0

Assumed knowledge

Detailed objectives for assumed knowledge

1 Summarise the main features of a data set (exploratory data analysis)

- 1.1 Summarise a set of data using a table or frequency distribution, and display it graphically using a line plot, a box plot, a bar chart, histogram, stem and leaf plot, or other appropriate elementary device.
- 1.2 Describe the level/location of a set of data using the mean, median, mode, as appropriate.
- 1.3 Describe the spread/variability of a set of data using the standard deviation, range, interquartile range, as appropriate.
- 1.4 Explain what is meant by symmetry and skewness for the distribution of a set of data.

2 Probability

- 2.1 Set functions and sample spaces for an experiment and an event.
- 2.2 Probability as a set function on a collection of events and its basic properties.
- 2.3 Calculate probabilities of events in simple situations.
- 2.4 Derive and use the addition rule for the probability of the union of two events.

Detailed objectives for assumed knowledge – continued

2.5 Define and calculate the conditional probability of one event given the occurrence of another event.

2.6 Derive and use Bayes' Theorem for events.

2.7 Define independence for two events, and calculate probabilities in situations involving independence.

3 Random variables

3.1 Explain what is meant by a discrete random variable, define the distribution function and the probability function of such a variable, and use these functions to calculate probabilities.

3.2 Explain what is meant by a continuous random variable, define the distribution function and the probability density function of such a variable, and use these functions to calculate probabilities.

3.3 Define the expected value of a function of a random variable, the mean, the variance, the standard deviation, the coefficient of skewness and the moments of a random variable, and calculate such quantities.

3.4 Evaluate probabilities associated with distributions (by calculation or by referring to tables as appropriate).

3.5 Derive the distribution of a function of a random variable from the distribution of the random variable.

0 Introduction

There are a lot of topics that are assumed knowledge for the CS1 exam. This material can be tested implicitly (*ie* part of a question on the remainder of the CS1 syllabus relies on the assumed knowledge) or explicitly (*ie* a whole question in the CS1 exam tests material contained in the assumed knowledge).

Either way, it is essential to understand the objectives listed on the previous pages.

The rest of this chapter gives a brief outline of the material and some questions to help understanding.

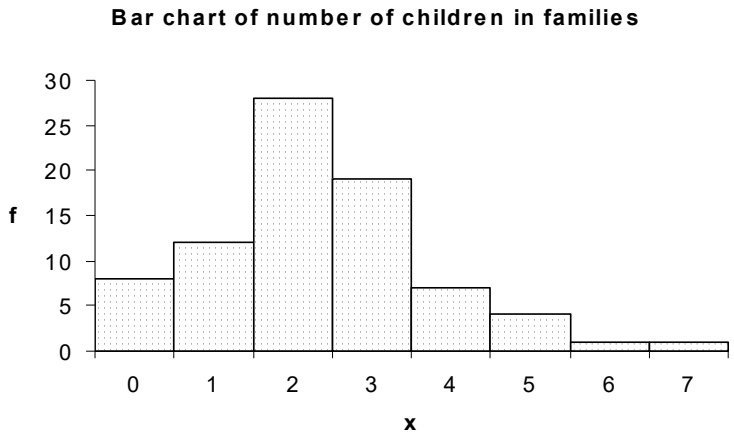
1 Tables and graphs

1.1 Frequency distribution

The data values from a discrete distribution can be summarised using a frequency distribution. For example, the number of children in a sample of 80 families can be shown as a frequency distribution as follows:

Number of children under 16, x	Number of families in sample, f
0	8
1	12
2	28
3	19
4	7
5	4
6	1
7	1
8 or more	0

A bar chart can be used to display the data values graphically.



1.2 Histograms and grouped frequency distributions

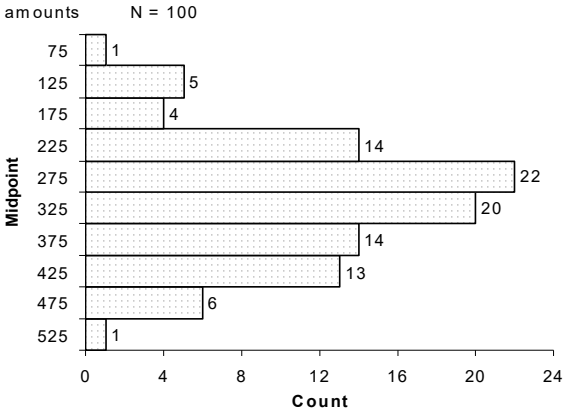
A sample of 100 claims for damage due to water leakage on an insurance company's household contents policies is as follows:

243	306	271	396	287	399	466	269	295	330
425	324	228	113	226	176	320	230	404	487
127	74	523	164	366	343	330	436	141	388
293	464	200	392	265	403	372	259	426	262
221	355	324	374	347	261	278	113	135	291
176	342	443	239	302	483	231	292	373	346
293	236	223	371	287	400	314	468	337	308
359	352	273	267	277	184	286	214	351	270
330	238	248	419	330	319	440	427	343	414
291	299	265	318	415	372	238	323	411	494

These data values can be summarised in the following grouped frequency distribution:

Group	Frequency
$50 \le x < 100$	1
$100 \le x < 150$	5
$150 \le x < 200$	4
$200 \le x < 250$	14
$250 \le x < 300$	22
$300 \le x < 350$	20
$350 \le x < 400$	14
$400 \le x < 450$	13
$450 \le x < 500$	6
$500 \le x < 550$	1

The data values can also be presented in a histogram. The continuous scale has been broken down into categories by dividing the data values into bands. The bar labelled '125', for example, represents the 5 data values (113, 127, 141, 113 and 135) that are in the group 100–. Histograms can be presented with vertical or horizontal rectangles.



The above grouped frequency distribution and histogram have equal group widths. In some situations it may be convenient to have one or two wider groups at the extremes of the distribution. For such cases it is the areas of the rectangles that are proportional to the frequencies not their heights.

1.3 Stem and leaf diagrams

A stem and leaf diagram gives a visual representation similar to a histogram but does not lose the detail of the individual data points. A stem and leaf diagram for the water leakage data claim amounts is as follows:

0	7
1	11344
1	6888
2	0122333344444
2	566677777778899999999
3	00011122222233334444
3	55556677777799
4	000001122333444
4	677899
5	2

The stems are on the left with units of 100 and the leaves are on the right with units of 10 (this can be shown as a key on the diagram). The individual data points are represented, although they are rounded to the nearest 10.

These diagrams are useful to observe the general shape of the distribution of data and can be used to calculate values such as the median or interquartile range. We will cover these ideas later in this chapter.

1.4 Line plots

For smaller data sets another alternative diagram is the *dotplot* or *lineplot* in which the data points are plotted as 'dots' or 'crosses' along a line with a scale.

