Mathematics 2Y Spring 1995
Probability Theory

Contents

§1. Basic concepts. Sample space, events, inclusion-exclusion principle, probabilities. Examples.

§2. Independence, conditioning, Baye's formula, law of total probability. Examples.

§3. Discrete random variables. Expectation, variance, independence. Binomial, geometric and Poisson distributions and their relationships. Examples.

§4. Probability generating functions. Compound randomness. Applications.

§5. Continuous random variables. Distribution functions, density functions. Uniform, exponential and normal distributions.

§6. Moment generating functions. Statement of Central Limit Theorem. Chebyshev's inequality and applications (including the weak law of large numbers).

§7. Markov chains. Transition matrix, steady-state probability vectors, regularity, an ergodic theorem.

§8. Birth and Death processes. Steady states. Application to telecom circuits. M/M/1 queue.

If there is time we will go on to discuss reliability. Examples on the problem sheets will include some ideas associated with simulation.

Some recommended books

G.Grimmett & D.Welsh. Introduction to probability theory. *Oxford.*

P.King. Computer and communications systems performance modelling. *Prentice-Hall.*

H.F.Mattson Jr. Discrete Mathematics. *Wiley.*

S.B.Maurer & A.A.Ralston. Discrete Algorithmic Mathematics. *Addison-Wesley.*

J.Pitman. Probability. *Springer-Verlag.*

S.Ross. A first course in probability theory. *Collier-Macmillan.*

§1. Basic concepts

In this section of the course we will introduce some basic concepts of probability theory: sample spaces, events, inclusion-exclusion principle, probabilities.

Think of modelling an experiment. There are a number of different possible outcomes to the experiment and we wish to assign a 'likelihood' to each of these. We think of an experiment as being repeatable under identical conditions.

1.1. Definition

The set of all possible outcomes of our experiment is called the *sample space*. It is usually denoted Ω.

1.2. Examples

a. Suppose we flip a coin. $\Omega = \{H, T\}$.

b. Suppose that we roll a six-sided die. $\Omega = \{1, 2, 3, 4, 5, 6\}$

c. Rolling a die twice, $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$

1.3. Definition

Any subset of the sample space is called an *event*.

1.4. Example

If I roll a fair die, the *event* that I roll an even number ($\{2, 4, 6\} \subseteq \Omega$) has *probability* one half.

Discrete probability theory is concerned with the modelling of experiments which have a finite or countable number of possible outcomes. The simplest case is when there are a finite number of outcomes all of which are equally likely to happen. (For example rolling a fair die.) In general we assign a probability ('likelihood') p_i to each element ω_i of the sample space. i.e. to each possible outcome of the experiment. The probability of the event $A = \{\omega_1, \omega_2 \ldots \omega_n\}$ is then the sum of the probabilities corresponding to the outcomes which make up the event $(p_1 + p_2 + \ldots + p_n)$.

1.5. Examples

a. For rolling a fair die we already calculated (without necessarily realising it) that

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

b. Suppose that a certain component will fail during its nth minute of operation with probability $1/2^n$. The chance that the component fails within an hour is then

$$\sum_{n=1}^{60} \frac{1}{2^n} = (1 - \frac{1}{2^{60}})$$

1.6. IMPORTANT CHECK

When you set up your model be sure that all of the probabilities are non-negative less than or equal to one and the sum of the probabilities is equal to one.

1.7. Notation

It is customary to use some notation from set-theory. The probability that both events A and B happen is written $\mathbb{P}(A \cap B)$. The probability that at least one of the events A and B happens is written $\mathbb{P}(A \cup B)$.

1.8. The principle of inclusion and exclusion

This principle looks rather daunting in full generality, so here first is the statement for $n = 2$: for events A, B

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

If we take B to be the event A does not happen (in set-theoretic notation $B = A^c$), then this says

$$1 = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

i.e.

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

In words this is just "the probability that A does not happen is one minus the probability that A does happen".

Here then is the general form of the inclusion-exclusion principle:
For events $A_1, A_2, \ldots, A_n$,

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{1 \le i \le n} \mathbb{P}(A_i) \quad &- \quad \sum_{1 \le i < j \le n} \mathbb{P}(A_i \cap A_j) \\ &+ \quad \ldots (-1)^n \mathbb{P}(A_1 \cap A_2 \cap \ldots A_n)\end{aligned}$$

You'll come across this again in the enumeration course next term.

1.9. Example

Suppose that we roll two fair dice. What is the probability that the sum of the numbers thrown is even or divisible by three (or both)?

Solution: Let A be the event that the sum is even and B be the event that the sum is divisible by three. Then $A \cap B$ is the event that the sum is divisible by six and we seek $\mathbb{P}(A \cup B)$. Using inclusion-exclusion we see

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}$$

§2. Conditioning and Independence

In this section we will discuss conditioning, Baye's formula and the law of total probability.

The idea of conditional probability is fundamental in probability theory. Suppose that I know that something has happened, then I might want to reevaluate my guess as to whether something else will happen. For example, if I know that there has been a snowstorm then I think it (even) more likely that my train will be late than I might have thought had I not heard a weather report.

A proper understanding of conditioning can save us from some bad mistakes. Here is a famous example:

2.1. Example

You visit the home of an acquaintance who says "I have two kids". From the boots in the hall you guess that at least one is a boy. What is the probability that your acquaintance has two boys?

Well before, the sample space was (in an obvious notation)

$$\{(b,b),(b,g),(g,b),(g,g)\}$$

but if there is at least one boy, we know that in fact the only possibilities are

$$\{(b,b),(b,g),(g,b)\}$$

All of these are about equally likely, so the answer to our question is about one third.

We write $\mathbb{P}(A|B)$ for the probability of the event A given that we *know* that the event B happens.

2.2. Baye's formula

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

One way to think of this is that since we know that B happens, the possible outcomes of our experiment are just the elements of B. i.e. we change our sample space. In the example above we knew we need only consider families with at least one boy.

Independence

Heuristically, two events are independent if knowing about one tells you nothing about the other. That is $\mathbb{P}(A|B) = \mathbb{P}(A)$. Usually this is written in the following equivalent way:

2.3. Definition

a. Two events A and B are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

b. Sets $A_1, A_2, \ldots, A_n$ are *pairwise independent* if for every choice of i, j with $1 \leq i < j \leq n$

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$$

c. Sets $A_1, A_2, \ldots, A_n$ are *independent* if for all choices $i_1, i_2, \ldots, i_m$ of *distinct* integers from $\{1, 2, \ldots, n\}$

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \ldots \mathbb{P}(A_{i_m})$$

PAIRWISE INDEPENDENT DOES NOT IMPLY INDEPENDENT

2.4. Example

A bitstream when transmitted has

$$\mathbb{P}(0 \text{ sent}) = \tfrac{4}{7} \qquad \mathbb{P}(1 \text{ sent}) = \tfrac{3}{7}$$

Owing to noise:

$$\mathbb{P}(1 \text{ received } | 0 \text{ sent}) = \tfrac{1}{8}$$
$$\mathbb{P}(0 \text{ received } | 1 \text{ sent}) = \tfrac{1}{6}$$

What is $\mathbb{P}(0 \text{ sent}|0 \text{ received})$?

