PROBABILITY

Paul Taylor (Trinity)

In the classical view of probability an experiment has a finite set, Ω, of outcomes, ω (eg $\Omega=\{1,2,3,4,5,6\}$ for a fair die), which are equally likely. An event is a subset, A, of the sample space (eg $A=\{1,3,5\}$, the event that the roll, ω, is odd). A occurs if $\omega\in A$. Ω and $\emptyset$ are the certain and impossible events, and $\Omega\setminus A$ is the event not A. Also if A,B are events, $A\cap B$ and $A\cup B$ are the events A and B and A or B. The probability of A is $P(A)=\#A/\#\Omega$, so

(α) $0=P(\emptyset)\le P(A)\le P(\Omega)=1$

(β) $A\cap B=\emptyset \Rightarrow P(A\cup B)=P(A)+P(B)$.

If $A\cap B=\emptyset$, A and B are (mutually) exclusive; a set of events $A_1,A_2,\ldots,A_n$ such that $A_i\cap A_j=\emptyset$ if $i\neq j$, and $A_1\cup A_2\cup\ldots\cup A_n=\Omega$ is a partition of Ω. If $A\subseteq B$, we say A implies B, and then $P(A)\le P(B)$.

If, however, we wish to model irrational probabilities or an infinite sample space, the classical view breaks down. Instead we adopt an axiomatic view. We have a sample space, Ω, and a collection of subsets, $\{A\}$, called events, to which we assign probabilities satisfying (α) and (β). Examples:

(i) $\Omega=\mathbb{N}$, $\omega\in\Omega$ is the number of α-particles from a radioactive source detected by a Geiger tube in one minute. A typical event is $A=\{1,2,\ldots\}$, that at least one particle is detected.

(ii) $\Omega=\mathbb{R}^+$, $\omega\in\Omega$ is the height in metres of an undergraduate. An event might be $A=\{\omega: 1.5<\omega<1.6\}$.

(iii) $\Omega=2^{\mathbb{N}}=\{\omega=(a_1,a_2,\ldots): a_i\in\{H,T\}\}$, where $\omega\in\Omega$ is an infinite sequence of tosses of a fair coin. An event might be $A=\{\omega: a_1=H\}$, that heads is tossed first time.

In the first example the probability of an event is the sum of the probabilities of the points, but this is not so in the other two cases: indeed the points are unrealisable and have probability zero (they are almost impossible). In general we consider only events, not points.

From the axioms we deduce $P(\Omega\setminus A)=1-P(A)$ and $P(A\cup B)=P(A)+P(B)-P(A\cap B)$. More generally we have the inclusion-exclusion principle:

for events $A_1,A_2,\ldots,A_n$,

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^n P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

If B is an event with $P(B) > 0$, the conditional probability of A given B is

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

If $B_1, B_2, \ldots, B_n$ is a partition of Ω, $P(A) = \sum_i P(A|B_i) P(B_i)$. Also, (Bayes' theorem),

$$P(B_i|A) = \frac{P(A|B_i) P(B_i)}{\sum_j P(A|B_j) P(B_j)}$$

A random variable on a space Ω is just a function, $X: \Omega \to \mathbb{R}$. Then if B is a subset of $\mathbb{R}$,

$$P(X \in B) = P(\omega : \omega \in \Omega \text{ and } X(\omega) \in B)$$

Examples: number of sixes in n rolls of a die; the square of the number of α-particles decaying in one minute; the upper density of the frequency of tosses of heads in an infinite sequence ($\limsup H_n/n$ as $n \to \infty$, where H_n is the number of heads thrown in the first n throws). Any real-valued function of a random variable is also a random variable.

If Ω is finite or countable, X has a probability distribution $P(x_i) = P(X = x_i) = P(\{\omega \in \Omega : X(\omega) = x_i\})$.

If $\Omega = \mathbb{R}$, X has a distribution function: $F(x) = P(X \le x) = P(\{\omega \in \Omega : X(\omega) \le x\})$.

Notice that F is monotone increasing, so if $x \le y$, $0 \le F(x) \le F(y) \le 1$. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$. F is also right-continuous, so if $x \to x_0+$, $F(x) \to F(x_0)$.