Here is a lineplot for the first row of 10 amounts from the claims data:

If there are repeats of any data points (*eg* two values of 300) in a lineplot, the crosses are placed above each other.

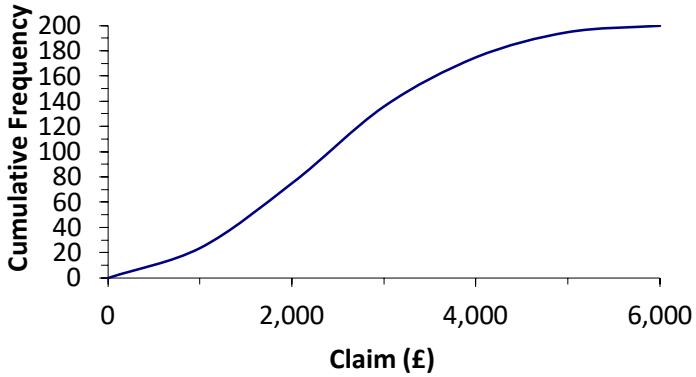
A lineplot can be used to check the shape of a data set and to check whether two data sets have a similar spread.

1.5 Cumulative frequency tables and graphs

Cumulative frequencies are obtained by summing the frequencies to give the total number of observations up to and including the value or group in question. The following table shows the cumulative frequencies for 200 motor insurance claims received by an office in a month.

Claim size	Cumulative frequency
up to £1,000	24
up to £2,000	75
up to £3,000	136
up to £4,000	175
up to £5,000	195
up to £6,000	200

A cumulative frequency diagram can be drawn from these data values as follows:



The cumulative frequency is plotted against the *highest* claim size in the group.

A cumulative frequency table or diagram can be used to determine the median or interquartile range of the data. The interquartile range is covered in Section 2.7 of this chapter.

2 Measures of location and spread

2.1 The sample mean

For a set of observations denoted by $x_1, x_2, \ldots, x_n$ or $x_i, i = 1,2,\ldots,n$ the mean is defined by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

For a frequency distribution with possible values $x_1, x_2, \ldots, x_k$ with corresponding frequencies $f_1, f_2, \ldots, f_k$, where $\sum f_i = n$, the mean is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\, x_i$$

Question

Calculate an approximate value of the mean of the water leakage data (repeated below for convenience).

Group	Frequency
$50 \le x < 100$	1
$100 \le x < 150$	5
$150 \le x < 200$	4
$200 \le x < 250$	14
$250 \le x < 300$	22
$300 \le x < 350$	20
$350 \le x < 400$	14
$400 \le x < 450$	13
$450 \le x < 500$	6
$500 \le x < 550$	1

Solution

An approximate value of the mean is:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1\times 75 + 5\times 125 + \cdots + 1\times 525}{1 + 5 + \cdots + 1} = \frac{31{,}250}{100} = 312.5$$

This figure is an approximation because we have used the midpoints of each group and we have assumed that the values in each group are evenly distributed over the group.

2.2 The median

The median is a value that splits the data set of n values into two equal halves, so that half the observations are less than the median and half are greater than the median.

If n is odd, then the median is the middle observation. If n is even, then the median is the midpoint of the middle two observations. This is expressed as the $(n+1)/2$ th observation.

One of the advantages of the median is that it is resistant to the effects of extreme observations These can cause problems because they can have a disproportionate effect on the calculated value of some statistics.

We can't calculate the median of the water leakage data from the previous question *exactly* because we don't know what the actual data values are. We would assume that the cumulative frequency increases linearly over each class and use interpolation. Some statisticians then argue that, under these assumptions, the position of the median is better chosen to be $\frac{1}{2}n$.

The median is the value corresponding to a cumulative frequency of 50% so it can be read from a plot of the cumulative frequency distribution.

Question

The ages of seven policyholders in a portfolio of insurance policies are as follows:

26 28 34 34 39 41 70

(i) Determine the median age of the policyholders in this portfolio.

(ii) Another policyholder aged 41 years is added to the portfolio. Determine the median age of policyholders in the portfolio.

(iii) Explain why the mean is a poor measure of location for these data values.

Solution

(i) The data values are in order so the median is the $(7+1)/2 = 4$ th observation, which is 34.

(ii) The ages in increasing order of magnitude are now:

26 28 34 34 39 41 41 70

The median is the $(8+1)/2 = 4.5$ th observation, which is $(34+39)/2 = 36.5$.

(iii) The mean for the original seven policyholders is 38.86. This figure has been affected by the extreme value of 70, whereas the median is more *robust* (resistant to the effects of extreme observations).

2.3 The mode

Another measure of location is the mode. It is defined as the value that occurs with the greatest frequency. Its use in practice is limited but there are occasions when, for example, a company is interested in the most typical policyholder.

2.4 The sample standard deviation and variance

The symbol s^2 is used to denote the sample variance. For a data set x_i, $i = 1,2,\ldots,n$ with mean $\bar{x}$, the sample variance is:

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

The standard deviation is the positive square root of the variance.

The sample standard deviation and sample variance can be calculated more easily using the alternative formula:

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - n\bar{x}^2\right]$$

We divide by $n-1$ to make the sample variance an *unbiased estimator* of the population variance. This concept will be covered in Subject CS1.

If we have the raw data we can use the statistical functions on a calculator to calculate the sample standard deviation and variance.

Question

For a set of data we have:

$$n = 100 \quad \sum x = 31{,}353 \quad \sum x^2 = 10{,}687{,}041$$

Calculate the:

(i) sample mean

(ii) sample standard deviation.

Solution

(i) $\bar{x} = \frac{31{,}353}{100} = £313.53$

(ii) $s^2 = \frac{1}{99}\left[10{,}687{,}041 - 100 \times 313.53^2\right] = 8{,}655.91$

So the sample standard deviation is:

$$s = \sqrt{8{,}655.91} = £93.04$$

2.5 Moments

The mean and variance are special cases of a set of summary measures called the *moments* of a set of data. In general the k th-order moment about the value α is defined by:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^k$$

So the mean is the first-order moment about the origin, and the variance is the second-order moment about the mean with a divisor of $n-1$ rather than n.

When α equals the mean of the distribution, we call the moment a *central moment*.

2.6 The range

The *range* is defined to be the difference between the largest and smallest observations in the data set.

$$R = \max_i(x_i) - \min_i(x_i)$$

The range is a poor measure of the spread of the data as it relies on the extreme values, which aren't necessarily representative of the data set as a whole.

2.7 The interquartile range

In order to calculate the interquartile range we must define the quartiles of a set of data. The quartiles divide a set of data into four quarters (just as the median divides a set of data into two halves). The quartiles are denoted by Q_1, Q_2 and Q_3. Q_2 is the median, Q_1 is called the lower quartile and Q_3 is called the upper quartile.