Solution

Take the sample space to be $\Omega = \{(0,0), (0,1), (1,0), (1,1)\}$ where for example $(1,0)$ denotes (0 sent, 1 received).

$$\mathbb{P}(0 \text{ sent}|0 \text{ received}) = \frac{\mathbb{P}(0 \text{ sent and } 0 \text{ received})}{\mathbb{P}(0 \text{ received})}$$

Now

$$\mathbb{P}(0 \text{ received}) = \mathbb{P}(0 \text{ sent and } 0 \text{ received}) + \mathbb{P}(1 \text{ sent and } 0 \text{ received})$$

In our notation this is $\mathbb{P}((0,0)) + \mathbb{P}((1,0))$. Now we use Baye's formula in the slightly unfamiliar form

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

Thus

$$\begin{aligned} \mathbb{P}((0,0)) &= \mathbb{P}(0 \text{ received}|0 \text{ sent})\mathbb{P}(0 \text{ sent}) \\ &= (1 - \mathbb{P}(1 \text{ received}|0 \text{ sent}))\mathbb{P}(0 \text{ sent}) \\ &= (1 - \frac{1}{8})\frac{4}{7} = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} \mathbb{P}((1,0)) &= \mathbb{P}(1 \text{ received}|0 \text{ sent})\mathbb{P}(0 \text{ sent}) \\ &= \frac{1}{8} \cdot \frac{4}{7} = \frac{1}{14} \end{aligned}$$

Putting these together gives

$$\mathbb{P}(0 \text{ received}) = \frac{1}{2} + \frac{1}{14} = \frac{8}{14}$$

and

$$\mathbb{P}(0 \text{ sent}|0 \text{ received}) = \frac{\frac{1}{2}}{\frac{8}{14}} = \frac{7}{8}$$

We used some important ideas in the above solution. In particular, we used the following result which formalises the idea that if you can't immediately calculate a probability then split it up:

2.5. Theorem (The law of total probability)

If $B_1, B_2, \ldots$ are a finite or countable number of disjoint events (no two can happen together) whose union is all of Ω (one of them must happen) then for any event A

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \ldots \\ &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \ldots \end{aligned}$$

The law of total probability is often used in the analysis of algorithms. The general strategy is as follows. Recall that a major design criterion in the

development of algorithms is efficiency as measured by the quantity of some resource used. For example we might wish to analyse the run-time of an algorithm. Of particular interest will be the 'average' case. Given an algorithm $\mathcal{A}$

a. Identify events $E_1, E_2, \ldots, E_n$ that effect the run-time of $\mathcal{A}$

b. Find $p_i = \mathbb{P}(E_i)$

c. Find the conditional run-time $t_i|E_i$

d. The average case run-time is then $p_1t_1 + p_2t_2 + \ldots + p_nt_n$

Several times already we have used $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. This method can be extended (proof by induction) to give what I call 'the method of hurdles'.

2.6. The method of hurdles

For any events $A_1, A_2, \ldots, A_n$

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\ldots\mathbb{P}(A_n|A_1 \cap A_2 \ldots \cap A_{n-1})$$

2.7. Example

A bag contains 26 tickets–one with each letter of the alphabet. If six tickets are drawn at random from the bag (without replacement), what is the chance that they can be rearranged to spell CALVIN ?

2.8. Example

Calvin has seven pairs of socks–all different colours. Every Sunday night he washes all his socks and throws them (unpaired) into his drawer. Each morning he pulls out two socks at random from the clean socks left in the drawer. What is the chance that his socks match every weekday, but don't match at the weekend?

2.9. Some basic rules for calculating probabilities

AND $\Longrightarrow$ method of hurdles (multiplication)

OR $\Longrightarrow$ if the events are mutually exclusive then add the probabilities, otherwise try taking *complements* (the probability that at least one of $A_1, \ldots A_n$ happens is one minus the probability that none of them happens).

If you can't calculate a probability directly then try splitting it up and using the law of total probability.

§3. Discrete random variables

For our purposes discrete random variables will take on values which are natural numbers, but any countable set of values will do.

3.1. Definition

A *random variable* is a function on the sample space.

It is basically a device for transferring probabilities from complicated sample spaces to simple sample spaces where the elements are just natural numbers.

3.2. Example

Suppose that I am modelling the arrival of telephone calls at an exchange. Modelling this directly is very complicated–my sample space should include all possible times of arrival of calls and all possible numbers of calls. If instead I consider the *random variable* which counts how many calls arrive before time t (for example) then the sample space becomes $\Omega = \{0, 1, 2, \ldots\}$. We'll return to this example later.

3.3. Definitions

a. Let X be a random variable which takes values $0, 1, 2, \ldots$ with probabilities $p_0, p_1, p_2, \ldots$. (Formally, we take Ω to be the set of events $X = 0, X = 1, \ldots$.) The values p_k are called the *distribution* of X.

b. For any function f define

$$\mathbb{E}[f(X)] = \text{ 'the expectation of } f(X)\text{'} = \sum_{k=0}^{\infty} f(k)p_k$$

In particular

$$\mathbb{E}[X] = \text{ 'expectation of } X\text{'} = \sum_{k=0}^{\infty} kp_k$$

The expectation of X is the 'average' value which X takes –if we repeat the experiment that X describes many times and take the average of the outcomes then we should expect that average to be close to $\mathbb{E}[X]$.

3.4. Example

Suppose that X is the number obtained when we roll a fair die.

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + \ldots + 6 \cdot \mathbb{P}(X = 6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = 3.5\end{aligned}$$

Of course, you'll never throw 3.5 on a single roll of a die, but if you throw a lot of times you expect the average number thrown to be close to 3.5.

3.5. Properties of Expectation

If we have two random variables X and Y, then $X+Y$ is again a random variable and

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Similarly, if α is a constant then αX is a random variable and

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

3.6. Example

Roll two dice and let Z be the sum of the two numbers thrown. Then

$$\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$$

where X is the number on the first die and Y the number on the second. By our previous example and the property above we see $\mathbb{E}[Z] = 7$. (Compare this with calculating this expectation directly by writing out the probabilities of different values of Z.)

The problem with expectation is it is too blunt an instrument. The average case may not be typical. To try to capture more information about 'how spread out' our distribution is we introduce the variance.

