set A. It then follows for any finite linear combination of such functions, and then by usual methods of Lebesgue integral theory (monotone convergence, e.g. RAP, Proposition 4.1.5 and Theorem 4.3.2) for any nonnegative measurable function, thus for g^+ and g^-, then from the definitions whenever Eg is defined.

Here are three (families of) examples of probability densities on $\mathbb{R}$:
(i) For any constant $c > 0$, an *exponential density* is defined by $f(x) = ce^{-cx}$ for all $x \geq 0$ and $f(x) = 0$ for $x < 0$. Let P be the probability measure with density f and let X be the identity function, $X(x) = x$ for all x. Then by integration by parts one can check that $EX = 1/c$.
(ii) For any $\mu \in \mathbb{R}$ and $0 < \sigma < \infty$, the *normal density* for the probability distribution $N(\mu, \sigma^2)$ is defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

If X has this density it is not hard to show that $EX = \mu$.
(iii) The *Cauchy density* on $\mathbb{R}$ is defined by $f(x) := 1/(\pi(1 + x^2))$ for all x. Let P be the probability measure with this density and again let X be the identity function on $\mathbb{R}$. Then EX is not defined because $EX^+ = EX^- = +\infty$.

Some notations for orders of magnitude, O, o, O_p and o_p. For two functions f and g of a real variable, $f = O(g)$ or $f(x) = O(g(x))$ as $x \to +\infty$ will mean that $g(x) > 0$ for x large enough and f/g remains bounded as $x \to +\infty$. Here "O" seems to refer to "order of magnitude;" $f = O(g)$ means that f is of the same order of magnitude as g, or smaller. $f = O(g)$ can also be defined for any other limiting behavior of x, for example, $f = O(g)$ as $x \downarrow 0$ means that $g(x) > 0$ for $x > 0$ small enough and f/g remains bounded as $x \downarrow 0$. $f = O(1)$ thus means in either case that f remains bounded.

On the other hand, $f(x) = o(g(x))$ as $x \to +\infty$ will mean that $g(x) > 0$ for x large enough and $f(x)/g(x) \to 0$ as $x \to +\infty$. $f = o(g)$ as $x \downarrow 0$ is defined analogously. Thus $f = o(g)$ means that f is of smaller order of magnitude than g under a given limiting behavior of x.

Both O and o notations are also defined for functions of an integer n as $n \to +\infty$.

Now for orders of magnitude in probability, let X_n be a sequence of random variables. Recall that X_n is said to converge to 0 *in probability* if for every $\varepsilon > 0$, $P(|X_n| > \varepsilon) \to 0$ as $n \to \infty$. Then $X_n = o_p(1)$ will mean that $X_n \to 0$ in probability. A sequence Y_n of random variables will be said to be *bounded in probability* if for every $\varepsilon > 0$ there is an $M < \infty$ such that $P(|Y_n| > M) < \varepsilon$ for all n. Then one writes $Y_n = O_p(1)$.

If X_n and Y_n are two sequences of random variables, then $X_n = O_p(Y_n)$ will mean that $Y_n > 0$ almost surely for n large enough and $X_n/Y_n = O_p(1)$, in other words, X_n/Y_n is bounded in probability. $X_n = o_p(Y_n)$ will mean that $Y_n > 0$ almost surely for n large enough and $X_n/Y_n = o_p(1)$, in other words, $X_n/Y_n \to 0$ in probability.

NOTE

The axiomatization of probability became widely known from and is generally attributed to a book by A. N. Kolmogorov published in 1933. Actually the axiomatization appeared a bit earlier in much less known publications, one by Kolmogorov himself and one by S. Ulam. See RAP, notes to section 8.1.

The O and o notations in analysis were apparently invented by the British mathematician G. H. Hardy. I don't know who first defined the O_p and o_p notations.