Probability and Statistics *

Paul C. Kainen
Department of Mathematics
Georgetown University
Washington, D.C. 20057-1233

DRAFT Feb. 7, 2005; 10:30 pm

1 Introduction

These notes are rather sketchy. But at least they should be readable. Those who are up to the challenge can gain extra credit by pointing out the errors! The notes are only in partial form now but they will be extended as time goes on.

2 Probability and distributions, Part I

In the following, S is a finite set but later we'll deal with infinite sets. Think of S as the set of "outcomes" of some "experiment." The outcomes are elementary indivisible results of the experiment; we'll make this more precise through examples below. The set S must be nonempty.

Let $S \neq \emptyset$ be a finite set. A function P which maps the subsets of S to real numbers will be called *probability* provided it satisfies the following three conditions:

(i) $0 \leq P(A) \leq 1$ for all subsets A of S;
(ii) $P(S) = 1$;
(iii) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ if $A_1 \cap A_2 = \emptyset$ for all subsets A_1, A_2 of S.

The third condition is called *additivity* of probability.

*notes for Math 040

A family $A_1, \ldots, A_k$ of subsets of B will be called a *partition* of B provided each distinct pair of sets in the partition are disjoint and B is their union; that is, if $A_i \cap A_j = \emptyset$ when $i \neq j$ and $B = A_1 \cup \ldots \cup A_k := \bigcup_{i=1}^{k} A_i$. This is denoted $B = A_1 + \ldots + A_k$. We call a partition *nontrivial* provided that each A_i has nonzero probability.

Using mathematical induction, one can show that (iii) can be extended to

(iv) If $B = A_1 + \ldots + A_k$ is any partition, then $P(B) = P(A_1) + \ldots P(A_k)$.

Since $A \cup B = (A - B) + A \cap B + (B - A)$ (draw a Venn diagram), by (iv) $P(A \cup B) = P(A - B) + P(A \cap B) + P(B - A)$. But $A = (A - B) \cup (A \cap B)$ and similarly $B = (B - A) \cup (A \cap B)$ so

(v)$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

It is also convenient to note that by (ii) and (iii)

(vi)$P(S - A) := P(A') = 1 - P(A)$.

In particular, $P(\emptyset) = 0$. Also, it is a good exercise to check that we have

(vii)If $A \subseteq B$, then $P(A) \leq P(B)$.

This property is usually called *monotonicity* of probability. In particular, $P(A \cap B) \leq P(A) \leq P(A \cup B)$ for all subsets A, B of S.

Note that if $S = A_1 + \ldots + A_k$, then for any subset B of S,

$$B = B \cap A_1 + \ldots + B \cap A_k;$$

that is, a partition of sample space induces one on any subset. Indeed, each b in B is also in S so $b \in A_i$ for a unique i, $1 \leq i \leq k$. Hence, $b \in B \cap A_i$ for this value of i and no other.

The notion of probability corresponds to the intuitive idea that, if the experiment were repeated many times, each outcome $s \in S$ would have some typical frequency of occurrence which corresponds to $P(\{s\})$, denoted $P(s)$ for convenience. By (iv) if $A \subseteq S$ and $A = \{s_1, \ldots, s_k\}$ for some k, then $P(A) = P(s_1) + \ldots + P(s_k) := \sum_{i=1}^{k} P(s_i)$.

When S is given with a probability P defined on its subsets, we call S a *sample space*. Subsets of sample space are called *events*.

Suppose that it is known that some event has occurred. How should this fact affect the probability of other events? We capture this notion with the definition of *conditional probability*. If B is an event with nonzero probability, then the conditional probability $P(A|B)$ of A given B is defined by

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

Intuitively, conditional probability amounts to restricting the sample space by discarding all outcomes to the experiment which do not lie in the subset B. The conditional probability of A given B only involves the portion of A which is also in B.

Two events are said to be independent provided that the probability of their intersection is equal to the product of the separate event probabilities. That is, for A and B subsets of S, A and B are *independent* if and only if

$$P(A \cap B) = P(A)P(B).$$

Since intersection is symmetric, independence is a symmetric relation. Also, if A, B are independent, then A, B', A', B, and A', B' are also independent (exercise).

Note that if either A or B is empty (or merely has zero probability), then their intersection $A \cap B$ has the same property using the fact that a subset of the empty set is empty (or the monotonicity property (vii)). Hence, when either A or B has zero probability, then the two events are independent. But this case is usually not very interesting.

On the other hand, if neither A nor B has zero probability, we can consider both $P(A|B)$ and $P(B|A)$. Then A and B are independent if and only if (iff) $P(A|B) = P(A)$ iff $P(B|A) = P(B)$.

