Chapter 0

Basic Probability Theory (Assumed Knowledge)

This is a review of some basic probability material assumed for this course. I suggest that students read through all sections to make sure that they have a firm grasp of the various concepts. With regard to the individual distributions given at the end, you will find that exponential distribution occurs frequently in the work we will be doing in this course. It is the distribution that gives us the times between events in a Poisson process (a commonly used process that describes arrivals)

0.1 Random Experiments, Sample Space and Events, Probability

We will be concerned with "random" experiments where the outcome of the experiment has a degree of unpredictability but is nevertheless still subject to analysis.

Examples:

1. The holding time of a voice/video call.
2. The number of arriving calls in a fixed time period.
3. The number of calls blocked because of an overloaded network.
4. The cell to be visited next by a mobile unit.
5. The amount of bandwidth required by the next user in a B-ISDN network.
6. Telstra's profit/loss in a given day.

Although we cannot predict the outcome of a random experiment with certainty we usually can predict the set of all possible outcomes.

Definition 0.1.1: (Sample space)

The *sample space* S of a random experiment is the set of all possible outcomes of the experiment.

Examples:

1. The holding time of a call $S = \{\text{time } t; t \geq 0\} = \mathbb{R}^+$.
2. The number of arriving calls $S = \{0, 1, \cdots\} = \mathbb{Z}^+$.
3. Telstra's profit/loss in a given day $S = \{c : -\infty < c < \infty\} = \mathbb{R}$.
4. The cell to be visited next by a mobile unit $S = \{\text{all labels of cells}\}$.

Often we are not interested in a single outcome but in whether or not one of a group of outcomes occurs.

Definition 0.1.2: (Event)

An *event* is a subset of the sample space that is a collection of some possible outcomes of the experiment.

Notation: Events will be denoted by capital letters $A, B, C, \ldots$. We say event A occurs if the outcome of the experiment is one of the elements in A.

Examples:

1. The event that a particular household member hogs the phone for at least one hour $A = \{t : t \geq 60 \text{ minutes}\}$.
2. The event "Telstra makes a profit" $A = \{c : c > 0\}$.
3. The event that the number of calls in a morning exceeds 1500 $A = \{n : n > 1500\}$.

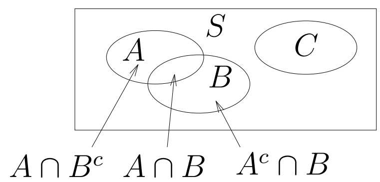
Mathematically, events are sets.

1. The event $A \cup B$ (A union B) is the event that A *or* B *or* both occur.
2. The event $A \cap B$ (A intersection B) is the event that A *and* B both occur.
3. The event A^c (A complement) is the event that A does not occur.
4. The event containing no outcomes is denoted $\emptyset$ (the impossible event).
   Note that the event S always occurs because the outcome to the experiment will always belong to the sample space S and hence $\emptyset^c = S$.
5. Two events A_1, A_2 which have no outcomes in common, that is, $A_1 \cap A_2 = \emptyset$, are called mutually exclusive or disjoint events.
   (For example A and A^c are two disjoint events.) Similarly, events $A_1, A_2, \ldots, A_n$ are disjoint if no two have outcomes in common, $A_i \cap A_j = \emptyset, \ \forall i \neq j$.

6. Two events are exhaustive if they contain all possible outcomes between them, e.g. $A_1 \cup A_2 = S$. Example: $A \cup A^c = S$. Therefore events A and A^c are exhaustive. Similarly, events $A_1, A_2, \ldots, A_n$ are exhaustive if their union is the whole sample space. For example $A_1 \cup A_2 \ldots \cup A_n = S$.

Example 0.1.1: (Venn diagrams)

Since events are *sets* of outcomes, they may be represented in the usual way by Venn diagrams. e.g. Suppose A, B, and C are events in a sample space S represented by.



From the Venn diagram we can see that

(i) $A \cap C = \emptyset$ and therefore events A and C are disjoint.

(ii) $(A \cap B^c) \cap (A^c \cap B) = \emptyset$ and hence events $A \cap B^c$ and $A^c \cap B$ are disjoint.

Note: *Venn diagrams can be used to help visualise things but cannot be used to prove things.* $\triangleleft$

0.2 Probability Axioms

What do we mean when we say The probability that a toss of a coin will result in heads is 1/2?

This is a very complicated and subtle question and people who do research into the foundations of probability are still arguing about it. There is, however, an interpretation that is accepted by most people for practical purposes, that such statements are made based upon some information about relative frequencies. In the experience of most of us, heads comes up about half the time in coin tosses, so we say the probability of a head is 1/2.

Therefore having listed all possible outcomes of a random experiment we seek a measure which tells us how likely it is that a particular event will occur. This measure is called the *probability* of an event and is a measure of expected relative frequency of that event.

We can now postulate the existence of a number $P(A)$ called the probability of event A and define a set of *probability axioms* which we think are reasonable for the probabilities to obey.

Definition 0.2.1: (Probability of an event)

Consider a random experiment whose sample space is S.
If for each event A of the sample space, a number $P(A)$ can be defined which satisfies the following axioms then we refer to $P(A)$ as the probability of event A.

