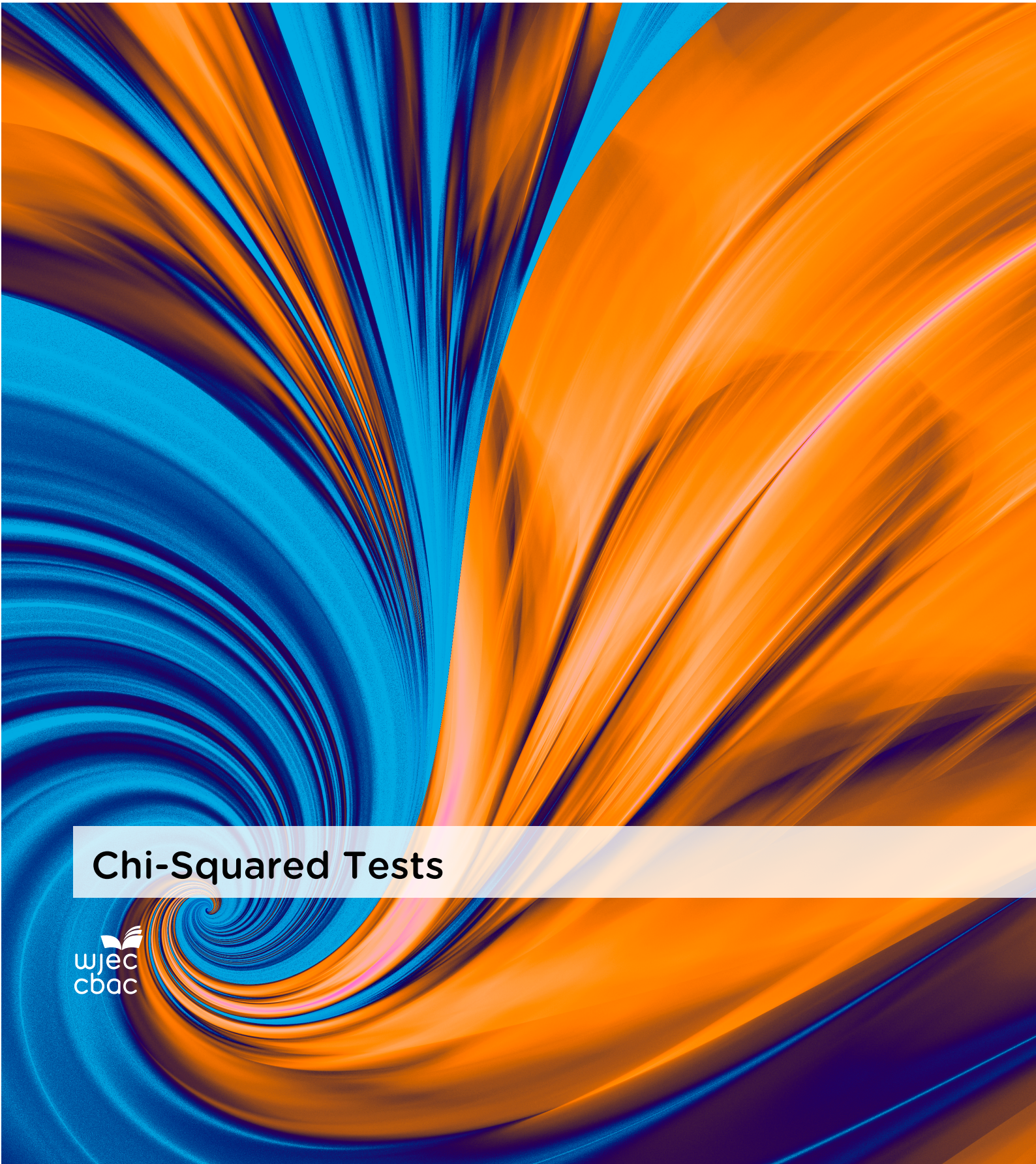
GCE

FURTHER MATHEMATICS
AS UNIT 2: FURTHER STATISTICS A



Chi-Squared Tests

Further Mathematics AS Unit 2: Further Statistics A

Chi-Squared Tests

You have previously met the binomial, Poisson and discrete uniform distributions. These are examples of discrete probability distributions. This unit also introduces the exponential distribution, which by contrast is a continuous probability distribution.

For this section, we introduce the chi-square distribution. This is a continuous probability distribution and is used in several statistical hypothesis tests. Any statistical test whose test statistic follows, or approximately follows, the chi-squared distribution is referred to as a chi-squared test.