When we have discrete ungrouped data, we determine the quartiles in a similar way to that used to calculate the median. The median is the $(n+1)/2$ th observation, Q_1 can be defined to be the $(n+2)/4$ th observation counting from below and Q_3 as the same counting from above, with relevant interpolation if needed. There are several different formulae for the position of the quartiles. For example, one alternative is:

Q_1 is the $\frac{n+1}{4}$ th observation counting from below

Q_3 is the $\frac{n+1}{4}$ th observation counting from above

When we have grouped data, the quartiles can be read from the cumulative frequency diagram or calculated using interpolation (and some statisticians then argue that, under these assumptions, the position of the quartiles are better chosen to be $\frac{1}{4}n$ and $\frac{3}{4}n$).

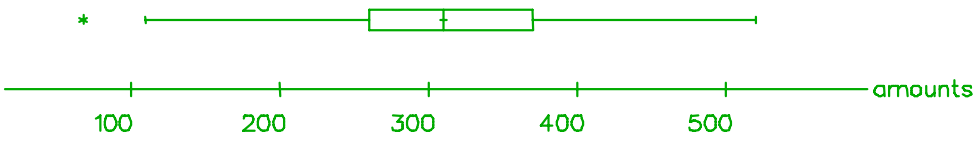
The lower quartile, the median and the upper quartile are also referred to as the 25th, 50th and 75th percentiles. The p th percentile corresponds to the $\left(\frac{p}{100}\times n+\frac{1}{2}\right)$ th data value.

The *interquartile range* (IQR) is defined to be $Q_3 - Q_1$. The IQR is a measure of spread is not affected by extreme values of the data.

Boxplots

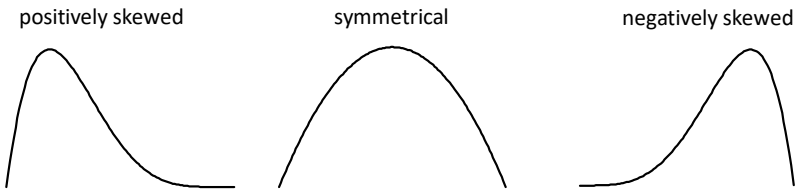
Now that we have looked at quartiles we can consider using a *boxplot* to present a data set graphically. It consists of a box or rectangle with ends at Q_1 and Q_3 divided with a line at the median Q_2. Then lines are drawn from Q_1 and Q_3 out to the extreme values in the data set.

Here is a boxplot of some claim amount data. Potential 'outliers' (observations which are in some way detached from the main bulk of the data set) are shown by an asterisk. In actuarial work outliers may sometimes be values that have been recorded incorrectly, *eg* a pensioner's age given as 27 instead of 72, or an amount of money typed with an extra or missing 0 on the end. However, they may also be a genuine part of the data.



3 Symmetry and skewness

The approximate shape of a distribution can be determined by looking at a histogram or dotplot. The approximate shapes of a positively skewed, symmetrical and negatively skewed distribution are as follows:



Positively skewed distributions are the most common ones used in actuarial work because we are often dealing with quantities, such as claim amounts, which must be positive but have no upper limit. Examples of positively skewed distributions are the Poisson, exponential, gamma and lognormal distributions.

A small data set (say 10 observations) is unlikely to show up the skewness of a population unless its skewness is very severe, whereas a large data set (say 200 observations) will better reflect the shape of the population.

One measure of skewness is based on the third moment about the mean:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$$

The cubic power in this formula gives a positive or negative value depending on which side of the mean x_i is. Consequently, positively skewed distributions with a long tail on the right give a positive value, and negatively skewed distribution with a long tail on the left give a negative value.

The coefficient of skewness is a scaled version of this moment, obtained by dividing it by the second moment about the mean raised to the power of 1.5. This is not quite the same as dividing by the sample variance to the power of 1.5 because the denominator of the second moment is n whereas the denominator of the sample variance is $n-1$. The coefficient of skewness is a dimensionless measure.

Question

For a set of data we have:

$$n=100 \quad \sum(x_i-\bar{x})^2=856{,}934.91 \quad \sum(x_i-\bar{x})^3=-11{,}949{,}848.3946$$

Calculate the skewness and the coefficient of skewness.

Solution

The skewness is given by:

$$\frac{\sum(x_i-\bar{x})^3}{n}=-\frac{11,949,848.3946}{100}=-119,498.483946$$

The second central moment is:

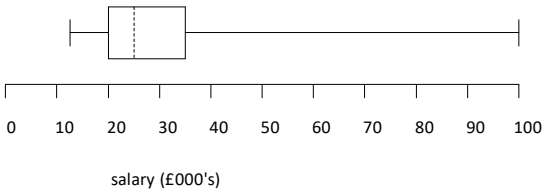
$$\frac{\sum(x_i-\bar{x})^2}{n}=\frac{856,934.91}{100}=8,569.3491$$

Hence the coefficient of skewness is given by:

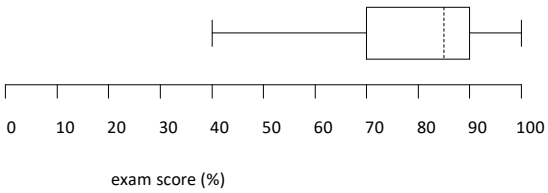
$$-\frac{119,498.483946}{8,569.3491^{1.5}}=-0.15064$$

We will now revisit boxplots from the previous section and use them to consider the shape of a data set. We will consider two example data sets.

The annual salaries of the employees of a medium-sized company are likely to have a positively skewed distribution. A possible boxplot could be:



The scores obtained by a class of students in an easy exam are likely to have a negatively skewed distribution. A possible boxplot could be:



4 Probability

4.1 Set notation and operations

A set is defined as a collection of objects and each individual object is called an element of that set.

If an experiment is defined as an operation whose outcome cannot be predicted in advance with certainty, the sample space S is the set of all possible outcomes that might be observed. For example, if the experiment is one roll of a normal six-sided die, the sample space would be defined as $S = \{1, 2, 3, 4, 5, 6\}$, *ie* all the possible numbers that can be rolled.

An event A is defined as a subset of the sample space S, containing any element of S. It is written as $A \subset S$.

In the example above, we could define throwing an even number as an event, therefore $A = \{2, 4, 6\}$. The event is said to occur if any one of its elements is the outcome observed, for example if a 2 is rolled.

The null set $\varnothing$ is the set with no elements.

An event is said to be complementary to another, A, in a sample space, S, if it contains all the elements of S that are not in A. The complement of A is normally written as A' or $\overline{A}$. In the example of rolling a die, where $S = \{1, 2, 3, 4, 5, 6\}$, if $A = \{2, 4, 6\}$ then $A' = \{1, 3, 5\}$.