Axiom 1: $0 \leq P(A) \leq 1$.
Axiom 2: $P(S) = 1$.
Axiom 3: For any two events A_1, A_2 which are disjoint.

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

Proposition 0.2.1

$P(\emptyset) = 0$.

Proof: We know $\emptyset \cap S = \emptyset$ therefore $\emptyset$ and S are disjoint. Also $\emptyset \cup S = S$. Therefore $P(S) = P(\emptyset \cup S) = P(\emptyset) + P(S)$ by axiom 3. Axiom 2 implies that $\Rightarrow P(\emptyset) + 1 = 1$ and so $\Rightarrow P(\emptyset) = 0$. ∎

Proposition 0.2.2

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

Proof: If $A \subseteq B$ then $B = A \cup (A^c \cap B)$, and A and $A^c \cap B$ are disjoint events.
Therefore by axiom 3 we have $P(B) = P(A) + P(B \cap A^c)$ and then by axiom 1 we see that $P(A^c \cap B) \geq 0$ so that $P(B) \geq P(A)$. ∎

Proposition 0.2.3

For any event A, $P(A^c) = 1 - P(A)$. (Sometimes it is more convenient to calculate the probability of the complementary event.)

Proof: A^c and A are mutually exclusive and $A^c \cup A = S$, therefore $P(A^c \cup A) = P(S)$.
Hence

$$\begin{aligned} P(A^c) + P(A) &= P(S) \\ &= 1 \end{aligned}$$

using axioms 2 and 3. Hence in a mobile environment, during a handover from one cell to a neighbouring cell

$$P(\text{at least one frequency is available}) = 1 - P(\text{no frequency is available}) .$$

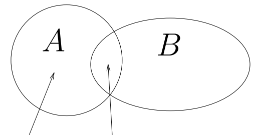
∎

Proposition 0.2.4

For any two events A and B, $P(A \cap B^c) = P(A) - P(A \cap B)$.

Proof:

We know that (see the diagram for an illustration) $(A \cap B^c) \cap (A \cap B) = \emptyset$ and so $\Rightarrow (A \cap B^c)$ and $(A \cap B)$ are disjoint events.
Therefore from Axiom 3, $P\Big((A \cap B^c) \cup (A \cap B)\Big) = P(A \cap B^c) + P(A \cap B)$.
Now $A = (A \cap B^c) \cup (A \cap B)$ and so $P(A) = P(A \cap B^c) + P(A \cap B)$ and finally $P(A \cap B^c) = P(A) - P(A \cap B)$. ∎

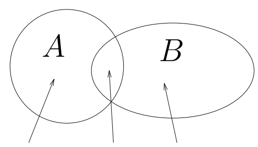


$A \cap B^c \quad A \cap B$

Proposition 0.2.5

For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: $A \cap B^c$, $A \cap B$, and $A^c \cap B$ are disjoint events and $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$ (see diagram)



$A \cap B^c \quad A \cap B \quad A^c \cap B$

Therefore using Axiom 3 twice and Proposition 0.2.4 twice

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(B \cap A^c) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(B \cap A^c) \\ &= P(A) + P(B) - P(A \cap B)\ . \end{aligned}$$

∎

Theorem 0.2.1

If $B_1, B_2, \ldots, B_n$ are disjoint and exhaustive events (they partition S), then for any event A,

$$P(A) = \textstyle\sum_{i=1}^n P(A \cap B_i).$$

Proof: If the events $B_1, B_2, \ldots, B_n$ are exhaustive then $S = B_1 \cup B_2 \cup \ldots \cup B_n$.
Therefore the event $A = A \cap S = A \cap (B_1 \cup B_2 \cup \ldots \cup B_n) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)$.
Also $(A \cap B_i) \cap (A \cap B_j) = \emptyset$, so that $(A \cap B_i)$ and $(A \cap B_j)$ are disjoint for all $i \neq j$.
Hence, by Axiom 3

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)) \\ &= P(A \cap B_1) + P(A \cap B_2) \ldots + P(A \cap B_n) \ = \sum_{i=1}^n P(A \cap B_i)\ . \end{aligned}$$

∎

0.3 Conditional Probability

Example 0.3.1: (Toss two fair dice)

If we know that the first die is a 3 what is the probability that the sum of the two dice is 8?

Solution: $S = \{(1,1)\ldots(6,6)\}$ and given that the first die is a 3 there are only six outcomes of interest being $\{(3,1),(3,2)\ldots(3,6)\}$ and therefore the probability of $(3,5)$ occurring is 1/6 (conditional on the first die being a 3). $\triangleleft$

Let A denote the event "the sum on the dice is 8" and B denote the event "the first die is a 3". The probability we have calculated is called the conditional probability that A occurs given B has occurred and is denoted $P(A|B)$. If B occurs then in order for A to occur it is necessary that the actual occurrence be an outcome in *both* A and B, so that it must be in $A \cap B$. Since we know B has occurred, B may be regarded as our new sample space. Hence the probability of "A given B" is the probability of $A \cap B$ relative to the probability of B, or