Suppose $S = A_1 + \ldots + A_k$ is a nontrivial partition (that is, each A_i has nonzero probability). Then

$$P(B) = P(B \cap A_1) + \ldots + P(B \cap A_k) = P(B|A_1)P(A_1) + \ldots + P(B|A_k)P(A_k);$$

in words: the probability of an event B is the weighted average of the conditional probabilities $P(B|A_i)$ where the A_i, $1 \leq i \leq k$, are a partition of sample space and the weights are the probabilities of the A_i.

Bayes' theorem allows one to make inferences for any A about $P(A|B)$ provided that one knows the conditional probabilities $P(B|A_1), \ldots, P(B|A_k)$, where $S = A_1 + \ldots + A_k$ is a nontrivial partition, A is one of the members A_i of the partition, and that one also knows the probabilities $P(A_i)$. In that case, we have *Bayes formula*:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \ldots + P(B|A_k)P(A_k)}$$

3 Random variables and distribution functions

A function $X : S \to \Re$ from a sample space S to the real numbers is called a *random variable*. Random variables (r.v. for short) correspond to measurements made on the outcomes of an experiment. While functions are *deterministic* (that is, they always give the same result when given the same input), the input to the function corresponding to an r.v. is the outcome of an experiment and so the output value is *stochastic* (i.e., has a probabilistic or "random" nature).

Random variables, being functions from S to $\Re$, can be added, subtracted, and multiplied, and they can be composed with other functions from $\Re$ to $\Re$. A special type of r.v. with many uses is the 0/1 r.v. which takes on the values of either 1 or 0.

If X is a r.v. on S, we define the (probability) *distribution function* $f(x)$ corresponding to X as follows: For each real number x, let

$$f(x) = P(\{s \in S : X(s) = x\} := P(X = x);$$

this is just the probability that the r.v. X assumes the value x. (Random variables are also sometimes denoted by boldface letters so one would then have $f(x) = P(\mathbf{x} = x)$. For convenience, we will refer to the probability distribution function of some random variable X as the pdf of X.

Note that the pdf f of X is a nonnegative function defined for all real numbers and the sum of all its nonzero values is 1. Indeed, probabilities are always nonnegative (by property (i)). Since S is finite and X is a function, $f(x) = P(X = x)$ has a nonzero value for only finitely many x, say $x_1, \ldots, x_k$, so $S = (X = x_1) + \ldots + (X = x_k)$ and, hence (by properties (ii) and (iv)), we have

$$1 = P(S) = f(x_1) + \ldots + f(x_k).$$

The *cumulative distribution function* of a random variable X is the function $F(t) = P(\{s \in S : X(s) \leq t\} := P(X \leq t)$. Hence, for any real number t, $F(t)$ is the sum of $f(x)$ for all $x \leq t$; this sum only involves a finite set of nonzero summands. (Later we will consider *continuous* r.v. where the corresponding notion will involve an integral.)

Let $\{x_1, \ldots, x_k\}$ denote the set of values which are achieved on some nonempty subset of S by a r.v. X. The *expected value* of X is the number

$$\mathbf{E}(X) := x_1 f(x_1) + \ldots + x_k f(x_k) := \sum_{i=1}^{k} P(X = x_i) x_i$$

which is the sum of the random variable's values weighted by the probability of the value occurring. Expected value is also called mean value and is sometimes denoted μ_X.

Here are two easily proved facts about expected value:

(i) $\mathbf{E}(X + b) = \mathbf{E}(X) + b$

(ii) $\mathbf{E}(aX) = a\mathbf{E}(X)$

Indeed, adding a constant to a random variable has the effect of adding the constant averaged over the probabilities of the random variable achieving its possible values. But averaging a constant is the same constant. Multiplying the values by a fixed factor a clearly also multiplies their average by the same constant.

Putting (i) and (ii) together, we have

(iii) $\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$.

Hence, if $a = 0$, then the expectation of a random variable with constant value on all members of the sample space is just the constant value. Note that the existence of expected value is not a problem for finite sample spaces but this is not the case when the sample space is infinite - as we shall see later.

The notation $aX + b$ means the random variable with value $aX(s) + b$ on the element s in sample space. One can also define random variables by adding, subtracting, or multiplying other random variables. For example, for X, Y r.v. on S, $(X + Y)(s) = X(s) + Y(s)$.

We now consider when these operations can be interchanged with the expectation operator $\mathbf{E}$.

(iv) $\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y)$.