The *union* of two sets, A and B, written as $A \cup B$, is the set that consists of all the elements that belong to A or B or both. For example, if $A = \{2, 4, 6\}$ and $B = \{1, 4\}$ then $A \cup B = \{1, 2, 4, 6\}$.

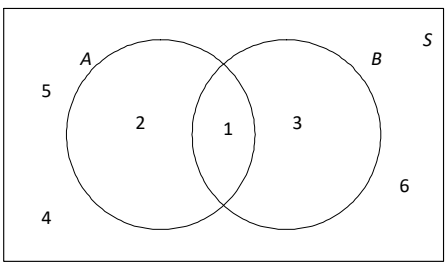
The *intersection* of two sets, A and B, written as $A \cap B$ is the set that consists of all elements that belong to both A and B. So in the example above, $A \cap B = \{4\}$.

If there are no elements common to both sets, they are known as *mutually exclusive*. For example, if $A = \{2, 4, 6\}$ and $C = \{1, 5\}$ then $A \cap C = \varnothing$, the null set.

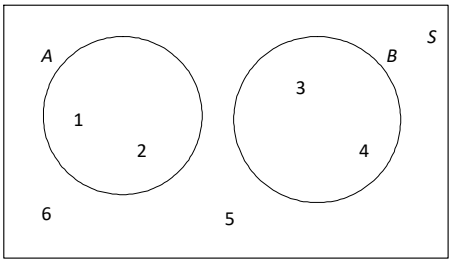
4.2 Venn diagrams

A convenient way to represent sets is by drawing Venn diagrams.

The Venn diagram below for the roll of a die shows the sets $A = \{1, 2\}$ and $B = \{1, 3\}$. Since 1 is in both sets it is placed in the overlap (the intersection) between A and B.



The Venn diagram below for the roll of a die shows the sets $A = \{1, 2\}$ and $B = \{3, 4\}$. These sets are mutually exclusive (*ie* they have no elements in common). They are drawn with no overlap.



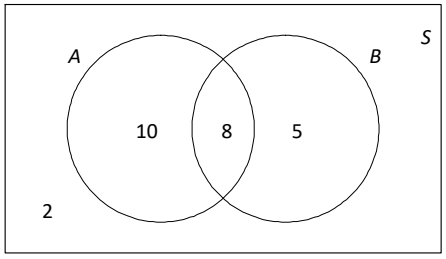
Instead of writing the elements on the diagram we could write the total number of elements (or the probability) in each region.

Question

In a group of 25 people, 18 have a mortgage, 13 own some shares and 2 people have neither a mortgage nor any shares. Determine how many people have both.

Solution

Adding up the number of people who have a mortgage and own shares, we should get 23, since there are two people who have neither. The actual figure is 31, but this includes the people who have a mortgage and own shares twice. Therefore there must be 8 people in the intersection.



There are 8 people with both a mortgage and some shares.

4.3 Probabilities

If each of the elements in the sample space are equally likely, then we can define the probability of event *A* as:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } S}$$

Question

Determine the probability of rolling an even number on an ordinary die.

Solution

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Defining 'throwing an even number' as event *A*, we have $A = \{2, 4, 6\}$.

So the probability of throwing an even number is given by:

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

4.4 Basic probability axioms

The three basic probability axioms can be summarised as follows:

1 $P(S) = 1$

It follows that for event *A* from sample space *S* that $P(A') = 1 - P(A)$.

2 $P(A) \geq 0$ for all $A \subset S$

Rules 1 and 2 together are telling us that probabilities lie between 0 (impossible) and 1 (certain).

3 $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$

If two events cannot occur simultaneously, *ie* they are mutually exclusive, the probability of the event defined by their union is equal to the sum of the probabilities of the two events. This property is known as *additivity*.

4.5 The addition rule

Where two sets are not necessarily mutually exclusive, the general case of the addition rule is:

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

The addition rule can be extended to three events A, B and C:

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

Question

Students at a performing arts college can choose to study one or more classes of acting, dance or singing. The probability that a student is studying acting is 0.5, dance 0.65, singing 0.55, acting or dancing 0.8, acting and singing 0.25, dancing and singing 0.25. Calculate the probability that a student studies all three classes.

Solution

We are given:

$$P(A)=0.5 \quad P(D)=0.65 \quad P(S)=0.55$$

$$P(A\cup D)=0.8 \quad P(A\cap S)=0.25 \quad P(D\cap S)=0.25$$

Since students must study at least one of these classes, we also have:

$$P(A\cup D\cup S)=1$$

We require $P(A\cap D)$:

$$P(A\cup D)=P(A)+P(D)-P(A\cap D) \quad \Rightarrow \quad 0.8=0.5+0.65-P(A\cap D) \quad \Rightarrow \quad P(A\cap D)=0.35$$

Hence, using the addition rule:

$$1=0.5+0.65+0.55-0.35-0.25-0.25+P(A\cap D\cap S) \quad \Rightarrow \quad P(A\cap D\cap S)=0.15$$

4.6 Conditional probabilities

Consider two events, A and B. The probability that event A occurs, given that event B occurs is known as a conditional probability and is written as:

$$P(A\,|\,B)$$

It is calculated using the formula:

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

Question

A card is picked from an ordinary pack of 52 playing cards and events are defined as follows:

$A=\{\text{pick a spade}\}$ $\quad\quad$ $B=\{\text{pick an 8}\}$

Calculate $P(A\,|\,B)$.

Solution

$A \cap B = \{8 \text{ of spades}\} \Rightarrow P(A \cap B) = \frac{1}{52}$

$B = \{\text{pick an 8}\} \Rightarrow P(B) = \frac{4}{52}$

$\Rightarrow P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$

Rearranging the formula for conditional probabilities, we obtain:

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

This is known as the *multiplication rule*. $A \cap B$ can be interpreted as A and B, so we can use this rule to calculate the probability of events A and B both happening.

4.7 Independent events

Events A and B are said to be *independent* if whether or not event B has occurred gives us no information on whether event A will occur. This can be expressed algebraically as follows:

$$P(A) = P(A \mid B) = P(A \mid B')$$

If A and B are independent then:

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

Question

Calculate the probability of rolling a 5 on both dice when two dice are thrown.

Solution

$A = \{\text{roll a 5 on the 1st die}\} \Rightarrow P(A) = \frac{1}{6}$ $B = \{\text{roll a 5 on the 2nd die}\} \Rightarrow P(B) = \frac{1}{6}$

Since these events are independent:

$$P(A \cap B) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

4.8 Tree diagrams

A tree diagram is a convenient representation of probabilities.