3.7. Definition

For a random variable X with $\mathbb{E}[X] = \mu$, say, the *variance* of X is given by

$$\begin{aligned} var(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2] - \mu^2 \end{aligned}$$

3.8. Remark A statistician would use the *standard deviation* $\sigma = \sqrt{var(X)}$. This has the advantage of having the same units as X.

3.9. Example

Roll a fair die and let X be the number thrown. We already calculated that $\mathbb{E}[X] = 3.5$.

$$\mathbb{E}[X^2] = 1 \cdot \mathbb{P}(X=1) + 4 \cdot \mathbb{P}(X=4) + \ldots + 36 \cdot \mathbb{P}(X=6) = \frac{91}{6}$$

Thus

$$var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

(the standard deviation is about 1.7).

3.10. Properties of variance

Since variance measures spread around the mean, if X is a random variable and a is a constant,

$$var(X + a) = var(X)$$

If α is another constant

$$var(\alpha X) = \alpha^2 var(X)$$

(Notice then that the standard deviation of αX is α times the standard deviation of X.)

What about the variance of the sum of two random variables X and Y?

Remembering the properties of expectation we calculate:

$$\begin{aligned} var(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - (\mathbb{E}[Y])^2 \\ &= var(X) + var(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \end{aligned}$$

The quantity $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is called the *covariance* of X and Y.

We need another definition. In the same way as we defined events A, B to be *independent* if knowing that A had or had not happened told you nothing about whether B had happened, we define random variables X and Y to be independent if the probabilities for different values of Y are unaffected by knowing the value of X. Formally we have the following:

3.11. Definition

Two random variables X and Y are *independent* if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \text{for all } x, y$$

i.e. the events $(X = x), (Y = y)$ are independent for all choices of x and y.

A simple consequence of the definition is that if X and Y are independent random variables, then for any functions f and g

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

In particular, if X and Y are independent

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

and so the covariance of X and Y is zero and from our previous calculation $var(X+Y) = var(X) + var(Y)$.

3.12. WARNING

$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ does NOT guarantee independence of X and Y.

By analogy with the definitions for events, we define random variables $X_1, X_2, \ldots, X_n$ to be independent if

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\ldots\mathbb{P}(X_n = x_n)$$

3.13. Lemma

If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$var(X_1 + X_2 + \ldots X_n) = var(X_1) + var(X_2) + \ldots + var(X_n)$$

Common examples of discrete random variables.

3.14. The Binomial Distribution.
Model: repeated trials, number of successes.

A string of N binary digits is constructed so that independently each digit is 0 with probability p and 1 with probability $1-p$. The random variable X given by the number of zero's in the string has the *binomial distribution*.

Any given sequence of k zeroes and $N-k$ ones has probability $p^k(1-p)^{N-k}$ of occurring and the number of such sequences is the number of ways of choosing the k slots in which to put zeroes from the N available:

$$\left(\begin{array}{c} N \\ k \end{array} \right) = \frac{N!}{k!(N-k)!}$$

So that the distribution of X is given by

$$\mathbb{P}(X = k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$$

$k = 0, 1, \ldots, n$.

The number of ways of choosing k from N is the famous binomial coefficient which is where this distribution gets its name. You will see this object again in enumeration next term, all that we need here is the Binomial Theorem:

The Binomial Theorem

$$(x+y)^N = \sum_{k=0}^{N} \left(\begin{array}{c} N \\ k \end{array} \right) x^k y^{N-k}$$

Notice that

$$\begin{aligned} \sum_{k=0}^{N} \mathbb{P}(X = k) &= \sum_{k=0}^{N} \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k} \\ &= (p + (1-p))^N = 1^N = 1 \end{aligned}$$

3.15. The Geometric Distribution.
Model: repeated trials, time until first success.

Suppose now that the above binary string is infinite. Let Y be the position of the first zero.

$$\begin{aligned}
\mathbb{P}(Y=1) &= \mathbb{P}(\text{1st digit } = 0) = p \\
\mathbb{P}(Y=2) &= \mathbb{P}(\text{1st digit } = 1, \text{2nd digit } = 0) = p(1-p) \\
\mathbb{P}(Y=3) &= \mathbb{P}(\text{1st and 2nd digits } = 1, \text{3rd digit} = 0) = p(1-p)^2
\end{aligned}$$

and so on. In general

$$\mathbb{P}(Y=k) = \mathbb{P}(\text{ first (k-1) digits } = 1, \text{kth digit } = 0) = p(1-p)^{k-1}$$

We say that Y has the geometric distribution.

3.16. The Poisson distribution.
Model: the number of calls to arrive at a telephone exchange in a fixed time period.

We suppose that calls arrive 'at rate λ'. i.e.

$$\mathbb{P}(\text{call arrives in small interval of time } \delta t) = \lambda \cdot \delta t + o(\delta t)$$

Take a time period $[0,T]$ and let Z be the number of calls arriving in $[0,T]$. To find the distribution of Z divide $[0,T]$ into N small intervals of time of length $\delta t = T/N$. If we assume that δt is small enough that the probability of two or more calls arriving in a time-interval of length δt is negligible, then the number of calls arriving in $[0,T]$ has binomial distribution with $p = \lambda \delta t = \lambda T/N$. Using our previous calculation we have

$$\begin{aligned}
\mathbb{P}(k \text{ calls arrive in } [0,T]) &= \binom{N}{k} \left(\frac{\lambda T}{N}\right)^k \left(1 - \frac{\lambda T}{N}\right)^{N-k} \\
&= \frac{(\lambda T)^k}{k!} \frac{N!}{(N-k)!N^k} \left(1 - \frac{\lambda T}{N}\right)^N \left(1 - \frac{\lambda T}{N}\right)^{-k}
\end{aligned}$$

Now let $N \to \infty$. The expression above tends to

$$\frac{(\lambda T)^k}{k!} \cdot 1 \cdot e^{-\lambda T} \cdot 1$$

(using $\lim_{N\to\infty}(1 + \frac{a}{N})^N = e^a$). Thus

$$\mathbb{P}(Z=k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \qquad k = 0, 1, 2, \ldots$$

Z has the Poisson distribution. (One either says that Z is a Poisson point process of rate λ or setting $\mu = \lambda T$ that Z is Poisson parameter μ.)
Note: Letting $N \to \infty$ justifies our assumption that the probability that two or more calls arrive in an interval of time T/N is negligible.