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if} \quad P(B) > 0.$$

Definition 0.3.1: (Conditional probability)

Let A and B be two events with $P(B) > 0$. The conditional probability that A occurs given B occurs is denoted by

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

Sometimes we may know $P(B)$ and $P(A|B)$, but not $P(A \cap B)$. If this is the case we can use the definition of conditional probability to express the probability that A and B both occur, that is

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

We can also extend this definition so that we can work with probabilities that are conditional on more than one event, that is,

$$\begin{aligned} P(A|B,C) &= \frac{P(A \cap B|C)}{P(B|C)}, \\ \text{and so} \quad P(A \cap B|C) &= P(A|B,C)P(B|C). \end{aligned}$$

0.4 Independent Events

The conditional probability $P(A|B)$ is not generally equal to $P(A)$, (the unconditional probability of A). When $P(A|B) = P(A)$ we say A is independent of B i.e. A is independent of B if the knowledge that B has occurred, does not change the probability that A occurs.

Definition 0.4.1: (Independence)

Two events A and B are *independent* if

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

This implies that A and B are *independent* if

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B)$$

Two events that are not independent are said to be *dependent*.

Example 0.4.1: (Toss a coin and then a die.)

Show that the events "6 on the die" and "H on the coin" are independent.

Solution:
The sample space is

$$S = \{(H,1)(H,2),\ldots,(H,6),(T,1)(T,2),\ldots,(T,6)\} \ .$$

Let $A =$ "6 on a dice" $= \{(H,6),(T,6)\}$.
Let $B =$ "H on coin" $= \{(H,1),(H,2),\ldots(H,6)\}$.
The event $A \cap B =$ "6 on a die and H on a coin" $= \{(H,6)\}$, therefore $P(A) = 1/6$, $P(B) = 1/2$ and $P(A \cap B) = 1/12$.
Note:
$P(A \cap B) = 1/12 = P(A)P(B)$, therefore A and B are independent. ◁

0.5 Law of Total Probability

Theorem 0.5.1 (Law of Total Probability)

Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space S. Then, for any event A of S,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \ .$$

Proof:
By Theorem 0.2.1, since $B_1, B_2, \ldots, B_n$ is a partition of S and hence are disjoint and exhaustive events, for any event A in S

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \quad \text{(conditional probability.)} \end{aligned}$$

∎

Example 0.5.1: (Elections)

In a city there are four Federal voting divisions (1,2,3, and 4). At the last election the proportion of electors that voted Democrat in each division was 0.09, 0.13, 0.12, 0.10 respectively. A voter is selected by choosing a Division and then choosing a voter. What is the probability that the voter selected voted Democrat.

Solution: The sample space is the set of all voters in the city.
Let $A_i,\ i = 1, 2, 3, 4$ be the event that "the selected voter lives in Division i" and let D be the event " the selected voter voted Democrat". We wish to find $P(D)$.
It is given that $P(A_i) = 1/4\ i = 1, 2, 3, 4$ and $P(D/A_1) = 0.09$, $P(D|A_2) = 0.13$, $P(D|A_3) = 0.12$ and $P(D|A_4) = 0.10$.
Also $A_i \cap A_j = \emptyset\ \ (i \neq j)$ (the selected voter cannot live in two divisions) and
$A_1 \cup A_2 \cup A_3 \cup A_4 = S$ (the selected voter must live in one division). Thus the events A_i partition S, giving us that

$$P(D) \;=\; \sum_{i=1}^{4} P(D|A_i)P(A_i) \;=\; \frac{1}{4}(0.09 + 0.13 + 0.12 + 0.10) \;=\; 0.11\ . \qquad \triangleleft$$

The Law of Total Probability is used regularly in Applied Probability, it is used to set up and solve equations concerning queues, teletraffic networks, epidemics, genetics and computer processors to name just a few applications.

0.6 Bayes' Theorem

Sometimes we want to write $P(B|A)$ in terms of $P(A|B)$. This is easily done from the definition of $P(B|A)$.

Theorem 0.6.1 (Bayes' Theorem)

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

Proof:

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(A \cap B)}{P(B)} \, \frac{P(B)}{P(A)} \\ &= \frac{P(A|B)P(B)}{P(A)} \end{aligned}$$

∎

Use of Bayes Theorem can yield some counter-intuitive facts.

Example 0.6.1: (HIV Test)

Suppose a test for HIV is 90% effective in the sense that if a person is HIV positive then the test has a 90% chance of saying they are HIV positive. On the other hand, if they are not HIV positive there is a 5% chance that the test says that they are. Assume that the percentage of the population that is HIV positive is (a very small) .01%.
Suppose a person tests positive for HIV. What is the probability they actually are HIV positive?

Let HIV+ be the event that the person is HIV positive and Test+ be the event that the person *tests* positive.

We know: $P(\text{Test}+ \mid \text{HIV}+) = 0.9$, $P(\text{Test}+ \mid \text{HIV}+^c) = 0.05$, $P(\text{HIV}+) = 0.0001$ and thus $P(\text{HIV}+^c) = 0.9999$ and we wish to determine $P(\text{HIV}+ \mid \text{Test}+)$.