A discrete random variable has step-function distribution, a continuous one has continuous distribution.

If $f: \mathbb{R} \to \mathbb{R}$ is a function such that $F(x) = \int_{-\infty}^{x} f(x')\,dx'$, it is called the density function of X. A discrete random variable may be said to have δ-function distribution [see Linear Systems].

If A is an event, the indicator function, I_A, is an example of a random variable, where $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ if $\omega \notin A$. We have

$$I_{A \cap B} = I_A I_B, \quad I_{\Omega \setminus A} = 1 - I_A \text{ and } I_{A \cup B} = 1 - (1 - I_A)(1 - I_B),$$

thus providing a neat proof of the inclusion-exclusion principle.

The expectation or mean of a random variable on a discrete space is

$$E(X) = \sum_i X(\omega) P(\omega)$$

and on a continuous space, $E(X) = \int_{-\infty}^{\infty} X(x) f(x)\,dx$, so long as the sum or integral converges absolutely.

otherwise X has infinite expectation. The expectation is linear, so if X, Y are random variables and $\lambda, \mu \in \mathbb{R}$ are constants,

$$E(\lambda X + \mu Y) = \lambda E(X) + \mu E(Y).$$

If $X = a$ with probability 1, $E(X) = a$ also, and if $X \geq 0$, $E(X) \geq 0$. $E(X)$ also has the important property that it is the constant, a, which minimises $E(X-a)^2$.

The expectation of an indicator function is the probability of the event, $E(I_A) = P(A)$, so this could provide an alternative definition of probability.

The n^{th} moment of X is $E(X^n)$ if this exists. The n^{th} absolute moment is $E(|X|^n)$. The second moment about the mean, $E((X-E(X))^2 = E(X^2) - (E(X))^2$ is called the variance, $V(X)$. If $\lambda \in \mathbb{R}$, $V(\lambda X) = \lambda^2 V(X)$ and $V(X+\lambda) = V(X)$. Finally $V(X) \geq 0$ with equality iff X is constant with probability 1. The positive square root of $V(X)$ is called the standard deviation.

If X, Y are random variables on a discrete space, their joint distribution is the set of probabilities $P(X = x_i, Y = y_j) = p_{ij}$. $P(X = x_i) = \sum_j p_{ij} = p_{i\cdot}$ is called the marginal distribution of X (similarly Y). If we fix y_j, the conditional distribution of X given $Y = y_j$ is $P(X = x_i | Y = y_j) = p_{ij} / p_{\cdot j}$ and the conditional expectation is

$$E(X | Y = y_j) = \sum_i \frac{x_i p_{ij}}{p_{\cdot j}}$$

This is itself a random variable, with $E(E(X|Y)) = E(X)$. $E(X|Y)$ is that function $a(Y)$ which minimises $E((X - a(Y))^2)$. Moments are defined in the obvious way, $E(X^r Y^s)$.

The covariance of X, Y is

$$\mathrm{cov}(X,Y) = E((X-E(X))(Y-E(Y))) = E(XY) - E(X)E(Y).$$

It satisfies $V(X+Y) = V(X) + V(Y) + 2\,\mathrm{cov}(X,Y)$.

Notice that $V(X) = \mathrm{cov}(X,X)$; the matrix $(a_{ij}) = (\mathrm{cov}(X_i, X_j))$ is called the (variance-) covariance matrix: if $Y = \sum c_i X_i$, $V(Y) = \sum_{ij} c_i a_{ij} c_j$.

The correlation coefficient is

$$\mathrm{corr}(X,Y) = \mathrm{cov}(X,Y) / [V(X)V(Y)]^{1/2};$$

it always lies between -1 and $+1$.

Two random variables are said to be independent if for any sets of real numbers, $A, B \subseteq \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A)\,P(Y \in B).$$

In this case their covariance is zero, so $E(XY) = E(X)E(Y)$ and $V(X+Y) = V(X) + V(Y)$, but the converse does NOT hold.

A sequence of random variables $X_1, X_2, \ldots, X_n$ is <u>totally independent</u> if, for any $A_1, A_2, \ldots, A_n \subseteq \mathbb{R}$,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1) P(X_2 \in A_2) \ldots P(X_n \in A_n).$$

This implies that they are pairwise independent, but NOT conversely (take $X_1, X_2, X_3 = X_1 + X_2$).