The above derivation actually suggests that we can use the Poisson distribution as an approximation to the binomial distribution if we are considering a very large number of trials with a very low success probability. If N is large, p is small and $N \cdot p = \lambda$ is 'reasonable' then setting $T = 1$ in the above we have shown that

$$\left(\begin{array}{c} N \\ k \end{array} \right) (\frac{\lambda}{N})^k (1 - \frac{\lambda}{N})^{N-k} \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

3.17. Example

Consider a single page of the Guardian newspaper–containing say 10^6 characters. Suppose that each character (independently of the others) is mis-set with probability about 10^{-5}. Then the number of errors on the page should have the binomial distribution with $N = 10^6, p = 10^{-5}$. Now $N \cdot p = 10$, so using the above we estimate

$$\mathbb{P}(\text{ number of errors } = k) \approx \frac{10^k e^{-10}}{k!}$$

For example

$$\mathbb{P}(\text{number of errors } = 10) \approx \frac{10^{10} e^{-10}}{10!} \approx 0.125$$

3.18. Remark and WARNING

We have followed Grimmett and Welsh in our definition of the geometric distribution. Many authors consider a slight variant $\tilde{Y}$ of this in which $\tilde{Y}$ is allowed to take values $0, 1, 2, \ldots$ and $\mathbb{P}[\tilde{Y} = k] = pq^k$ for $k = 0, 1, 2, \ldots$. Notice that we can recover the random variable $\tilde{Y}$ as $Y - 1$ where Y has (our version of) the geometric distribution. Be sure you know which definition is being used.

§4. Probability Generating Functions.

We can encode all the information about the distribution of a discrete random variable X in a single function, sometimes written $P_X(s)$, called the *probability generating function of* X. It is defined by

$$P_X(s) = \sum_{k=0}^{\infty} \mathbb{P}(X = k)s^k = \mathbb{E}[s^X]$$

4.1. Examples

(We retain the notation of §3.)

a. **Binomial Distribution:** (N trials, success probability p)

$$P_X(s) = (1 - p + ps)^N$$

b. **Geometric Distribution:** (success probability p)

$$P_Y(s) = \frac{ps}{(1 - (1 - p)s)}$$

Notation: It is reasonably standard to write $q = (1 - p)$ so that this becomes

$$P_Y(s) = \frac{ps}{(1 - qs)}$$

c. **Poisson Distribution:** (number of calls to arrive in $[0, T]$, arrivals at rate λ) Writing $\mu = \lambda T$,

$$P_Z(s) = e^{-\mu(1-s)}$$

4.2. Properties of probability generating functions.

For any discrete random variable X

a.

$$P_X(1) = \sum_{k=0}^{\infty} \mathbb{P}(X = k) = 1$$

b.

$$\begin{aligned} P'_X(1) = \frac{d}{ds}P_X(s)\Big|_{s=1} &= \sum_{k=0}^{\infty}\mathbb{P}(X=k)ks^{k-1}\Big|_{s=1} \\ &= \sum_{k=0}^{\infty}k\mathbb{P}(X=k) = \mathbb{E}[X] \end{aligned}$$

c.

$$\begin{aligned} P''_X(1) &= \sum_{k=0}^{\infty}k(k-1)\mathbb{P}(X=k)s^{k-2}\Big|_{s=1} \\ &= \sum_{k=0}^{\infty}(k^2-k)\mathbb{P}(X=k) \\ &= \mathbb{E}[X^2]-\mathbb{E}[X] \end{aligned}$$

If, for example, I want to calculate the variance of X, I calculate

$$\begin{aligned} \mathbb{E}[X^2]-(\mathbb{E}[X])^2 &= \mathbb{E}[X^2]-\mathbb{E}[X]+\mathbb{E}[X]-(\mathbb{E}[X])^2 \\ &= P''_X(1)+P'_X(1)-(P'_X(1))^2 \end{aligned}$$

In general this may be much easier to calculate via the probability generating function than directly.

4.3. Example Let X have the binomial distribution (N trials, success probability p). Calculating directly gives

$$var[X] = \sum_{k=0}^{\infty}k^2\left(\begin{array}{c}N\\k\end{array}\right)p^k(1-p)^{N-k} - (\sum_{k=0}^{\infty}k\left(\begin{array}{c}N\\k\end{array}\right)p^k(1-p)^{N-k})^2$$

which looks horrible. If however we use the generating function we have

$$P_X(s) = (1-p+ps)^N$$

$$P'_X(s) = Np(1-p+ps)^{N-1}$$

$$P''_X(s) = N(N-1)p^2(1-p+ps)^{N-2}$$

Thus $P'_X(1) = Np, P''_X(1) = N(N-1)p^2$ and substituting into our expression for variance in terms of generating functions we have

$$\begin{aligned} var(X) &= P''_X(1) + P'_X(1) - (P'_X(1))^2 \\ &= N(N-1)p^2 + Np - (Np)^2 \\ &= Np(1-p) = Npq \end{aligned}$$

In fact one can pursue this process much further. We can calculate $\mathbb{E}[X^n]$ for any n in terms of derivatives of $P_X(s)$ evaluated at $s = 1$. The quantities $\mathbb{E}[X^n]$ are called the *moments* of X.

4.4. Note: Integer-valued random variables X, Y have the same probability generating function if and only if $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for all k. i.e. the p.g.f. characterises the distribution.

4.5. Theorem

If X and Y are independent positive-integer valued random variables, then

$$P_{X+Y}(s) = P_X(s)P_Y(s)$$

It follows that the sum $S_n = X_1 + X_2 + \ldots + X_n$ of n independent positive integer valued random variables has probability generating function

$$P_{S_n}(s) = P_{X_1}(s)P_{X_2}(s)\ldots P_{X_n}(s)$$

(proof by induction).

In particular if $X_1, X_2, \ldots X_n$ are independent and all have the same distribution (in which case we say that they are i.i.d. for *independent identically distributed*)

$$P_{S_n}(s) = (P_{X_1}(S))^n$$

4.6. Compound randomness.

We now know how to calculate the p.g.f. of the sum of n i.i.d. random variables where n is some fixed integer, but what if n itself is random? For example, let X be the number of errors in one byte of data and N be the number of bytes in a session. What is the p.g.f. of W, the total number of errors in a session?