Bayes Theorem, $P(\text{HIV}+ \mid \text{Test}+) = \dfrac{P(\text{Test}+ \mid \text{HIV}+)P(\text{HIV}+)}{P(\text{Test}+)}$, tells us we need to determine

$$\begin{aligned} P(\text{Test}+) &= P(\text{Test}+ \mid \text{HIV}+)P(\text{HIV}+) + P(\text{Test}+ \mid \text{HIV}+^c)P(\text{HIV}+^c) \\ &= 0.9 \times 0.0001 + 0.05 \times 0.9999 = 0.050085. \end{aligned}$$

So $P(\text{HIV}+ \mid \text{Test}+) = \dfrac{0.9 \times 0.0001}{0.050085} = 0.0019966 \approx \dfrac{1}{501}$. $\triangleleft$

If we combine the Law of Total Probability with Bayes' Theorem we get a more general version of Bayes' Theorem. Recall that the Law of Total Probability is that if $B_1, B_2, \ldots, B_n$ are disjoint and exhaustive events then, for any event A
$P(A) = P(A|B_1)PrB_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)$.

Theorem 0.6.2 (Bayes' Theorem)
Let $B_1, B_2, \ldots, B_n$ be a set of disjoint and exhaustive events of a sample space S. Then, for any event A and any $j \in \{1, 2, \ldots, n\}$,

$$Pr(B_j|A) = \frac{Pr(A|B_j)Pr(B_j)}{Pr(A|B_1)Pr(B_1) + Pr(A|B_2)Pr(B_2) + \ldots + Pr(A|B_n)Pr(B_n)}.$$

0.7 Random Variables

In many random experiments we are interested in some *function* of the outcome rather than the actual outcome itself. In these cases we wish to assign a real number x to each outcome w of the sample space S, so that $x = X(w)$ is the value of a (real-valued) *function* X from the sample space to the real numbers, $\mathbb{R}$.

Definition 0.7.1: (Random variable)

Consider a random experiment with sample space S. A *function* $X = X(w)$, assigning to every outcome $w \in S$ a real number, is called a random variable.

Notation:
Random variables will be denoted by capital letters and the values they take by small letters.

Since the value of a random variable is determined by the outcome of the experiment we may assign probabilities to the possible values of the random variable.

Definition 0.7.2:

If the real number x is a value of the random variable X, let $P(X = x)$ denote the probability of the event $E = \{w : X(w) = x\}$.

Thus $P(X = x) = P(E)$ and similarly $P(X \leq x) = P(F)$ where $F = \{w : X(w) \leq x\}$.

Example 0.7.1: (Toss two dice.)

The sample space is $S = \{(1, 1), \ldots, (6, 6)\}$.

If X denotes the random variable defined as the sum of the two faces and assuming each outcome is equally likely

$$\begin{aligned} P(X = 2) &= P\Big(\{w : X(w) = 2\}\Big) = P\Big(\{(1, 1)\}\Big) = 1/36, \\ P(X = 3) &= P\Big(\{w : X(w) = 3\}\Big) = P\Big(\{(1, 2), (2, 1)\}\Big) = 2/36 \end{aligned}$$

$$\begin{array}{lll} P(X = 4) = 3/36, & P(X = 5) = 4/36, & P(X = 6) = 5/36, \\ P(X = 7) = 6/36, & P(X = 8) = 5/36, & P(X = 9) = 4/36, \\ P(X = 10) = 3/36, & P(X = 11) = 2/36, & P(X = 12) = 1/36. \end{array}$$

Since X must take on one of the values from 2 to 12 we must have

$$\sum_{n=2}^{12} P(X = n) = 1, \qquad \text{which is easily checked.}$$

◁

Example 0.7.2: (Traffic lights)

Consider the process of driving through traffic lights until stopped by a red light. Let R = "light is red" and G = "light is green". Let $P(R) = p$ and assume the lights are independent. Let N denote the number of lights encountered before being stopped.

The sample space of the experiment is the set of sequences of the form $\omega = G, G, G, G, \ldots, G, R$.

N is the function from the sample space to the set of real numbers such that $N(G, G, G, \ldots, G, R) = n$ when $(G, G, G, \ldots, G, R)$ contains $n - 1$ G's.

Hence N is a random variable and
$$\begin{aligned} P(N = 1) &= P\big(\{\omega : N(\omega) = 1\}\big) &&= P(R) = p, \\ P(N = 2) &= P\big(\{\omega : N(\omega) = 2\}\big) &&= P(G, R) = (1 - p)p, \\ P(N = n) &= P\big(\{\omega : N(\omega) = n\}\big) && \\ &= P(G, G, \ldots, G, R) &&= (1 - p)^{n-1}p. \end{aligned}$$

◁

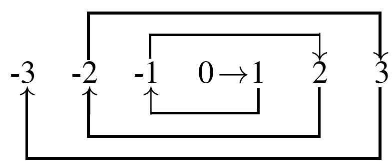
The above random variables all have a countable number of possible values.

Definition 0.7.3: (Countability)

A set is said to be *countable* if

- it is either finite or
- it can be put into a 1-1 correspondence with the set of natural numbers $\{1, 2, 3, \ldots\}$.

