Summary of basic probability theory, part 1

D. Joyce, Clark University

Math 218, Mathematical Statistics, Jan 2008

Sample space. A *sample space* consists of a *underlying* set S, whose elements are called *outcomes*, a collection of subsets of S called *events*, and a function P on the set of events, called a *probability* function, satisfying the following axioms.

1. The probability of any event is a number in the interval $[0, 1]$.

2. The entire set S is an event with probability $P(S) = 1$.

3. The union and intersection of any finite or countably infinite set of events are events, and the complement of an event is an event.

4. The probability of a disjoint union of a finite or countably infinite set of events is the sum of the probabilities of those events,

$$P(\bigcup_i E_i) = \sum_i P(E_i).$$

From these axioms a number of other properties can be derived including these.

5. The the complement $\overline{E} = S - E$ of an event E is an event, and

$$P(\overline{E}) = 1 - P(E).$$

6. The empty set is an event with probability $P(\emptyset) = 0$.

7. For any two events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F),$$

therefore

$$P(E \cup F) \leq P(E) + P(F).$$

8. For any two events E and F,

$$P(E) = P(E \cap F) + P(E \cap \overline{F}).$$

9. If event E is a subset of event F, then $P(E) \leq P(F)$.

10. Statement 7 above is called the *principle of inclusion and exclusion.* It generalizes to more than two events.

$$\begin{aligned} P(\bigcup_{r=1}^{n} E_r) &= \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) \\ &+ \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - \cdots \\ &+ (-1)^{n-1} P(E_1 \cap E_2 \cap \cdots \cap E_n) \end{aligned}$$

In words, to find the probability of a union of n events, first sum their individual probabilities, then subtract the sum of the probabilities of all their pairwise intersections, then add back the sum of the probabilities of all their 3-way interections, then subtract the 4-way intersections, and continue adding and subtracting k-way intersections until you finally stop with the probability of the n-way intersection.

Random variables notation. In order to describe a sample space, we frequently introduce a symbol X called a *random variable* for the sample space. With this notation, we can replace the probability of an event, $P(E)$, by the notation $P(X \in E)$, which, by itself, doesn't do much. But many events are built from the set operations of complement, union, and intersection, and with the random variable notation, we can replace those by logical operations for 'not', 'or', and 'and'. For instance, the probability $P(E \cup \overline{F})$ can be written as $P(X \in E \text{ but } X \notin F)$.

Also, probabilities of finite events can be written in terms of equality. For instance, the probability of a singleton, $P(\{a\})$, can be written as $P(X{=}a)$, and that for a doubleton, $P(\{a,b\}) = P(X{=}a \text{ or } X{=}b)$.

One of the main purposes of the random variable notation is when we have two uses for the same sample space. For instance, if you have a fair die, the sample space is $S = \{1,2,3,4,5,6\}$ where the probability of any singleton is $\frac{1}{6}$. If you have two fair dice, you can use two random variables, X and Y, to refer to the two dice, but each has the same sample space. (Soon, we'll look at the joint distribution of (X,Y), which has a sample space defined on $S \times S$.

Random variables and cumulative distribution functions. A sample space can have any set as its underlying set, but usually they're related to numbers. Often the sample space is the set of real numbers $\mathbf{R}$, and sometimes a power of the real numbers $\mathbf{R}^n$.

The most common sample space only has two elements, that is, there are only two outcomes. For instance, flipping a coin as two outcomes—Heads and Tails; many experiments have two outcomes—Success and Failure; and polls often have two outcomes—For and Against. Even though these events aren't numbers, it's useful to replace them by numbers, namely 0 and 1, so that Heads, Success, and For are identified with 1, and Tails, Failure, and Against are identified with 0. Then the sample space can have $\mathbf{R}$ as its underlying set.

When the sample space does have $\mathbf{R}$ as its underlying set, the random variable X is called a *real random variable*. With it, the probability of an interval like $[a,b]$, which is $P([a,b])$, can then be described as $P(a \le X \le b)$. Unions of intervals can also be described, for instance $P((-\infty,3) \cup [4,5])$ can be written as $P(X < 3 \text{ or } 4 \le X \le 5)$.

When the sample space is $\mathbf{R}$, the probability function P is determined by a cumulative distribution function (c.d.f.) F as follows. The function $F : \mathbf{R} \to \mathbf{R}$ is defined by

$$F(x) = P(X \le x) = P((-\infty,x]).$$

Then, from F, the probability of a half-open interval can be found as

$$P((a,b]) = F(b) - F(a).$$

Also, the probability of a singleton $\{b\}$ can be found as a limit

$$P(\{b\}) = \lim_{a \to b} (F(b) - F(a)).$$

From these, probabilities of unions of intervals can be computed. Sometimes, the c.d.f. is simply called the *distribution*, and the sample space is identified with this distribution.

Discrete distributions. Many sample distributions are determined entirely by the probabilities of their outcomes, that is, the probability of an event E is

$$P(E) = \sum_{x \in E} P(X{=}x) = \sum_{x \in E} P(\{x\}).$$

The sum here, of course, is either a finite or countably infinite sum. Such a distribution is called a *discrete distribution*, and when there are only finitely many outcomes x with nonzero probabilities, it is called a *finite distribution*.

A discrete distributions is usually described in terms of a probability mass function (p.m.f.) f defined by

$$f(x) = P(X{=}x) = P(\{x\}).$$

This p.m.f. is enough to determine this distribution since, by the definition of a discrete distribution, the probability of an event E is

$$P(E) = \sum_{x \in E} f(x).$$

In many applications, a finite distribution is *uniform*, that is, the probabilities of its outcomes are all the same, $1/n$, where n is the number of outcomes with nonzero probabilities. When that is the case, the field of combinatorics is useful in finding probabilities of events. Combinatorics includes various principles of counting such as the multiplication principle, permutations, and combinations.

Continuous distributions. When the cumulative distribution function F for a distribution is differentiable function, we say it's a *continuous distribution*. Such a distribution is determined by a probability density function f. The relation between F and f is that f is the derivative F' of F, and F is the integral of f.

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Conditional probability and independence. If E and F are two events, with $P(F) \neq 0$, then the *conditional probability* of E given F is defined to be

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Two events, E and F, neither with probability 0, are said to be *independent*, or *mutually independent*, if any of the following three logically equivalent conditions holds

$$\begin{aligned} P(E \cap F) &= P(E)\, P(F) \\ P(E|F) &= P(E) \\ P(F|E) &= P(F) \end{aligned}$$

Bayes' formula is useful to invert conditional probabilities. It says

$$\begin{aligned} P(F|E) &= \frac{P(E|F)\, P(F)}{P(E)} \\ &= \frac{P(E|F)\, P(F)}{P(E|F)\, P(F) + P(E|\overline{F})\, P(\overline{F})} \end{aligned}$$

where the second form is often more useful in practice.