We use the law of total probability. We can write

$$W = X_1 + X_2 + \ldots + X_N$$

where X_k is the number of errors in the kth byte. For the events B_n in the law of total probability we take $B_n = (N = n)$. We can then rewrite

$$\mathbb{P}(W = k) = \mathbb{P}(W = k|N = 0)\mathbb{P}(N = 0) + \mathbb{P}(W = k|N = 1)\mathbb{P}(N = 1) + \ldots$$

Then

$$P_W(s) = \sum_{k=0}^{\infty}\{\mathbb{P}(W = k|N = 0)\mathbb{P}(N = 0)+\mathbb{P}(W = k|N = 1)\mathbb{P}(N = 1)+\ldots\}s^k$$

Now rearrange this by gathering together terms involving $\mathbb{P}(N = n)$ for each i and we get

$$P_W(s) = \sum_{k=0}^{\infty} s^k\mathbb{P}(W = k|N = 0)\mathbb{P}(N = 0)+\sum_{k=0}^{\infty} s^k\mathbb{P}(W = k|N = 1)\mathbb{P}(N = 1)+\ldots$$

Now the nth sum is just

$$\sum_{k=0}^{\infty} s^k\mathbb{P}(X_1 + X_2 + \ldots + X_n = k) = P_{S_n}(s) = (P_X(s))^n$$

(by our previous calculation). Thus

$$P_W(s) = \sum_{n=0}^{\infty}(P_X(s))^n\mathbb{P}(N = n) = P_N(P_X(s))$$

4.7. Example

If the number of errors per byte has p.g.f. $P_X(s) = (ps+(1-p))^8$ and the number of bytes has p.g.f. $P_N(s) = (1-p)s/(1-ps)$ then the total number of errors has p.g.f.

$$P_W(s) = \frac{(1-p)}{1-p(ps+(1-p))^8}(ps+(1-p))^8$$

§5. Continuous Probability

Suppose that we are choosing a number at random from $[0,1]$ in such a way that any number is equally likely to be picked. How can we do this? Does the problem even make sense? Evidently we cannot assign a positive probability to each number –our probabilities wouldn't sum to one.

To get around this we don't define the probability of individual sample points, but only of certain events. For example, by symmetry we expect that $\mathbb{P}(X \leq 1/2) = 1/2$. More generally, we expect the probability that X lies in an interval $[a,b] \subseteq [0,1]$ to be equal to the length of that interval:

$$\mathbb{P}(X \in [a,b]) = b - a \qquad 0 \leq a < b \leq 1$$

We will just deal with continuous random variables whose values are real numbers. It will be enough then to specify $\mathbb{P}(X \leq t)$ for each $t \in \mathbb{R}$.

5.1. Definition

For a continuous real-valued random variable X we define the *distribution function* of X to be

$$F_X(t) = \mathbb{P}(X \leq t) \qquad t \in \mathbb{R}$$

Note that we will always have $F_X(-\infty) = 0$ and as t increases $F_X(t)$ increases to $F_X(\infty) = 1$.

Some important continuous distributions

5.2. The Uniform Distribution on $[0,1]$

This is the name given to the distribution which we discussed before. The idea is that 'every point of $[0,1]$ is equally likely to be picked'.

$$F_X(t) = \begin{cases} 0 & t \leq 0 \\ t & 0 \leq t \leq 1 \\ 1 & 1 \leq t \end{cases}$$

This ensures that the $\mathbb{P}(X \in [a,b]) = b - a$ as our intuition suggested we should require.

5.3. The exponential distribution

This is intimately connected with the Poisson process which we discussed in 3.16.. The Poisson process modelled the number of calls to arrive at a telephone exchange whereas the exponential distribution models the time between successive calls.

Recall that for a Poisson point process of rate λ, the probability of k calls arriving in time $[0,t]$ is given by

$$\mathbb{P}(Z=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

In particular then, the probability of *no* calls arriving by time t is given by

$$\mathbb{P}(Z=0)=e^{-\lambda t}$$

If we write X for the arrival time of the first call, then $\mathbb{P}(X>t)=e^{-\lambda t}$ so that

$$\mathbb{P}(X\leq t)=1-e^{-\lambda t} \qquad t\geq 0$$

The random variable X is said to have the exponential distribution.

The exponential distribution has an extremely important property from a theoretical point of view: '$\mathbb{P}(X>t+s|X>t)=\mathbb{P}(X>s)$'. I haven't defined anything formally, but what this says is that an exponentially distributed random variable has no memory. If the first call hasn't arrived at time t, you may as well start the clock again –it has the same probability of arriving in the next s minutes as it had of arriving in the first s minutes.

5.4. The normal distribution

To a statistician, a normally distributed random variable is a fundamental object. We'll see why when we write down an important theorem called the Central Limit Theorem. For now we content ourselves with recording the definition.

A random variable X is normally distributed if

$$F_X(t)=\mathbb{P}(X\leq t)=\int_{-\infty}^{t}\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})dx$$

Here μ and σ are parameters and they give the expectation and variance of X. One often writes $X\sim N(\mu,\sigma)$.

5.5. Expectation of continuous random variables

Recall that for discrete random variables

$$\mathbb{E}[f(X)]=\sum_{k=0}^{\infty}f(k)\mathbb{P}(X=k)$$

The continuous analogue of a sum is an integral.

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) dF_X(t)$$

where $dF_X(t) = F'_X(t)dt$ (which we assume exists). One way to think of this is as summing

$$\sum g(t)\mathbb{P}(X \in [t, t+\delta t))$$

over tiny intervals $[t, t+\delta t)$. In particular,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t dF_X(t)$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} t^2 dF_X(t)$$

(with the same interpretations as for discrete random variables).

5.6. Definition

The function $F'_X(t)$ (if it exists) is called the density function of the distribution of X. It is often denoted $f_X(t)$.

5.7. Examples

a. **Uniform distribution on** $[0,1]$

$$f_X(t) = \begin{cases} 1 & t \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

b. **Exponential distribution**

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

c. **Normal distribution**

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(t-\mu)^2}{\sigma^2}\right)$$

5.8. Notes

All density functions are positive and

$$\int_{-\infty}^{\infty} f_X(t)dt = 1$$

If X only takes non-negative values, then F_X and f_X will be zero for $t < 0$ and so $\int_{-\infty}^{\infty}$ can be replaced by $\int_0^{\infty}$ in calculations.
As in the discrete case, 'expectation is a linear operator'. i.e.

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \qquad \mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$$

Analogous to the discrete case we define variance by

$$var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

and the kth moment of X is $\mathbb{E}[X^k]$.

We also have the notion of independence for continuous random variables. Intuitively it is exactly as before –'X, Y are independent if knowing about X tells us nothing about Y'. Formally:

5.9. Definition

Random variables X and Y are independent if the events $(X \leq x), (Y \leq y)$ are independent for all x, y. i.e.

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$$

§6. Moment generating functions

Recall that for discrete random variables we were able to encode all the information about the distribution in a single function –the probability generating function. If we were able to identify this function by some means, then expanding it as a power series, the coefficient of s^k gave the probability that our random variable took the value k. Unfortunately, this only works for random variables taking only non-negative integer values. For more general random variables we consider a modification of the p.g.f. called the moment generating function.

6.1. Definition

For a random variable X the *moment generating function* $M_X(t)$ is defined by

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} dF_X(x)$$

(for all t for which this expectation exists).