That is, a set is countable if it is possible to list its elements in the form $x_1, x_2, \ldots$, for example $N = \{1, 2, \ldots\}$, $Z = \{0, 1, -1, 2, -2, 3, \ldots\}$ are both countable.



Otherwise a set is *uncountable*

- Random variables that take a countable number of possible values are called *discrete* random variables.
- Random variables that take an uncountable number of possible values are called *continuous* random variables.

Example random variables

1. The label of a cell used by a mobile phone. *(discrete)*
2. Number of circuits occupied on a trunk. *(discrete)*
3. Duration of a call. *(continuous)*
4. Ages of stellar bodies. *(continuous)*

0.8 Probability Mass Functions and Probability Distribution Functions

Let X be a discrete random variable.

Definition 0.8.1: (Probability mass function)

The *probability mass function* $p(x)$ of X is given by

$$p(x) = P(X = x) \ .$$

Definition 0.8.2: (Distribution function)

The *(cumulative) distribution function* $F(x)$ of X is given by

$$F(x) = P(X \leq x) \ .$$

The term *mass function* arises by considering a total mass of one unit to be distributed over the integers (or whatever countable set we are considering).

The numbers $p(x_i)$ represent the amount of mass at x_i, while the numbers $F(x_i)$ represent the amount of mass up to and including x_i.

Theorem 0.8.1
The probability mass function $p(x)$ of a discrete random variable X satisfies the following:
(i) $p(x) \geq 0$, *for all x,*
(ii) $\sum_{x_i} p(x_i) = 1.$

Proof:
(i) $p(x) \geq 0$ since each $p(x)$ is a probability,
(ii) $\sum_{x_i} p(x_i) = 1$ since the events $E_i = \{\omega : x(\omega) = x_i\}$ are exhaustive. ∎

Theorem 0.8.2

For a random variable X with probability mass function $p(x)$ and (cumulative) distribution function $F(x)$

1. $F(x) = \sum_{x_i \leq x} p(x_i)$.
2. $\lim_{x \to -\infty} F(x) = 0$.
3. $\lim_{x \to +\infty} F(x) = 1$.
4. *For $a < b$,* $P(a < X \leq b) = F(b) - F(a)$.
5. $F(x)$ *is a non-decreasing function of x.*

Proof:

1. The event A = $\{\omega : X(\omega) \leq x\}$ is the union of the disjoint events $B_i = \{\omega : X(\omega) = x_i, x_i \leq x\}$. Therefore $F(x) = P(A) = \sum_i P(B_i) = \sum_{x_i \leq x} p(x_i)$.

2. $\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} P(X \leq x) = P(\emptyset) = 0.$

3. $\lim_{x \to \infty} P(x) = \lim_{x \to \infty} P(X \leq x) = P(S) = 1$.

4. For $a < b$, $\{\omega : X(\omega) \leq b\} = \{\omega : X(\omega) \leq a\} \cup \{\omega : a < X(\omega) \leq b\}$, and the events on the right hand side are disjoint. Therefore $P(X \leq b) = P(X \leq a) + P(a < X \leq b)$ and so $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$.

5. If $a < b$ then by (iv) above $F(b) - F(a) = P(a < X \leq b)$ but $P(a < X \leq b) \geq 0$ by probability axiom 1 and so F is non-decreasing. ∎

Example 0.8.1: (Toss two fair dice.)

What is the probability mass function $p(x)$ and the cumulative distribution function $F(x)$ for the random variable X which is the sum of the two dice?

Solution: The sample space is $S = \{(1,1), \ldots, (6,6)\}$, then the probability mass function and cumulative distribution functions are given by

$$\begin{array}{lll}
P(2) = P(X=2) = 1/36, & P(3) = P(X=3) = 2/36, & P(4) = 3/36, \\
P(5) = 4/36, & P(6) = 5/36, & P(7) = 6/36, \\
P(8) = 5/36, & P(9) = 4/36, & P(10) = 3/36, \\
P(11) = 2/36, & P(12) = 1/36. &
\end{array}$$

$$\begin{array}{lll}
F(2) = 1/36, & F(3) = 3/36, & F(4) = 6/36, \\
F(5) = 10/36, & F(6) = 15/36, & F(7) = 21/36, \\
F(8) = 26/36, & F(9) = 30/36, & F(10) = 33/36, \\
F(11) = 35/36, & F(12) = 1. &
\end{array}$$

◁

0.9 Expectation of a Random Variable

Definition 0.9.1: (Expected value)

Let X be a discrete random variable with possible values $x_1, x_2, \ldots, x_n$ and probability mass function $p(x_i) = P(X = x_i)$, $\quad i = 1, 2, \ldots, n$. The *expected value* (or *mean value* of X), denoted by $E[X]$, is defined by

$$E[X] = \sum_{i=1}^{n} x_i p(x_i).$$

Example 0.9.1: (Toss a die)

Find $E[X]$ if X is the outcome of a toss of a fair dice.
Solution: Since $p(1) = p(2) \ldots = p(6) = 1/6$, we have

$$E[X] = 1(\frac{1}{6}) + 2(\frac{1}{6}) + \ldots + 6(\frac{1}{6}) = \frac{7}{2}.$$

$\triangleleft$

Note: $E[X]$ is not necessarily a possible outcome of the random experiment as in the previous example.