If $g_1, g_2, \ldots, g_n$ are arbitrary functions, $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are also totally independent.

We frequently discuss <u>independent, identically-distributed</u> random variables (IIDs): these correspond to the idea of a repeated trial. If $X_1, X_2, \ldots, X_n$ are IID then $V(X_1 + X_2 + \ldots + X_n) = n V(X_1)$ and the standard deviation increases as $\sqrt{n}$.

If $(a_0, a_1, \ldots)$ is a sequence of real numbers, the function $f(t)$ [for such values of t as the sum converges] is called the <u>generating function</u> of (a_n).

$$f(t) = \sum_{n=0}^{\infty} a_n t^n$$

If it exists, $f(t)$ uniquely determines the sequence.
If $(a_n), (b_n)$ are sequences, their <u>convolution</u> is (c_n), where

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}$$

If $(a_n), (b_n)$ have generating functions $f(t), g(t)$, the generating function of (c_n) is $f(t)g(t)$.

If $\Omega = \mathbb{N}$, a countable sample space, and $a_i = P(i)$, $f(t)$ always exists for $|t| \leq 1$ and is called the <u>probability generating function</u>, written $p(t)$. It is analytic in $|t| < 1$ and

$$E(X) = p'(1)$$
$$V(X) = p''(1) + p'(1) - [p'(1)]^2$$

if these exist; other moments are similarly calculated.

If X, Y are independent, the probability distribution of $X+Y$ is the convolution of those of X, Y, so the probability generating function (pgf) is the product. Hence if $X_1, \ldots, X_n$ are IID, the pgf of $X_1 + \ldots + X_n$ is $p(t)^n$.

$M_X(t) = E(e^{tX})$, if it exists, is called the <u>moment generating function</u> (mgf). It too uniquely determines the probability distribution, and the n-th moment of X is the coefficient of $t^n/n!$ in $M_X(t)$. Once again if X, Y are independent, $M_{X+Y} = M_X M_Y$. Observe $M(t) = p(e^t)$.

name	example or interpretation	distribution or density	mean	variance	probability generating function	moment generating function
Bernoulli (p)	Toss of biassed coin (p=prob^y)	$\begin{cases} q=1-p & r=0 \\ p & r=1 \\ 0 & r\geq 2 \end{cases}$	p	pq	$q+pz$	$q+pe^\theta$
Binomial B(n,p)	Heads in n tosses	$\binom{n}{r}p^r q^{n-r}$	np	npq	$(q+pz)^n$	$(q+pe^\theta)^n$
Geometric (p)	Tosses to first head (inclusive)	$\begin{cases} 0 & r=0 \\ q^{r-1}p & r\geq 1 \end{cases}$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pz}{1-qz}$	$\frac{pe^\theta}{1-qe^\theta}$
Negative Binomial (n,p)	Tosses to n^{th} head (inclusive)	$\binom{-n}{r-n}p^n(-q)^{r-n} = \binom{r-1}{n-1}p^n q^{r-n}$	$\frac{n}{p}$	$\frac{nq}{p^2}$	$\left(\frac{pz}{1-qz}\right)^n$	$\left(\frac{pe^\theta}{1-qe^\theta}\right)^n$
Hypergeometric (m_1, m_2, n)	No. of white balls out of n drawn from m_1 white & m_2 black without replacement	$\frac{\binom{m_1}{r}\binom{m_2}{n-r}}{\binom{m_1+m_2}{n}}$ $0\leq r\leq m_1$	$\frac{nm_1}{m_1+m_2}$	$\frac{nm_1m_2(n-m)}{m^2(m-1)}$ $(m=m_1+m_2)$	not known in closed form	
Poisson (λ)	No. of α-decays.	$e^{-\lambda}\lambda^r/r!$	λ	λ	$e^{-\lambda(z-1)}$	$e^{\lambda(e^\theta-1)}$
Uniform U(a,b)	Uniform in [a,b]	$\begin{cases} 1/(b-a) & a\leq x\leq b \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{2}(b+a)$	$\frac{1}{12}(b-a)^2$	—	$\frac{e^{b\theta}-e^{a\theta}}{b\theta-a\theta}$
Normal $N(\mu,\sigma^2)$	Normal or Gaussian.	$\frac{e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sigma\sqrt{2\pi}}$	μ	σ^2	—	$e^{(\theta\mu+\frac{1}{2}\theta^2\sigma^2)}$
Exponential (λ)	Lifetime of a radioactive particle	$\lambda e^{-\lambda x}$ $(x\geq 0)$	$1/\lambda$	$1/\lambda^2$	—	$\frac{\lambda}{\lambda-\theta}$