Suppose X is a r.v. with values $\{x_1, \ldots, x_k\}$ and Y is another r.v. on the same sample space S and Y has values $\{y_1, \ldots, y_t\}$. Writing $f(x_i, y_j)$ for the probability of the event that $X = x_i$ and simultaneously $Y = y_j$; that is, $f(x_i, y_j) = P(\{s \in S : X(s) = x_i, Y(s) = y_j\})$, it is straightforward to show that for each fixed i, $f(x_i) = \sum_{j=1}^{t} f(x_i, y_j)$ and for each fixed j, $f(y_j) = \sum_{i=1}^{k} f(x_i, y_j)$.

We now prove (iv) as follows.

$$\mathbf{E}(X) + \mathbf{E}(Y) = \sum_i x_i f(x_i) + \sum_j y_j f(y_j)$$

$$= \sum_i \sum_j [x_i f(x_i, y_j) + y_j f(x_i, y_j)] = \sum_i \sum_j (x_i + y_j) f(x_i, y_j) = \mathbf{E}(X + Y).$$

Two random variables X and Y are *independent* if for all i, j, we have $f(x_i, y_j) = f(x_i)g(y_j)$; that is, if each pair of events $X = x_i$, $Y = y_j$ are independent. Then we have

(v) $var(X + Y) = var(X) + var(Y)$ provided that X and Y are independent.

More generally, if we define the *covariance* of two random variables X, Y by the formula

$$cov(X, Y) = \mathbf{E}(X - \mu_X)\mathbf{E}(Y - \mu_Y).$$

Then we have

(v)' $var(X + Y) = var(X) + var(Y) + cov(X, Y)$.

This follows from the next result since $cov(X, Y) = E(XY) - E(X)E(Y)$, and the multiplicative property for expectation of independent random variables can be verified by a calculation.

(vi) $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$ provided that X and Y are independent.

We consider an example. Suppose you roll a fair die and toss a fair coin. Let X be the value shown on the die (so the pdf $f(x)$ of X is equal to $1/6$ if and only if x is one of the first six positive integers and otherwise $f(x) = 0$. Let Y be the value of the coin (say, 0 for Tails and 1 for Heads). Since the probability that X takes on some value is independent of the probability that Y takes on a given value, the two r.v. are independent. What is the expected value of XY.

4 The binomial distribution

Let $\binom{n}{k} := \frac{n!}{k!(n-k)!}$, where $0 \leq k \leq n$ and k, n are integers. Also, let $C(n, k)$ be the number of distinct k-element subsets in a set with n elements.

(i) $\binom{n}{k} = C(n, k)$.

When n is zero, this is certainly true since both numbers turn out to be 1: there is only one subset of the empty set (namely, the empty set), and the formula 0!/0!0! gives 1 (by definition, 0! = 1).

We proved in class that $C(n,k) + C(n,k+1) = C(n+1,k+1)$ since all $k+1$-element subsets of a set S with $n+1$ elements either contain "Bob" or not. Those that do are determined by choosing the remaining k elements from all the n members of S which are not Bob, while those subsets of S that don't contain Bob are determined by choosing $k+1$ elements from these non-Bob elements. Since the two collections of subsets are disjoint (either Bob is in one of them or not) and they include all subsets, we are done.

It is also possible to prove that the corresponding recursion holds for the binomial coefficients by using algebra. Indeed,

$$\binom{n}{k} + \binom{n}{k+1} = \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} = \frac{(k+1)(n!) + (n-k)(n!)}{(k+1)!(n-k)!};$$

and the last fraction is $\binom{n+1}{k+1}$.

Since $\binom{n}{k} = C(n,k)$ for $n = 0$, the recursion tells us that it also holds for $n = 1$; this can also be checked directly since $C(n,1) = n = \binom{n}{1}$ and $C(n,0) = 1 = \binom{n}{0}$. More generally, the recursion is exactly what one needs to use induction to prove that (i) holds for all integers n and k if $0 \leq k \leq n$. So the algebraic and combinatorial definitions agree.

It is convenient also to define $\binom{n}{k} := 0$ if $k > n$ or if $n < 0$. There are ways to extend these functions to much more general kinds of numbers and these numbers themselves have many special properties. However, we don't have time to explore them here.

There are three other discrete distributions worth noting. The geometric distribution was described in class and so was the uniform distribution. The remaining distribution is the Poisson distribution.

5 Probability and distributions, Part II

We've already seen two situations where an infinite sample space is required - for the geometric and Poisson distributions. In this section, we'll consider both the exponential and normal distributions where the sample space is not only infinite but "uncountable" (that is, the samples involve a continuum of choices). If one were describing an experiment involving visual perception, for example, one might need to quantify the outcome of an experiment by having the subject specify an apparent color. More later on this example and the topic ...