In some cases convergence of the integral can be a problem. We can get around this by introducing the *characteristic function* defined by $\mathbb{E}[e^{itX}]$ where i is $\sqrt{-1}$ which exists for all $t \in \mathbb{R}$, but we want to avoid complex numbers here. (If you know about Laplace and Fourier transforms, then you should think of the moment generating function as the Laplace transform and the characteristic function as the Fourier transform.)

Notice that the moment generating function is perfectly well defined for discrete random variables where

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \mathbb{P}(X = k) = P_X(e^t)$$

In this sense the m.g.f. really is just a modification of the p.g.f..

As you might guess, if you know the m.g.f. of a distribution then it is easy to recover the moments. Formally,

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \mathbb{E}[\frac{(tX)^k}{k!}] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k]$$

Thus

$$\frac{d}{dt} \mathbb{E}[e^{tX}] \Big|_{t=0} = \mathbb{E}[X]$$

$$\frac{d^2}{dt^2}\mathbb{E}[e^{tX}]\Big|_{t=0} = \mathbb{E}[X^2]$$

and in general

$$\frac{d^k}{dt^k}\mathbb{E}[e^{tX}]\Big|_{t=0} = \mathbb{E}[X^k]$$

6.2. Example (The moments of the exponential distribution)

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} dF_X(x) \\ &= \int_0^\infty e^{tx}\lambda e^{-\lambda x} dx \\ &= \int_0^\infty \lambda e^{(t-\lambda)x} dx \\ &= \frac{\lambda}{(\lambda - t)} \end{aligned}$$

This is certainly well-defined for $t < \lambda$ so we restrict our attention to such t. Then

$$\frac{\lambda}{(\lambda - t)} = \frac{1}{(1 - \frac{t}{\lambda})} = 1 + \frac{t}{\lambda} + (\frac{t}{\lambda})^2 + \ldots$$

Comparing coefficients in the powers series

$$\sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}[X^k] = \sum_{k=0}^{\infty}(\frac{1}{\lambda})^k t^k$$

gives

$$\mathbb{E}[X^k] = \frac{k!}{\lambda^k}$$

In particular

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathbb{E}[X^2] = \frac{2}{\lambda^2}$$

and

$$var(X) = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

6.3. Theorem

If X, Y are independent real-valued random variables, then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

A similar argument gives that if a, b are real constants then

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

This sort of relationship can be useful for identifying distributions once we have the following result:

6.4. Theorem

If $M_X(t) = \mathbb{E}[e^{tX}] < \infty$ for all t satisfying $-\delta < t < \delta$ for some $\delta > 0$, then if Y is another random variable with moment generating function $M_Y(t) = M_X(t)$, Y must have the same distribution function as X. (i.e. the moment generating function determines the distribution uniquely.)
(The proof is essentially the inversion Theorem for Laplace transforms and is omitted.)

6.5. Example (The normal distribution)

First suppose that $X \sim N(0,1)$. i.e. we have set the parameters $\mu = 0, \sigma = 1$. Then

$$M_X(t) = e^{\frac{1}{2}t^2}$$

If $Y \sim N(\mu, \sigma)$, then

$$M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

which we recognise as $e^{\mu t}M_X(\sigma t)$. Now using our Theorem we deduce that if $X \sim N(0,1)$ then $\sigma X + \mu \sim N(\mu, \sigma)$

The importance of the normal distribution stems from the following remarkable and powerful Theorem:

6.6. The Central Limit Theorem

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean μ and non-zero variance σ^2. Let

$$Z_n = \frac{((X_1 + X_2 + \ldots + X_n) - n\mu)}{\sigma\sqrt{n}}$$

Then as $n \to \infty$, for each $x \in \mathbb{R}$

$$\mathbb{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right) du$$

i.e. $\mathbb{P}(Z_n \leq x) \to \mathbb{P}(Y \leq X)$ where $Y \sim N(0,1)$.

The remarkable feature of this result is that it is independent of the distribution of the X_i's.

The 'noise' in a system is often modelled as a random variable. If the noise is the compound effect of small independent identically distributed random variables, then the Central Limit Theorem suggests that we should assume that it is normally distributed.

6.7. Example

A digitised system is transmitted as

$$0 \leftrightarrow -V \text{ volts}$$

$$1 \leftrightarrow +V \text{ volts}$$

At a given sampling instant the received voltage is interpreted as a

$$\begin{cases} 0 & \text{if it is } < 0 \\ 1 & \text{if it is } > 0 \end{cases}$$

If 0 is sent, then owing to noise the received voltage has $N(-V, \sigma^2)$ distribution. What is the probability that a 0 sent is interpreted as a 1?
If X is the received voltage then

$$\mathbb{P}(X > 0) = 1 - F_X(0) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty exp(-\frac{(x+V)^2}{2\sigma^2})dx$$

Often it is hard or even impossible to calculate probabilities exactly. If, for example, the probability is an error probability then it may be adequate just to obtain upper bounds. Here are some upper bounds –based on the same idea and often given in very general form, but here stated separately as often the simpler forms are all that is needed.

6.8. Chebyshev's inequality

If Y is a random variable and $\mathbb{E}[Y^2] < \infty$, then

$$\mathbb{P}(|Y| \geq a) \leq \frac{1}{a^2}\mathbb{E}[Y^2] \qquad \text{for all } a > 0$$

Proof

Using the law of total probability with events $(|Y| \geq a)$ and $(|Y| < a)$ we get

$$\mathbb{E}[Y^2] = \mathbb{E}[Y^2 \,\big|\, |Y| \geq a]\mathbb{P}(|Y| \geq a) + \mathbb{E}[Y^2 \,\big|\, |Y| < a]\mathbb{P}(|Y| < a)$$

The second term is certainly non-negative, so that

$$\mathbb{E}[Y^2] \geq \mathbb{E}[Y^2 \,\big|\, |Y| \geq a]\mathbb{P}(|Y| \geq a) \geq a^2\mathbb{P}(|Y| \geq a)$$

which yields the result.

Our other bounds are two of the many consequences of Chebyshev's inequality:

a. For any random variable X with mean μ and variance σ^2

$$\mathbb{P}(|X - \mu| \geq a] \leq \frac{\sigma^2}{a^2}$$

(Apply Chebyshev to $Y = X - \mu$.)

b. For any random variable X

$$\mathbb{P}(X > b) \leq e^{-bt} M_X(t) \qquad \text{for any } t > 0$$

where $M_X(t)$ is the moment generating function of X. [Proof:

$$\begin{aligned} \mathbb{P}(X \geq b) &= \mathbb{P}\left(\frac{tX}{2} \geq \frac{tb}{2}\right) \\ &= \mathbb{P}\left(exp(\frac{tX}{2}) \geq exp(\frac{tb}{2})\right) \\ &\leq \frac{\mathbb{E}[exp(tX)]}{exp(tb)} = e^{-tb} M_X(t) \end{aligned}$$

by applying Chebyshev's inequality to $Y = exp(tX/2)$.]