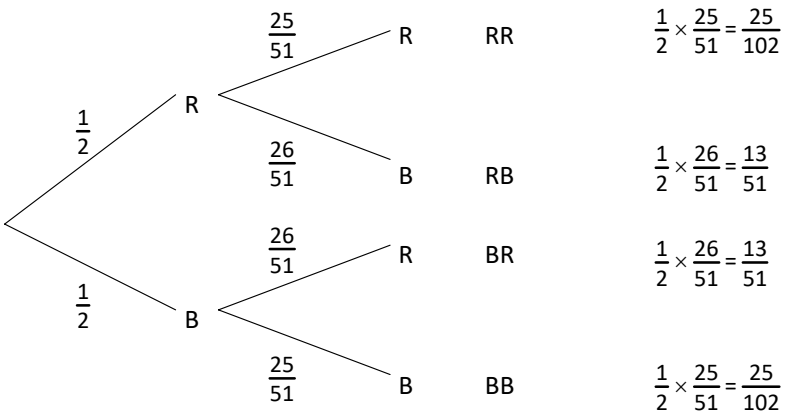
Question

A card is picked from an ordinary pack of 52 playing cards, *without replacement*, and then another one is picked. Calculate the probability of picking:

(i) two red cards

(ii) one of each colour.

Solution



(i) $P(RR) = \frac{1}{2} \times \frac{25}{51} = \frac{25}{102}$

(ii) $P(1\ R \text{ and } 1\ B) = P(RB) + P(BR) = \left(\frac{1}{2} \times \frac{26}{51}\right) + \left(\frac{1}{2} \times \frac{26}{51}\right) = \frac{13}{51} + \frac{13}{51} = \frac{26}{51}$

Tree diagrams can be used to calculate conditional probabilities.

Question

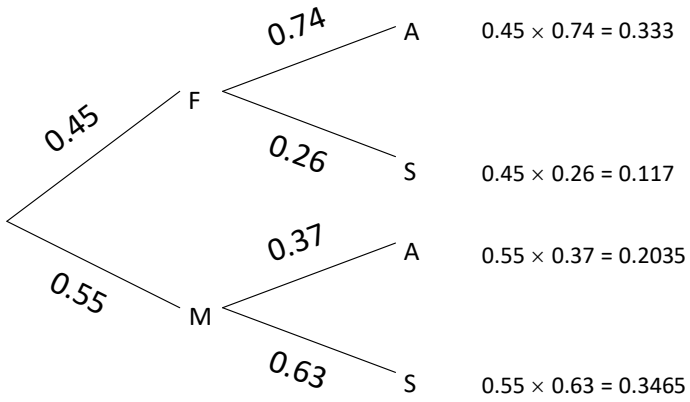
In a restaurant, 45% of the customers are female. 74% of females choose from the à la carte menu, whilst only 37% of males do. The rest choose from the set menu. Calculate the probability that:

(i) a customer orders from the set menu

(ii) a customer ordering from the à la carte menu is female.

Solution

If F is the event 'is female', M is the event 'is male', A is the event 'chooses from à la carte', and S is the event 'chooses from the set menu', then the tree diagram is as follows:



(i) $P(\text{S}) = 0.117 + 0.3465 = 0.4635$

(ii) This is a conditional probability as we are *given* that the customer *has* chosen from the à la carte menu:

$$P(F \mid A) = \frac{P(F \cap A)}{P(A)} = \frac{0.333}{0.333 + 0.2035} = 0.621$$

4.9 Law of total probability

Suppose that the set space, S, is divided into a partition of n mutually exclusive events, E_i where $i = 1, 2, 3, \ldots, n$, so that:

$$E_i \cap E_j = \varnothing$$

$$E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n = S$$

Then for any $A \subset S$:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3) \cup \ldots \cup (A \cap E_n)$$

and:

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + \cdots + P(A \cap E_n) = \sum_{j=1}^{n} P(A \cap E_j)$$

This result is known as the law of total probability.

4.10 Bayes' Theorem

Bayes' Theorem states that:

$$P(E_i \mid A) = \frac{P(E_i)P(A \mid E_i)}{\sum_{j=1}^{n} P(E_j)P(A \mid E_j)} \qquad i = 1, 2, 3, \ldots, n$$

This formula is given on page 5 of the *Tables* that are allowed in the actuarial exams. It allows us to calculate $P(E_i \mid A)$ given $P(A \mid E_i)$ *ie* turn around conditional probabilities.

Question

The punctuality of trains has been investigated by studying a number of train journeys. In the sample, 60% of trains had a destination of Manchester, 20% Birmingham and 20% Edinburgh. The probabilities of a train arriving late in Manchester, Edinburgh or Birmingham are 30%, 20% and 25% respectively.

If a late train is picked at random from the group under consideration, calculate the probability that it terminated in Manchester.

Solution

We want P(Manchester/Late).

If M is the event 'a train chosen at random terminated in Manchester' (and E and B have corresponding definitions), and L is the event 'a train chosen at random runs late', then:

$$P(M \mid L) = \frac{P(M)P(L \mid M)}{P(M)P(L \mid M) + P(E)P(L \mid E) + P(B)P(L \mid B)}$$

$$= \frac{0.6 \times 0.3}{(0.6 \times 0.3) + (0.2 \times 0.2) + (0.2 \times 0.25)} = 66.7\%$$

5 Random variables

5.1 Discrete random variables

A random variable X is a discrete random variable if the set of all possible values for x (the range) is a finite set or a countably infinite set. Each value of x has an associated probability.

The function $P(X = x)$ is known as the probability function of X. The requirements for a function to qualify as the probability function of a discrete random variable are:

$$P(X = x) \geq 0 \qquad \text{for all } x \text{ within the range of } X$$

$$\sum_x P(X = x) = 1$$

The cumulative distribution or cumulative distribution function (CDF) of X gives the probability that X takes a value that does not exceed x:

$$F_X(x) = P(X \leq x)$$

The mean, or expected value, of a discrete random variable is given by:

$$E[X] = \sum_x xP(X = x)$$

The expected value of a function of a discrete random variable is given by:

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

The variance of a discrete random variable is given by:

$$\text{var}[X] = E[X^2] - \left(E[X]\right)^2 = \sum_x x^2 P(X = x) - \left(\sum_x xP(X = x) \right)^2$$

Question

A discrete random variable has probability function:

x	0.1	0.15	0.2
$P(X = x)$	0.375	0.25	0.375

Determine:

(i) $P(X = 0.15)$

(ii) $F(0.15)$

(iii) $E[X]$

(iv) $\text{var}[X]$

(v) the standard deviation of X.

Solution

(i) From the table, $P(X = 0.15) = 0.25$.

(ii) By definition, $F(0.15) = P(X \leq 0.15)$:

$$F(0.15) = P(X \leq 0.15) = 0.375 + 0.25 = 0.625$$

(iii) The expected value is determined by using summation:

$$E[X] = 0.1 \times 0.375 + 0.15 \times 0.25 + 0.2 \times 0.375 = 0.15$$

By symmetry, this could have been written down immediately.