A probability function on an infinite sample space is not necessarily defined for every possible subset but only for those which are "measurable." Where it is defined, such a probability function must satisfy the same three axioms as in the discrete case. As an example, for a continuous r.v., the event $\{s \in S : a < X < b\} := (a < X < b)$ is measurable.

Continuous r.v. are defined in terms of their pdf's. A function $f : \Re \to \Re$ is a continuous pdf if it is (i) piece-wise continuous, (ii) always nonnegative, and (iii) the integral of f over the entire real line is unity:

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Given a sample space S with a probability measure, a function $X : S \to \Re = (-\infty, +\infty)$ is called a continuous r.v. with pdf f provided that there exists a continuous pdf f such that for all a, b,

$$P(a < X < b) = \int_a^b f(x)dx.$$

All our previous results and definitions go through, with a few minor adjustments, for continuous r.v. as well as discrete r.v.

For example, for a continuous r.v. X with pdf f we have

$$\mathbf{E}(X) := \int_{-\infty}^{+\infty} xf(x)dx, \text{ and}$$

$$var(X) := \mathbf{E}((x - \mathbf{E})^2) = \mathbf{E}(x^2) - (\mathbf{E}(x))^2.$$

There are three examples we will consider: the uniform, the exponential, and the normal distributions. The uniform pdf with parameters $a < b$ is the function

$$u(x) = \frac{1}{b-a}, a \le x \le b; u(x) = 0, \text{ else}.$$

This is a pdf since the area under the curve is a rectangle of area 1. A r.v. X is said to be uniformly distributed if it has $u(x; a, b)$ for some pair of real numbers $a < b$. It is not difficult to show that

(i) $E(X) = \frac{a+b}{2}$ when X has pdf $u(x; a, b)$.

Indeed, $\int_a^b x \cdot \frac{1}{b-a}dx = \frac{1}{b-a}x^2/2|_a^b = \frac{1}{2(b-a)}(b^2 - a^2) = \frac{b+a}{2}$. This argument uses the fundamental theorem of calculus and the fact that the derivative of $x^2/2$ is x. To find the variance of X, first we calculate $E(X^2) = \frac{1}{3(b-a)}x^3|_a^b = (1/3)(b^2 + ab + a^2)$. Hence,

(ii) $var(X) = (1/3)(b^2 + ab + a^2) - (1/4)(b^2 + 2ab + a^2) = \frac{1}{12}(b-a)^2$ when X has pdf $u(x; a, b)$.

This is reasonable since it says that the wider the interval $[a, b]$ in which the uniform r.v. can take its values, the larger the variance. Since standard deviation is the square root of the variance, we see that the SD for the uniform distribution depends linearly on the range $b - a$.

The second example of a r.v. with continuous pdf is the so-called *exponential* distribution which is defined by setting $f(x; \lambda) := 0$ for $x < 0$, and for $x > 0$ $f(x; \lambda) := \lambda e^{-\lambda x}$, where $\lambda > 0$ is a parameter. This is a pdf since $f(x; \lambda)$ is piece-wise continuous, nonnegative (it is either equal to zero or it is the product of two positive numbers), and has integral 1 over the real line since $\int_0^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}|_0^\infty = 1 - \lim_{t\to\infty} e^{-\lambda t} = 1$.

We can calculate the mean and SD for r.v. with this distribution.

(iii) If X has exponential pdf $f(x; \lambda)$, then $E(X) = \frac{1}{\lambda} = \sigma_X$.

The argument for the mean value uses integration-by-parts and l'Hospital's rule, and the derivation of the standard deviation is similar.

Using calculus techniques to evaluate simple integrals, we can actually calculate probabilities for exponentially distributed r.v. For example,

(iv) $P(X > x) = 1 - \int_0^x \lambda e^{-\lambda x} dx = e^{-\lambda x}$.

It can be shown that this is a rather good fit to the likelihood that electronic equipment will not fail prior to some time x.

Now suppose you are given an exponential distribution and you know that the expected time-to-failure for certain components is, say, 2000 hours. What is the probability that one of these components lasts more than 4000 hours? Using (iii), $\lambda = 1/2000$ (for t in hours) so by (iv) $P(X > 4000) = e^{-2}$. Similarly, if a laser has an exponentially distributed lifetime with mean 400 hours, there is a probability of e^{-2} that it will last more than 800 hours and a probability of $1 - e^{-3}$ that the laser will fail during its first 1200 hours of operation. I hope you see how easy it is to use these properties of the exponential distribution to actually calculate stuff. Once we have the basic facts (which need calculus for derivation), we can use them in a simple algebraic way.

The third distribution is the normal distribution. More on this one later.