In this chapter, two types of chi-squared tests shall be considered. The first of these is the chi-squared goodness of fit test, which compares sample data with expected frequencies from a known distribution. Samples that deviate too far from the expected frequencies will provide evidence to suggest that the hypothesised distribution is unlikely to be a suitable fit for the data.

The chi-square test for association in a contingency table shall also be considered. The purpose of this test is to identify whether there is an association between two categorical variables. To do this, a contingency table is used which shows all possible combinations of outcomes from both variables. Expected frequencies for each pair are computed by assuming there is no association between the variables. Sample frequencies that deviate too far from these expected frequencies provide evidence of an association between the variables.

Specification Content

- Understand and use the chi-squared distribution
- Conduct goodness of fit test using $\sum \frac{(O-E)^2}{E}$, or equivalent form, as an approximate χ^2 statistic (for use with categorical data)
  - for use with binomial, discrete uniform and Poisson distributions, for known parameters only
- Use χ^2 test to test for association in a contingency table and interpret results
  - to include pooling
  - not including Yates continuity correction.

Properties of Continuous Random Variables

The chi-squared distribution is an example of a continuous random variable. The rules for calculating probabilities using continuous random variables are different to those for discrete random variables.

Let X be a continuous random variable and let a, b, c and d be constants. Then:

- $P(X = a) = 0$
- $P(X \geq b) = 1 - P(X \leq b)$
- $P(c \leq X \leq d) = P(X \leq d) - P(X \leq c)$.

Contrast this with the rules for discrete distributions. Let X be a discrete random variable and let a, b, c and d be constants. Then:

- $P(X = a) = P(X \leq a) - P(X \leq a - 1)$
- $P(X \geq b) = 1 - P(X \leq b - 1)$
- $P(c \leq X \leq d) = P(X \leq d) - P(X \leq c - 1)$.

Further Mathematics AS Unit 2: Further Statistics A

The Chi-Squared Distribution

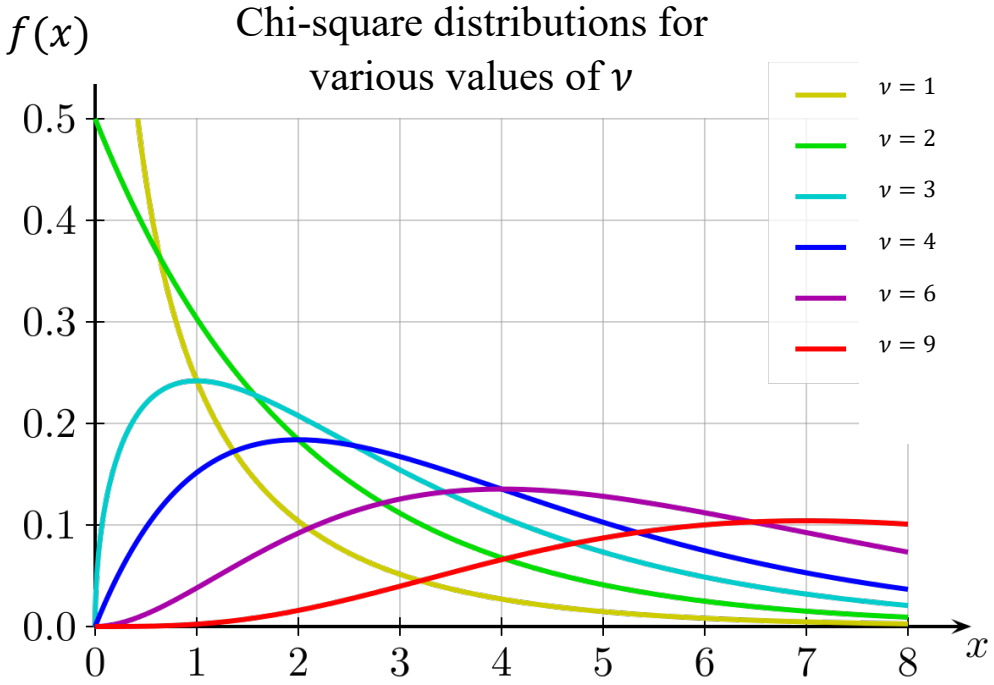
The chi-squared distribution depends on a quantity ν called the degrees of freedom. We shall write $X \sim \chi^2(\nu)$ to show that the random variable X follows the chi-square distribution with ν degrees of freedom.