(iv) The variance is determined by using summation:

$$E[X^2] = 0.1^2 \times 0.375 + 0.15^2 \times 0.25 + 0.2^2 \times 0.375 = 0.024375$$

$$\text{var}[X] = 0.024375 - 0.15^2 = 0.001875$$

(v) The standard deviation is the square root of the variance:

$$\sqrt{0.001875} = 0.0433$$

5.2 Continuous random variables

The range of a continuous random variable X is an interval (or a collection of intervals) on the real line.

The function $f_X(x)$ for a continuous random variable is known as the probability density function. Probabilities can be determined by integrating the PDF:

$$P(a < X < b) = \int_a^b f_X(x)\, dx$$

The conditions for a function to be a valid PDF are as follows:

$$f_X(x) \geq 0 \quad -\infty \leq x \leq \infty$$

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

The cumulative distribution or cumulative distribution function (CDF) of X gives the probability that X takes a value that does not exceed x:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt$$

The mean, or expected value, of a continuous random variable is given by:

$$E[X] = \int_{-\infty}^{\infty} x f_x(x)\, dx$$

The expected value of a function of a continuous random variable is given by:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_x(x)\, dx$$

The variance of a continuous random variable is given by:

$$\text{var}[X] = E[X^2] - \big(E[X]\big)^2 = \int_{-\infty}^{\infty} x^2 f_x(x)\, dx - \left(\int_{-\infty}^{\infty} x f_x(x)\, dx \right)^2$$

Question

The continuous random variable W has the PDF $f_W(w) = 12w^2(1-w)$, for $0 < w < 1$. Determine:

(i) $P(W < \frac{1}{2})$

(ii) an expression for $F_W(w)$

(iii) $E[W]$

(iv) $\text{var}[W]$.

Solution

(i) The probability is found by integrating the PDF over the relevant range of values, which here is 0 to $\frac{1}{2}$.

$$P(W < \tfrac{1}{2}) = \int_{0}^{\frac{1}{2}} 12w^2(1-w)\, dw = \Big[4w^3 - 3w^4 \Big]_0^{\frac{1}{2}} = 4 \times \frac{1}{8} - 3 \times \frac{1}{16} = \frac{5}{16}$$

(ii) The CDF is also found by integration, this time with an upper limit of w:

$$F_W(w) = \int_0^w 12t^2(1-t)\,dt = \left[4t^3 - 3t^4\right]_0^w = w^3(4-3w)$$

(iii) The expected value is determined by integrating:

$$E[W] = \int_0^1 w f_W(w)\,dw = \int_0^1 12w^3(1-w)\,dw = \int_0^1 12w^3 - 12w^4\,dw$$

$$= \left[3w^4 - \frac{12}{5}w^5\right]_0^1 = 0.6$$

(iv) The variance is determined by first calculating $E[W^2]$:

$$E[W^2] = \int_0^1 12w^4(1-w)\,dw = \int_0^1 12w^4 - 12w^5\,dw = \left[\frac{12}{5}w^5 - 2w^6\right]_0^1 = 0.4$$

The variance is $0.4 - 0.6^2 = 0.04$.

5.3 Linear functions of a random variable

Let $Y = aX + b$, where a and b are constants. Then:

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

$$\text{var}[Y] = a^2\,\text{var}[X]$$

Question

If a random variable X has a mean of 3 and standard deviation of 2, calculate:

(i) $E[2X - 4]$

(ii) $\text{var}[3X + 2]$

(iii) $\text{var}[3 - 4X]$

(iv) $E\left[\frac{3X + 2}{4}\right]$

Solution

(i) $E[2X - 4] = 2E[X] - 4 = 2$

(ii) $\text{var}(3X + 2) = 9\,\text{var}[X] = 36$

(iii) $\text{var}(3-4X) = 16\,\text{var}[X] = 64$

(iv) $E\left[\frac{3X+2}{4}\right] = \frac{3}{4}E[X] + \frac{1}{2} = 2.75$

5.4 Moments and other quantities

Moments

Earlier in this chapter we considered the moments and quartiles of a set of data. We now define the moments and quartiles of a random variable. $E[(X-c)^k]$ is the k th moment or k th-order moment of X about c. Moments about the mean are called central moments. A non-central moment refers to a moment about zero.

The third central moment or skewness, μ_3, is given by:

$$\mu_3 = E[(X-\mu)^3] = E[X^3] - 3\mu E[X^2] + 2\mu^3$$

where $\mu = E[X]$.

The coefficient of skewness is given by $\frac{\mu_3}{\sigma^3}$, where $\sigma^2 = \text{var}[X]$. It is a dimensionless measure.

Median and quartiles

The median, m, is defined for any random variable X by:

$$P(X < m) \le 0.5 \le P(X \le m)$$

In particular, if X is continuous the median, m, is defined as the solution of:

$$P(X < m) = \int_{-\infty}^{m} f_X(x)\,dx = 0.5$$

Quartiles can also be defined in an obvious way. The lower and the upper quartiles of a random variable X are the values q_1 and q_3 such that:

$$P(X < q_1) = 0.25 \qquad P(X < q_3) = 0.75$$

So the interquartile range, IQR, is given by:

$$IQR = q_3 - q_1$$

Question

The random variable X has PDF:

$$f_X(x) = 2e^{-2x} \qquad x > 0$$

Calculate the:

(i) median

(ii) interquartile range.

Solution

(i) The median, m, is the value such that:

$$P(X < m) = \int_0^m 2e^{-2x}\, dx = 0.5$$

Integrating gives:

$$\left[-e^{-2x}\right]_0^m = 1 - e^{-2m} = 0.5 \quad \Rightarrow \quad m = -\tfrac{1}{2}\ln 0.5 = 0.3466$$

(ii) The lower and upper quartiles are the values q_1 and q_3 such that:

$$P(X < q_1) = \int_0^{q_1} 2e^{-2x}\, dx = \left[-e^{-2x}\right]_0^{q_1} = 1 - e^{-2q_1} = 0.25$$

$$P(X < q_3) = \int_0^{q_3} 2e^{-2x}\, dx = \left[-e^{-2x}\right]_0^{q_3} = 1 - e^{-2q_3} = 0.75$$

Hence:

$$q_1 = -\tfrac{1}{2}\ln 0.75 = 0.1438$$

$$q_3 = -\tfrac{1}{2}\ln 0.25 = 0.6931$$

$$IQR = 0.6931 - 0.1438 = 0.5493$$

5.5 Functions of a continuous random variable

Suppose that $Y = u(X)$ is a function of the random variable X. We can derive the cumulative distribution function of Y from the cumulative distribution function of X.