The second inequality can be improved by optimising over $t > 0$. This is sometimes called the Chernoff bound.

As an application we'll prove the 'law of averages' –usually called the law of large numbers. There are stronger versions of this theorem.

6.9. The Weak Law of Large Numbers

If $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $var(X_i) = \sigma^2$ and $s_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, then for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|s_n - \mu| > \epsilon) = 0$$

Proof

From the properties of expectation $\mathbb{E}[s_n] = \mu$. Also $var(s_n) = \frac{1}{n^2} var(X_1 + X_2 + \ldots + X_n)$ and since the X_i's are independent the variance of their sum is the sum of their variances and so $var(s_n) = \frac{1}{n}\sigma^2$. Now applying the first inequality above we see

$$\mathbb{P}(|s_n - \mu| > \epsilon) \leq \frac{1}{n} \cdot \frac{\sigma^2}{\epsilon^2}$$

and for each fixed ϵ this obviously tends to zero as $n \to \infty$.

§7. Markov Chains

7.1. Definitions

A *Markov chain* is a system which can be in a certain finite number of states labelled say $0, 1, 2, \ldots$. At each 'tick of a clock', times $t = 1, 2, 3, \ldots$, it moves from its current state i to a new state j. Which state j it moves into is determined by probabilities

$$p_{ij} = \mathbb{P}(\text{system changes from state } i \text{ to state } j)$$

Obviously we must have that $p_{ij} \geq 0$ for every pair i, j and $\sum_j p_{ij} = 1$ for each i. The *transition probabilities* are recorded in a matrix

$$P = \begin{pmatrix} p_{00} & p_{01} & \ldots \\ p_{10} & p_{11} & \ldots \\ p_{20} & p_{21} & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}$$

called the *transition matrix* of the Markov chain.

Any matrix with non-negative entries and whose rows sum to one is called a *stochastic matrix*.

Given that we know P, the only information we need about the Markov chain to determine the probability of it being in any given state after the next transition is its current state –its past history is of no importance. This property is called the *lack of memory* property or the *Markov* property.

Useful Markov chains often have a large number of states. Here we restrict ourselves to small examples which illustrate how Markov chains are analysed.

7.2. Example

A computer system has three states: 0 (down), 1 (usable), 2 (overloaded) with

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

For example, if it is usable at time n with probability $3/4$ it will also be usable at time $n + 1$.

We don't know what state the system will be in at time n but we can write down a probability vector

$$\pi^{(n)} = (\pi_0^{(n)}, \pi_1^{(n)}, \pi_2^{(n)})$$

where $\pi_i^{(n)}$ denotes the probability that the system is in state i at time n. We assume that $\pi^{(0)}$ is known. (If the system starts in state 0, then $\pi^{(0)} = (1,0,0)$.) We can calculate $\pi^{(n+1)}$ in terms of $\pi^{(n)}$ as follows:

$$\begin{aligned}
\pi_j^{(n+1)} &= \mathbb{P}(\text{system in state } j \text{ at time } t = n+1) \\
&= \sum_{i=0}^{2} \mathbb{P}(\text{in state } j \text{ at } t = n+1 | \text{ in state } i \text{ at } t = n)\mathbb{P}(\text{in state } i \text{ at } t = n) \\
&= \sum_{i=0}^{2} p_{ij}\pi_i^{(n)}
\end{aligned}$$

This is the dot product of the jth row of the matrix with $\pi^{(n)}$. So writing this in terms of vectors and matrices we see

$$\pi^{(n+1)} = \pi^{(n)}P$$

The computer printout of $\pi^{(n)}$ for $n = 1, \ldots 19$ with $\pi^{(0)} = (1,0,0), (0,1,0), (0,0,1)$ suggests that $\pi^{(n)}$ tends to a limit *independent of* $\pi^{(0)}$ as n increases. If $\pi^{(n)}$ *does* tend to a limit, π say, then π must satisfy

$$\pi = \pi P$$

This equation can be solved for π to give

$$\pi = (\frac{2}{13}, \frac{8}{13}, \frac{3}{13})$$

Computer simulations also suggest that the number of visits before time n (large) of a typical realisation of the Markov chain to each state will be roughly in proportion to the steady state probabilities. Thus after 1300 transitions, we would expect to have been in state 0 about 200 times, in state 1 about 800 times and in state 2 about 300 times. Results which relate steady-state probabilities to frequency of visits in realisations of Markov chains are called *ergodic theorems*.

In the above example the system had a unique steady state and no matter where we started from it settled down to that steady state in the sense that the probability of being in state 0 was about 2/13, in state 1 about 8/13 and in state 2 about 3/13 at large times. This is not true of all Markov chains. There are essentially two things which can go wrong:

a. the first is obvious. The system can have 'traps'. For example, the states may split up into distinct groups in such a way that the system cannot get from one group of states to another. We shall therefore require that the chain has the property that every state can be reached (in one or more transitions) from every other state. Such a chain is said to be *irreducible*.

b. the second barrier is more subtle. It may for example be the case that from some initial states, the possible states split up into two groups –one visited only at even times and the other only at odd times. In general, there may be states that are only visited by the chain at times divisible by some integer k. A Markov chain with this property is said to be periodic. We require that the Markov chain be *aperiodic*. That is there are no states with this property for any $k = 2, 3, \ldots$.

7.3. Theorem

If a Markov chain is irreducible and aperiodic then it has a unique steady state probability vector π such that $\pi = \pi P$. As $n \to \infty$, the probability vector $\pi^{(n)}$ tends to π, independent of the initial vector $\pi^{(0)}$.

A chain which is irreducible and aperiodic is said to be *regular*. A necessary and sufficient condition for a chain to be regular is that for some n it is possible to get from any state to any other in exactly n transitions. That is, for some n all the entries of the matrix P^n are positive.

§8. Birth and Death Processes

8.1. Definitions

So far we have considered only discrete time Markov chains (transitions took place only at times $t = 1, 2, 3, \ldots$). We can modify this model to allow transitions at *any* time. This modification is a *continuous time* Markov chain. Transitions are as follows:
If the system is in state i at time t then the probability that in a small interval of time $[t, t + \delta t)$ it moves to state j is $\lambda_{i,j}\delta t$ (c.f. the Poisson process).

To simplify matters we consider only a special class of continuous time Markov chains called *birth and death processes*. Here the states are just $0, 1, \ldots, N$ and transitions are only possible from i to $i+1$ $(i \neq N)$ and from i to $i-1$ $(i \neq 0)$. We write

$$\lambda_{i,i+1} = b_i \qquad i = 0, 1, \ldots, N-1$$

$$\lambda_{i,i-1} = d_i \qquad i = 1, 2, \ldots, N$$

and all other $\lambda_{i,j} = 0$.