Example 0.9.2: (Manufacturing)

A manufacturer produces items of which 10% are defective and 90% are non-defective. If a defective item is produced the manufacturer loses $1 while a non-defective item yields a profit of $5. If X is the net profit per item then X is a random variable with expected value

$$E[X] = -1(0.1) + 5(0.9) = \$4.40$$

$\triangleleft$

Suppose we are interested in calculating the expected value of a function $g(X)$ of the random variable X.

Definition 0.9.2:

If X is a discrete random variable with probability mass function $p(x)$, then for any real-valued function g

$$E[g(X)] = \sum_x g(x)p(x)$$

Example 0.9.3: (Toss a fair die)

Find $E[X^2]$ if X is the outcome of the toss of a fair die.
Solution: $E[X^2] = 1^2\frac{1}{6} + 2^2\frac{1}{6} + 3^2\frac{1}{6} + \ldots + 6^2\frac{1}{6} = \frac{91}{6}$.
Note: $\frac{91}{6} = E[X^2] \neq (E[X])^2 = (\frac{7}{2})^2 = \frac{49}{4}$. $\triangleleft$

Theorem 0.9.1 *If X is a discrete random variable and a and b are constants then*

$$E[aX + b] = aE[X] + b.$$

Proof:

$$\begin{aligned} E[aX + b] &= \sum_x (ax + b)p(x) \\ &= \sum_x axp(x) + \sum_x bp(x) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE[X] + b\ . \end{aligned}$$

∎

Definition 0.9.3: (Variance)

The *variance* of a random variable X, which measures the spread or dispersion of the values of X, denoted by var(X) is defined by

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

It may be regarded as a measure of the *consistency* of outcome, a smaller value of var(X) implies that X is more often near $E[X]$ than for a larger value of var(X).

Theorem 0.9.2

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

Proof:

$$\begin{aligned}\text{var}(X) &= E[(X - E[X])^2] \;=\; E[X^2 - 2XE[X] + [E[X]]^2] \\ &= \sum_x (x^2 - 2E[X]x + [E[X]]^2)p(x) \;=\; E[X^2] - [E[X]]^2.\end{aligned}$$ ∎

Example 0.9.4: (Cricket)

Consider the batting performance of two cricketers, one of whom hits a century (exactly 100) with probability 0.5 or who gets a duck (0) with probability 0.5. The other always scores 50 runs.

If X_1 is the number of runs for the first cricketer and X_2 the number of runs for the second then $E[X_1] = E[X_2] = 50$.

However

$$\begin{aligned}\text{var}(X_1) &= 0.5(50-0)^2 + 0.5(50-100)^2 = 2,500 \\ \text{var}(X_2) &= 0\end{aligned}$$

which reflects the fact that the second cricketer is more consistent. ◁

Example 0.9.5: (Toss a fair die)

Calculate var(X) when X is the outcome of the toss of a fair die.
From above, $E[X] = 7/2$ and $E[X^2] = 91/6$.
Therefore var$(X) = 91/6 - 49/4 = 35/12$. ◁

Example 0.9.6: (Manufacturing)

In the machine manufacturing example $E[X] = 4.40$ and $E[X^2] = (-1)^2(0.1)+(5)^2(0.9) = 22.60$.
Therefore var$(X) = 22.6 - 19.36 = 3.24$. ◁

Theorem 0.9.3

$$\text{var}(aX + b) = a^2\text{var}(X)$$

Proof:
Let $Y = aX + b$ and so $E[Y] = aE[X] + b$.

$$\text{Now} \qquad \Big(Y - E[Y]\Big)^2 = a^2\Big(X - E[X]\Big)^2$$

$$\begin{aligned} \text{and} \qquad \text{var}(aX + b) &= \text{var}(Y) \\ &= E\Big[(Y - E[Y])^2\Big] \\ &= a^2E\Big[(X - E[X])^2\Big] \\ &= a^2\text{var}(X). \end{aligned}$$

∎

0.10 Continuous Random Variables

For a continuous random variable, the probability mass associated with any single point is zero and the mass function must be replaced by a density function. The density function, when integrated over an interval of the possible values of the random variable, gives the probability that the random variable takes one of the values in that interval.

If the random variable X has probability density function $f(x)$ then

$$P(a < X \leq b) = \int_a^b f(u)du$$

and the distribution function, $F(x)$, is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du.$$

The above definitions, properties and theorems for discrete random variables can be parallelled directly for continuous random variables by replacing sum by integral and mass function by density function.

If X is a continuous random variable with density function $f(x)$ and distribution function $F(x)$ then the following results hold.