We write "$X\sim B(n,p)$" for "X is a random variable with binomial distribution with parameters n,p", etc.

Some important analytical results:

Stirling's formula $\log n! = \log\sqrt{2\pi} + (n+\frac{1}{2})\log n - n + O(n^{-1})$ as $n\to\infty$

Convergence of Binomial to Poisson $np=\lambda$, fixed, for $k=0,1,2,\ldots$

$$\binom{n}{k}p^k(1-p)^{n-k} \longrightarrow e^{-\lambda}\lambda^k/k! \quad \text{as } n\to\infty$$

Chebyshev's inequality X is a random variable with finite mean $\mu=E(X)$ and variance $\sigma^2=V(X)$. For $\varepsilon>0$

$$P(|X-\mu|\geq\varepsilon) \leq \sigma^2/\varepsilon^2$$

note $E(X^2)<\infty \Rightarrow E(X)<\infty$ as $E(X^2)-E(X)^2=V(X)\geq 0$.

Weak law of large numbers for Bernoulli trials. Let S_n be the number of heads thrown in n tosses of a biassed coin (prob' p). Then for $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) \to 0 \qquad \text{as} \quad n \to \infty$$

Indeed for any IID $X_1, X_2, \ldots$ with mean μ and variance σ^2 put $S_n = \sum_1^n X_i$. Then for $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0 \qquad \text{as} \quad n \to \infty.$$

Note the natural interpretation of the expectation.

Strong Law of large numbers, corresponding [if anything does!] to the intuitive "law of averages". Let $X_1, X_2, \ldots$ be IID Bernoulli (ie coin tosses) and put $S_n = \sum_1^n X_i$. Then

$$P\left(\frac{S_n}{n} \to p\right) = 1$$

ie it is almost certain. More generally let $X_1, X_2, \ldots$ be IID with mean μ and $E(X_1^4) < \infty$. Put $S_n = \sum_1^n X_n$, then

$$P\left(\frac{S_n}{n} \to \mu\right) = 1$$

De Moivre-Laplace theorem. Let $X_1, X_2, \ldots$ be IID Bernoulli, $S_n = \sum_1^n X_i$. Let $a, b \in \mathbb{R}$ be fixed. Then

$$P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx \qquad \text{as } n \to \infty.$$

Normal approximation to Poisson Let $X(\lambda)$ be Poisson, $P(\lambda)$, $\lambda > 0$. Then for $a, b \in \mathbb{R}$, fixed,

$$P\left(a \le \frac{X(\lambda) - \lambda}{\sqrt{\lambda}} \le b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx \qquad \text{as } \lambda \to \infty$$

Note that $Y(\lambda) = (X(\lambda) - \lambda)/\sqrt{\lambda}$ is $X(\lambda)$ normalised to have mean 0 and variance 1.

Lemma Let $X_1, X_2, \ldots$ be IID whose mgf exists. Let μ, σ^2, S_n be as before. Then if $Z_n = (S_n - n\mu)/(\sqrt{n}\sigma)$, its mgf, $M_n(\theta)$, converges pointwise to that of the normal distribution. ie, $M_n(\theta) = E(e^{\theta Z_n}) \to e^{\frac{1}{2}\theta^2}$ as $n \to \infty$ for all $\theta \in \mathbb{R}$.

Central Limit theorem Let $a, b \in \mathbb{R}$ be fixed, then

$$P(a \le Z_n \le b) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx \qquad \text{as } n \to \infty.$$

The remainder of this course is about random walks, Markov chains, branching processes and Poisson processes, which are dealt with at the beginning of the IB Markov Chains course summary, by Ian White. Some notes on solving linear recurrence relations will be found appended to Potential Theory.