If $y = u(x)$ is such that a unique inverse $x = w(y) = u^{-1}(y)$ exists and $u(x)$ is an increasing function, then:

$$F(y) = P(Y < y) = P[u(X) < y] = P[X < w(y)] = F_x[w(y)]$$

Having determined $F(y)$, we can obtain the PDF $f(y)$, if required, by differentiation.

In the case that $u(x)$ is decreasing:

$$F(y) = P(Y < y) = P[u(X) < y] = P[X > w(y)] = 1 - F_x[w(y)]$$

In both cases we rearrange the inequality $P[u(X) < y]$ so that X is on the left-hand side.

Question

(i) Determine the cumulative distribution function for the random variable X with PDF:

$$f(x) = 2\beta x e^{-\beta x^2}, \qquad x > 0$$

where β is a positive constant.

(ii) Hence, derive the PDF of $Y = X^2$.

Solution

(i) The cumulative distribution function is:

$$F(x) = \int_0^x 2\beta t e^{-\beta t^2} dt = \left[-e^{-\beta t^2} \right]_0^x = 1 - e^{-\beta x^2}$$

(ii) So the distribution function of Y is:

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = P(X \le \sqrt{y})$$

since X can only take positive values. This is equal to $F_X(\sqrt{y}) = 1 - e^{-\beta y}$.

Therefore:

$$f_Y(y) = F_Y'(y) = \beta e^{-\beta y}$$

Practice questions start on the next page so that you can separate the questions and solutions.

Chapter 0 Practice Questions

There are some questions below for further practice.

0.1 A life insurance company examines the ages last birthday of the last 100 policyholders to take out endowment assurance policies. The results are shown below:

Age (years)	0 – 14	15 – 19	20 – 24	25 – 34	35 – 54	55 – 79
Frequency	9	28	21	16	14	12

(i) Draw a histogram of policyholders ages and use it to comment on the shape of the distribution.

(ii) Estimate the mean age for these policyholders.

(iii) Estimate the median age for these policyholders.

0.2 A small shop has a 'pick-n-mix' counter where customers may choose wine gums, jelly beans or cola bottles. The probability that a customer purchases cola bottles is 0.45, jelly beans and wine gums 0.19, cola bottles and jelly beans 0.15, cola bottles and wine gums 0.25, cola bottles **or** jelly beans 0.6, cola bottles **or** wine gums 0.84, and at least one of them 0.9.

Calculate the probability that a customer purchases:

(i) jelly beans

(ii) wine gums

(iii) all three

(iv) none of them.

0.3 A discrete random variable X has probability function given by:

x	0	1	2
$P(X = x)$	0.3	0.5	0.2

Calculate:

(i) $E(X)$

(ii) $\text{var}(X)$

(iii) the coefficient of skewness.

0.4 In a claims department of a motor insurance company, a sample of 12 claims had a mean of £845 and a standard deviation of £208.

It was then discovered that two mistakes were made; a claim of £526 was classified wrongly and is removed from the set of 12 claims, and a new claim of £1,034 is added to make the sample back up to 12.

Calculate the sample mean and standard deviation of the modified sample of 12 claims.

0.5 An actuarial recruitment company places adverts in three publications with probabilities of 0.2, 0.3 and 0.5 respectively.

The probability that the recruitment company gets an enquiry from an advert in the first publication is 0.001. The probabilities for the other two publications are 0.002 and 0.004 respectively.

The company has just received an enquiry. Calculate the probability that it came from an advert in the first publication.

0.6 A continuous random variable Y has PDF:

$$f(y)=\begin{cases} y(y-1)(y-2)+0.4 & 0\le y\le 2 \\ c & 2<y\le 4 \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant. Determine:

(i) the value of c

(ii) $E[Y]$

(iii) the standard deviation of Y.

0.7 The random variable U has PDF:

$$f(u)=1, \quad 0<u<1$$

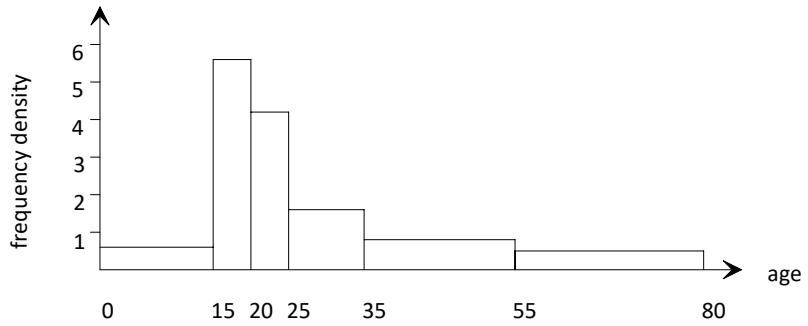
Determine the PDF of U^2.

Chapter 0 Solutions

0.1 (i) ***Histogram***

To draw a histogram we calculate the heights using $\frac{\text{frequency}}{\text{width}}$. The width of the $0-14$ group is 15, as someone aged 14 could be aged just under 15 years.

The heights are $\frac{9}{15}=0.6,\ \frac{28}{5}=5.6,\ \frac{21}{5}=4.2,\ \frac{16}{10}=1.6,\ \frac{14}{20}=0.7,\ \frac{12}{25}=0.48$.



The distribution appears to be positively skewed.

(ii) ***Mean***

Using $\bar{x}=\frac{\sum fx}{\sum f}$ where x is the midpoint of the groups gives:

$$\bar{x}=\frac{9\times 7.5+28\times 17.5+21\times 22.5+16\times 30+14\times 45+12\times 67.5}{9+28+21+16+14+12}=\frac{2{,}950}{100}=29.5 \text{ years}$$

(iii) ***Median***

The position of the median is given by:

$$\frac{n}{2}=\frac{100}{2}=50$$

The cumulative frequencies are:

Age (years)	0 – 14	15 – 19	20 – 24	25 – 34	35 – 54	55 – 79
Cum Freq	9	37	58	74	88	100

So the median lies in the 20-24 group. Interpolation gives a median of :

$$20+\frac{13}{21}\times 5=23.1 \quad \text{years}$$

0.2 Let C be the event 'cola bottles are purchased', W the event 'wine gums are purchased' and J the event 'jelly beans are purchased'.