Consider the time spent by the process in state i before there is a transition. The time before a birth has the exponential distribution with parameter b_i and the time before a death has the exponential distribution with parameter d_i. So we have two random alarm clocks running and when the first one goes off we have either a birth or a death (according to which of the clocks has gone off).

8.2. The steady state

In the same way as 'nice' discrete time Markov chains had a steady state probability, so do 'nice' continuous time Markov chains. If the b_i and d_i are all non-zero, then the birth and death process has a steady-state probability vector π which is independent of the initial state of the process. We shall assume this without proof and find out what that steady state vector must be.

Let $\pi(t) = (\pi_0(t), \pi_1(t), \ldots, \pi_N(t))$ be the probability vector for the Markov chain at time t. (We assume that $\pi(0)$ is given. If we are already in the steady state, then $\pi(t + \delta t) = \pi(t)$ for all $\delta t > 0$. i.e. $\pi_i(t + \delta t) = \pi_i(t)$ for all i. Now for $i \neq 0, N$

$$\begin{aligned} \pi_i(t+\delta t) &= \mathbb{P}(\text{in state } i \text{ at } t+\delta t) \\ &= \mathbb{P}(\text{in state } i \text{ at } t+\delta t \text{ and state } i \text{ at } t) \end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}(\text{in state } i \text{ at } t+\delta t \text{ and state } i+1 \text{ at } t)\\
&+\mathbb{P}(\text{in state } i \text{ at } t+\delta t \text{ and state } i-1 \text{ at } t)\\
= \;& \mathbb{P}(\text{in state } i \text{ at } t+\delta t|\text{state } i \text{ at } t)\mathbb{P}(\text{state } i \text{ at } t)\\
&+\mathbb{P}(\text{in state } i \text{ at } t+\delta t|\text{state } i+1 \text{ at } t)\mathbb{P}(\text{state } i+1 \text{ at } t)\\
&+\mathbb{P}(\text{in state } i \text{ at } t+\delta t|\text{state } i-1 \text{ at } t)\mathbb{P}(\text{state } i-1 \text{ at } t)\\
= \;& (1-(b_i+d_i)\delta t)\pi_i(t) + d_{i+1}\delta t\pi_{i+1}(t) + b_{i-1}\delta t\pi_{i-1}(t)\\
= \;& \pi_i(t) - \delta t\left((b_i+d_i)\pi_i(t) - d_{i+1}\pi_{i+1}(t) - b_{i-1}\pi_{i-1}(t)\right)
\end{aligned}
$$

But this must equat $\pi_i(t)$ so for $i \neq 0, N$

$$(b_i+d_i)\pi_i(t) = d_{i+1}\pi_{i+1}(t) + b_{i-1}\pi_{i-1}(t)$$

For $i = 0, N$ the same technique leads to

$$b_0\pi_0(t) = d_1\pi_1(t)$$

$$d_N\pi_N(t) = b_{N-1}\pi_{N-1}(t)$$

These equations can be visualised in terms of 'probability flow' = state probability multiplied by rate of transition. In the steady state flow out of state i = flow into state i. In fact the net flow along across any transition is zero, giving

$$b_{i-1}\pi_{i-1} = d_i\pi_i$$

This allows us to solve for π_i in terms of π_0:

$$\pi_i = \frac{b_0 b_1 \ldots b_{i-1}}{d_1 d_2 \ldots d_i}\pi_0$$

and π_0 is determined from $\sum \pi_i = 1$.

We now consider two important applications of birth and death processes.

8.3. Carrying capacity of Telecom circuits

Suppose that we have N circuits and new calls are arriving (wanting to use a new circuit) at rate λ (with arrivals forming a Poisson process). If all N circuits are in use, then the call is lost. Each call is assumed to have an exponentially distributed duration ('holding time') with parameter μ. Hence, if there are i calls in progress at time t,

$$\mathbb{P}(\text{some call terminates in } [t, t+\delta t)) = i\mu\delta t$$

Thus we have a birth and death process with

$$b_i = \lambda \qquad d_i = i\mu$$

The steady state is given (using the formula in 8.2) by

$$\pi_i = \frac{\lambda^i}{\mu^i i!}\pi_0$$

and since $\sum \pi_i = 1$ we find

$$\pi_i = \frac{(\lambda/\mu)^i}{i!\left(1 + (\lambda/\mu) + (\lambda/\mu)^2/2! + \ldots + (\lambda/\mu)^N/N!\right)}$$

In particular, π_N gives the steady state probability that *all circuits* are occupied and hence that an arriving call will be lost. π_N is oftem written $E(N, \lambda/\mu)$ and our expression for it is called *Erlang's loss formula.*

The ratio λ/μ is actually the mean number of calls that would be available if there were infinitely many circuits available (and so no calls are lost).

Returning to the case of a finite number of circuits we have the following interpretation of $E(N, \lambda/\mu)$.

8.4. Lemma

The expected number of busy circuits in the steady state is given by

$$\frac{\lambda}{\mu}\left(1 - E(N, \lambda/\mu)\right)$$

i.e. $E(N, \lambda/\mu)$ gives us the expected fraction of traffic lost.

8.5. The M/M/1 queue

The notation M/M/1 is an abbreviation for Markovian (memoryless) arrivals, Markovian service times and 1 server.

The idea is that customers arrive at a server according to a Poisson process of rate λ, say. Each customer takes an exponentially distributed amount of time to serve with parameter μ, say. We model this as a birth and death process with state i corresponding to there being i customers (including the one being served) in line. Then $b_i = \lambda, d_i = \mu$. Here there are infinitely many states.

If $b_i > d_i$ then customers are arriving faster than they are being served and there is no steady state (in the long term the line just keeps growing).

If $\lambda < \mu$ then our previous analysis remains valid and we obtain

$$\pi_i = \frac{b_0 b_1 \ldots b_{i-1}}{d_1 d_2 \ldots d_i}\pi_0 = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

and $\sum \pi_i = 1$ gives $\pi_0 = 1 - \lambda/\mu$. Thus

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right) \qquad i = 0, 1, 2, \ldots$$

This shows that queue lengths have the *geometric distribution* in the steady state.

Using the p.g.f. for the geometric distribution in the usual way gives that the mean queue length in the steady state is

$$\frac{\lambda/\mu}{1 - \lambda/\mu}$$

For instance, if $\lambda/\mu = 0.9$ the mean queue length is 9 and the closer the ratio λ/μ of arrival rate to service rate gets to one, the longer the mean queue length gets.