1. $f(x) \geq 0 \quad$ for all $x \in (-\infty, \infty)$
2. $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1$
3. $\displaystyle\lim_{x \to -\infty} F(x) = 0$
4. $\displaystyle\lim_{x \to \infty} F(x) = 1$
5. $P(a < X \leq b) = F(b) - F(a) \quad$ for $a < b$
6. $E[X] = \displaystyle\int_{-\infty}^{\infty} x \ f(x)dx$
7. $E[g(X)] = \displaystyle\int_{-\infty}^{\infty} g(x)f(x)dx$
8. $[aX + b] = aE[X] + b$
9. $\text{var}(X) = E\Big[\big(X - E[X]\big)^2\Big]$
10. $\text{var}(X) = E[X^2] - \Big[E[X]\Big]^2$
11. $\text{var}(aX + b) = a^2\text{var}(X)$

0.11 Some Common Discrete and Continuous Distribution Functions

Certain distribution functions occur sufficiently often to make their study useful. These distribution functions tend to occur in families, where they are indexed by one or more unknown parameters.

In these notes we will briefly introduce some of the more common distributions, together with an indication of their applications.

Recall that you will find that the exponential distribution occurs very frequently in the work we will be doing in this course and it is therefore a good idea to know this distribution very well. For example, the Poisson process has times between events that are exponentially distributed.

0.11.1 The Binomial Distribution

Consider a sequence of *Bernoulli Trials* (for example, experiments that result either in "success" or "failure"). If X is the random variable which counts the number of successes in n trials and trials are independent with $P(\text{success}) = p$ then we say X is *binomially distributed with parameters* n *and* p and write

$$X \sim b(n,p)\,.$$

The probability mass function of X is $p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. This follows because the probability of a given string including x successes and $n-x$ failures, e.g.

$$ssffssfs\ldots$$

is

$$p^x(1-p)^{n-x}$$

and there are $\displaystyle\binom{n}{x} = \frac{n!}{(n-x)!\,x!}$ such strings.

A binomially distributed random variable X has the following properties:

$$E[X] = np$$

$$\text{var}(X) = np(1-p).$$

Proofs:
Left to the reader.

An elegant proof is to use the properties of sums of *independent* random variables. Recall that the Binonmial random variable consists of counting the number of successes in n *independent* Bernoulli trials and the probability of success in a Bernoulli trial is p.

0.11.2 The Geometric Distribution

Again we look at a sequence of Bernoulli Trials but count a different thing.

Let $X =$ number of trials needed before the first success occurs. Then

$$p(x) = P(X = x) = (1-p)^{x-1}p\ ,$$

since the only string that has the required form is

$$\underbrace{fff\ldots f}_{x-1}\,s$$

and this has probability $(1-p)^{x-1}p$.
Such a random variable X is said to be *geometrically distributed* with parameter p.

A *geometrically distributed* random variable X has the following properties:

$$E[X] = \frac{1}{p}$$

$$\text{var}(X) = \frac{(1-p)}{p^2}.$$

Proof: (of $E[X]$)

$$\begin{aligned}
E[X] &= \sum_{x=1}^{\infty} x(1-p)^{x-1}p \;=\; p\sum_{x=1}^{\infty} x(u)^{x-1}\,, \qquad \text{where } u = 1-p \\
&= p\frac{d}{du}\sum_{x=0}^{\infty} u^x \;=\; p\frac{d}{du}\frac{1}{1-u} \\
&= p\frac{1}{(1-u)^2} \;=\; \frac{1}{p}.
\end{aligned}$$

∎

0.11.3 The Poisson Distribution

The Poisson distribution first arose (Poisson (1837)) as an approximation for the binomial distribution for large n. If we let $n \to \infty$ and $p \to 0$ with $np = \lambda$ held fixed then

$$\begin{aligned}
&\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} \qquad = \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty}\frac{n!}{x!(n-x)!}\lambda^x\left(\frac{1}{n^x}\right)\left(1-\frac{\lambda}{n}\right)^{-x}\left(1-\frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^x}{x!}\lim_{n\to\infty}\frac{n!}{(n-x)!}\frac{1}{(n-\lambda)^x}\left(1-\frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^x}{x!}\lim_{n\to\infty}\frac{n}{n-\lambda}\frac{n-1}{n-\lambda}\cdots\frac{n-x+1}{n-\lambda}\left(1-\frac{\lambda}{n}\right)^{n} \qquad = \frac{\lambda^x}{x!}e^{-\lambda}
\end{aligned}$$

In approximating $X \sim b(n,p)$, Poisson suggested using

$$P(X=x) \simeq \frac{(np)^x}{x!}e^{-np}$$

if n is large and p small.

However, it turns out that the probability mass function above can be used directly in many problems in which there are many possible events which may occur with each having a very small probability.

The Poisson distribution has been used to model.

1. Deaths/year due to a horse kick in the Prussian cavalry.
2. Number of telephone calls arriving at a telephone exchange in a given time.
3. Number of accidents on a stretch of road per unit time.

4. Number of surviving octopusses in a given area.

If X is a Poisson random variable with parameter λ, it has probability mass function

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}\ , \quad x \geq 0,$$

in which case X has the following properties:

$$E[X] = \lambda\,.$$

$$\text{var}(X) = \lambda\,.$$

Proof:
The properties follow either from the limits of the mean and variance of the binomial distribution or directly from the probability mass function above. ∎

0.11.4 The Uniform Distribution

The uniform distribution is a continuous distribution that assigns equal probability density to all points between a and b and zero to other points.