We are given:

$$P(C)=0.45 \qquad P(J\cap W)=0.19 \qquad P(C\cap J)=0.15 \qquad P(C\cap W)=0.25$$

$$P(C\cup J)=0.6 \qquad P(C\cup W)=0.84 \qquad P(C\cup J\cup W)=0.9$$

(i) ***Probability purchase jelly beans***

We require $P(J)$. Using the addition rule:

$$P(C\cup J)=P(C)+P(J)-P(C\cap J) \;\Rightarrow\; 0.6=0.45+P(J)-0.15 \;\Rightarrow\; P(J)=0.3$$

(ii) ***Probability purchase wine gums***

We require $P(W)$. Using the addition rule again:

$$P(C\cup W)=P(C)+P(W)-P(C\cap W) \;\Rightarrow\; 0.84=0.45+P(W)-0.25 \;\Rightarrow\; P(W)=0.64$$

(iii) ***Probability purchase all three***

We require $P(C\cap J\cap W)$. Using the addition rule:

$$P(C\cup J\cup W)=P(C)+P(J)+P(W)-P(C\cap J)-P(C\cap W)-P(J\cap W)+P(C\cap J\cap W)$$

$$\Rightarrow\; 0.9=0.45+0.3+0.64-0.15-0.25-0.19+P(C\cap J\cap W) \;\Rightarrow\; P(C\cap J\cap W)=0.1$$

(iv) ***Probability purchase none of them***

We require $P(C'\cup J'\cup W')$. The easiest way to get this is to realise that the probability of buying none is the complement of choosing at least one:

$$P(C'\cup J'\cup W')=1-P(C\cup J\cup W)=1-0.9=0.1$$

0.3 (i) ***E(X)***

$$E(X)=\mu=\sum_x xP(X=x)=(0\times 0.3)+(1\times 0.5)+(2\times 0.2)=0.9$$

(ii) ***Variance***

$$E(X^2)=\sum_x x^2P(X=x)=(0^2\times 0.3)+(1^2\times 0.5)+(2^2\times 0.2)=1.3$$

$$\Rightarrow\; \text{var}(X)=E(X^2)-\big(E(X)\big)^2=1.3-0.9^2=0.49$$

(iii) ***Coefficient of skewness***

The skewness is given by:

$$skew(X)=E\Big[(X-\mu)^3\Big]=(0-0.9)^3\times 0.3+(1-0.9)^3\times 0.5+(2-0.9)^3\times 0.2=0.048$$

Alternatively, we could use $skew(X) = E(X^3) - 3\mu E(X^2) + 2\mu^3$ *where* $E(X^3) = 2.1$.

The coefficient of skewness is given by:

$$\frac{skew(X)}{\left[\text{var}(X)\right]^{1.5}} = \frac{0.048}{0.49^{1.5}} = 0.140$$

0.4 We are given that $n = 12, \bar{x} = 845$ and $s = 208$.

$$\bar{x} = \frac{\sum x}{n} \quad \Rightarrow \quad \sum x = 12 \times 845 = 10{,}140$$

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right) \quad \Rightarrow \quad \sum x^2 = 11 \times 208^2 + 12 \times 845^2 = 9{,}044{,}204$$

Now subtracting the £526 claim and adding the £1,034 claim gives:

$$\sum x = 10{,}140 - 526 + 1034 = 10{,}648$$

$$\sum x^2 = 9{,}044{,}204 - 526^2 + 1{,}034^2 = 9{,}836{,}684$$

So the new results are:

$$\bar{x} = \frac{10{,}648}{12} = £887.33$$

$$s^2 = \frac{1}{11}\left(9{,}836{,}684 - 12 \times 887.33^2\right) = 35{,}305.33 \quad \Rightarrow \quad s = \sqrt{35{,}305.33} = £187.90$$

0.5 Let A be the event 'an advert is placed in publication A', B be the event 'an advert is placed in publication B', C be the event 'an advert is placed in publication C' and E be the event 'an enquiry is received'.

We have:

$$P(A) = 0.2 \qquad P(B) = 0.3 \qquad P(C) = 0.5$$

$$P(E \mid A) = 0.001 \quad P(E \mid B) = 0.002 \quad P(E \mid C) = 0.004$$

We want $P(A \mid E)$:

$$P(A \mid E) = \frac{P(E \mid A)P(A)}{P(E)} = \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid B)P(B) + P(E \mid C)P(C)}$$

Putting in the values from the question, we get:

$$P(A \mid E) = \frac{0.001 \times 0.2}{0.001 \times 0.2 + 0.002 \times 0.3 + 0.004 \times 0.5} = 0.0714$$

0.6 (i) ***Determine c***

To determine the value of c we need to integrate over the whole range of y and set the value of the integral equal to 1. Since the definition is different for different ranges of y, we need to carry out separate integrations:

$$\int_0^2 \big(y(y-1)(y-2)+0.4\big)\,dy + \int_2^4 c\,dy = 1 \quad \Rightarrow \quad \int_0^2 \Big(y^3-3y^2+2y+0.4\Big)\,dy + \int_2^4 c\,dy = 1$$

$$\Rightarrow \quad \left[\frac{y^4}{4}-y^3+y^2+0.4y\right]_0^2 + \big[cy\big]_2^4 = 1 \quad \Rightarrow \quad (4-8+4+0.8)+(4c-2c)=1 \quad \Rightarrow \quad c=0.1$$

(ii) ***Expectation***

To calculate $E[Y]$ we need to multiply the density function by y and then integrate:

$$E[Y] = \int_0^2 (y^4-3y^3+2y^2+0.4y)\,dy + \int_2^4 0.1y\,dy = \left[\frac{y^5}{5}-\frac{3y^4}{4}+\frac{2y^3}{3}+0.2y^2\right]_0^2 + \left[\frac{0.1y^2}{2}\right]_2^4$$

$$= \left(\frac{32}{5}-12+\frac{16}{3}+0.8\right)+\big(0.8-0.2\big) = 1.1333$$

(iii) ***Standard deviation***

First we need $E[Y^2]$:

$$E[Y^2] = \int_0^2 (y^5-3y^4+2y^3+0.4y^2)\,dy + \int_2^4 0.1y^2\,dy = \left[\frac{y^6}{6}-\frac{3y^5}{5}+\frac{2y^4}{4}+\frac{0.4}{3}y^3\right]_0^2 + \left[\frac{0.1y^3}{3}\right]_2^4$$

$$= \left(\frac{64}{6}-\frac{96}{5}+8+\frac{3.2}{3}\right)+\left(\frac{6.4}{3}-\frac{0.8}{3}\right) = 2.4$$

So the standard deviation is given by:

$$\sqrt{\text{var}(X)} = \sqrt{E(X^2)-E^2(X)} = \sqrt{2.4-1.1333^2} = \sqrt{1.1156} = 1.056$$

0.7 Let $V = U^2$. Since U only takes positive values, the distribution function of V is:

$$F_V(v) = P(V \le v) = P(U^2 \le v) = P(U \le \sqrt{v})$$

We can calculate this using integration:

$$P(U < \sqrt{v}) = \int_0^{\sqrt{v}} 1\,du = \big[u\big]_0^{\sqrt{v}} = \sqrt{v}$$

So the PDF of V is:

$$f_V(v) = F_V'(v) = \frac{d}{dv} v^{½} = ½v^{-½}$$

Since U can take values in the range $0 < U < 1$, V can take values in the range $0 < V < 1$.