It has probability density function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

The uniform distribution function is

$$F(x) \;=\; \int_{-\infty}^{x} f(t)dt \;=\; \begin{cases} 0 & \text{if } x \leq a \\ \displaystyle\int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a} & \text{if } a < x \leq b \\ 1 & \text{if } x > b. \end{cases}$$

If X is a uniformly distributed random variable then

$$E[X] = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] = \frac{a+b}{2}$$

and

$$\begin{aligned}
\text{var}(X) &= E[X^2] - E[X]^2 = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a}\left[\frac{b^3}{3} - \frac{a^3}{3}\right] - \frac{a^2+2ab+b^2}{4} \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}\,.
\end{aligned}$$

0.11.5 The Exponential Distribution

The exponential distribution models the time that elapses until a randomly generated event occurs (for example an arrival at a queue). We can derive the probability density function for the exponential distribution from the probability mass function of the geometric distribution in the following way.

Assume events happen at rate λ per unit time, and that we sub-divide the time axis into intervals of length Δt small enough so that the probability of more than one event in a given subinterval is negligible. The probability of an event occurring in a given subinterval is then $\lambda\Delta t$. Therefore the probability that it will take x intervals for the first event to occur is

$$\lambda\Delta t\Big(1 - \lambda\Delta t\Big)^{x-1} \qquad \text{(geometric)}\,.$$

Approximating a continuous distribution, with density function $f(t)$, by this discrete distribution means we are using $f(t)\Delta t$ in place of the probability $\lambda\Delta t(1-\lambda\Delta t)^{x-1}$ at least in the limit as $\Delta t \to 0$. Now let $x \to \infty$ and $\Delta t \to 0$ in such a way that $x\Delta t = t$ for some t. Then

$$f(t) \;= \lim_{x\to\infty} \lambda \left(1 - \frac{\lambda t}{x}\right)^{x-1} \;= \lim_{x\to\infty} \lambda \left(1 - \frac{\lambda t}{x}\right)^{x} \left(\frac{x}{x-\lambda t}\right) \;= \lambda e^{-\lambda t}\,.$$

A random variable T with probability density function given by

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

is said to be exponentially distributed with parameter λ.

If T is an exponentially distributed random variable, it has the following properties

$$E[T] = \frac{1}{\lambda}\,.$$

$$\text{var}(T) = \frac{1}{\lambda^2}\,.$$

Proof:

$$\begin{aligned} E[T] &= \int_0^\infty t\lambda e^{-\lambda t}dt \;=\; \lambda\left[\frac{te^{-\lambda t}}{-\lambda}\right]_0^\infty - \lambda\int_0^\infty \frac{e^{-\lambda t}}{\lambda}dt \\ &= 0+\int_0^\infty e^{-\lambda t}dt \;=\; \left[\frac{e^{-\lambda t}}{-\lambda}\right]_0^\infty = \frac{1}{\lambda}. \\ \text{var}(T) &= E[T^2]-[E[T]]^2 \;=\; \int_0^\infty t^2\lambda e^{-\lambda t}dt - \frac{1}{\lambda^2} \\ &= \lambda\left[t^2\frac{e^{-\lambda t}}{-\lambda}\right]_0^\infty - \lambda\int_0^\infty \frac{2te^{-\lambda t}}{-\lambda}dt - \frac{1}{\lambda^2} \\ &= 0+2\int_0^\infty te^{-\lambda t}dt - \frac{1}{\lambda^2} \\ &= \frac{2}{\lambda}E[T]-\frac{1}{\lambda^2} \;=\; \frac{2}{\lambda^2}-\frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2}\,. \end{aligned}$$

∎

0.11.6 The Gamma or Erlang Distribution

The *Erlang distribution* of order r models the distribution of time until r randomly generated events have occurred. Its density function can be derived from that of the negative binomial in a similar way to that in which the exponential density function is derived from the geometric probability mass function and is

$$f(t)=\begin{cases} 0, & t<0, \\ \lambda e^{-\lambda t}\dfrac{(\lambda t)^{r-1}}{(r-1)!}, & t\geq 0\,. \end{cases}$$

Its distribution function can be derived by integrating this but is very messy to write down.

Since the length of time until r independent events has occurred can be viewed as r independent lengths of time until a single event has occurred, that is r exponential random variables, we have, for an Erlang random variable T,

$$\begin{aligned} E[T] &= rE[\text{ exponential random variable}] \\ &= \frac{r}{\lambda}\,, \qquad (1) \\ \text{var}(T) &= r\text{ var}[\text{ exponential random variable}] \\ &= \frac{r}{\lambda^2}\,. \qquad (2) \end{aligned}$$

The Erlang distribution can be generalised to the case where r is not an integer. We do this by introducing the continuous analogue of $(r-1)!$ which is the gamma function $\Gamma(r)$ defined as

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx \, .$$

Thus if T is a random variable with probability density function

$$f(t) = \frac{\lambda(\lambda t)^{r-1} e^{-} \lambda t}{\Gamma(r)} \, ,$$

we say that T is distributed according to the *gamma distribution* with parameters r and λ and it obeys properties 1 and 2 above.