The probability density function of the chi-square distribution is given by:

$$f(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}.$$

Fortunately, we don't have to use this PDF in practice, as Table 6 of the Elementary Statistical Tables gives percentage points of the chi-square distribution.

The shape of the distribution changes depending on the value of ν. The chi-square distribution is not symmetrical; however, it becomes more symmetrical as the value of ν increases. If $X \sim \chi^2(\nu)$, then it can be shown that $E(X) = \nu$ and $\mathrm{Var}(X) = 2\nu$. Plots of various chi-square distributions are shown below.



The percentage points table for the chi-square distribution gives critical values for both small and large probabilities. The text with the table states:

"The table gives values of x satisfying $P(X \leq x) = p$ where X is a χ^2 random variable with ν degrees of freedom".

We shall write $x_{p,\nu}$ to denote the observed value from X, the chi-square distribution with ν degrees of freedom, such that $P(X \leq x_{p,\nu}) = p$.

The first column of the table gives values of ν, the degrees of the freedom, ranging from 1 to 40 and further multiples of five up to 100. The first row gives values of the probability p. The first five probabilities are all less than or equal to 0.1, whilst the last five probabilities are all greater than or equal to 0.9. The vertical line separates the small values of p from the large values of p.

Further Mathematics AS Unit 2: Further Statistics A

Worked Example 1
Find the following values and write them as probability statements of the form $P(X \leq x) = p$, stating clearly the distribution of X.

1. $x_{0.9,10}$
2. $x_{0.005,60}$
3. $x_{0.975,12}$
4. $x_{0.05,4}$
5. $x_{0.995,100}$

Solution:

1. $x_{0.9,10} = 15.987$, $P(X \leq 15.987) = 0.9$ where $X \sim \chi^2(10)$.
2. $x_{0.005,60} = 35.534$, $P(X \leq 35.534) = 0.005$ where $X \sim \chi^2(60)$.
3. $x_{0.975,12} = 23.337$, $P(X \leq 23.337) = 0.975$ where $X \sim \chi^2(12)$.
4. $x_{0.05,4} = 0.711$, $P(X \leq 0.711) = 0.05$ where $X \sim \chi^2(4)$.
5. $x_{0.995,100} = 140.169$, $P(X \leq 140.169) = 0.995$ where $X \sim \chi^2(100)$.

Whilst the percentage points tables are useful for calculating critical values, they are less helpful when calculating probabilities. Suppose we wish to calculate $P(X \leq 0.8)$ where $X \sim \chi^2(4)$. Since 1 is not listed in the tables, it is not possible to find the exact probability. Instead, based on the values given in the table, we can compute an interval for the probability.

Let $X \sim \chi^2(4)$. From the tables, we know that $P(X \leq 0.711) = 0.05$ and $P(X \leq 1.064) = 0.1$. Therefore, as 0.8 lies between 0.711 and 1.064, we know that $P(X \leq 0.8)$ must lie somewhere between 0.05 and 0.1. Therefore, we can say that:

$$0.05 < P(X \leq 0.8) < 0.1.$$

Suppose we now wish to calculate $P(X \leq 4)$. From tables, we know that $P(X \leq 3.841) = 0.95$ and $P(X \leq 5.024) = 0.975$. Therefore, since 4 lies between 3.841 and 5.024, we know that:

$$0.95 < P(X \leq 4) < 0.975.$$

Worked Example 2
Let $X \sim \chi^2(4)$. Find an interval for $P(X \geq 8)$. Give your answer as an inequality in the form $a < P(X \geq 8) < b$, where a,b are values to be determined.

Solution:
From tables, we observe that $P(X \leq 7.779) = 0.9$ and $P(X \leq 9.488) = 0.95$. Therefore, $P(X \geq 7.779) = 1 - 0.9 = 0.1$, and $P(X \geq 9.488) = 1 - 0.95 = 0.05$.

Since 8 lies between 7.779 and 9.488, we know that $P(X \geq 8)$ must lie somewhere between 0.05 and 0.1. Hence:

$$0.05 < P(X \geq 8) < 0.1.$$

Worked Example 3
Let $X \sim \chi^2(29)$. Find an interval for $P(X \geq 60)$. Give your answer as an inequality in the form $a < P(X \geq 60) < b$, where a, b are values to be determined.

Solution:
From ta bles, we observe that $P(X \leq 52.336) = 0.995$. Therefore, $P(X \geq 52.336) = 0.005$. Since 60 is greater than 52.336, we know that $P(X \geq 60)$ must be smaller than 0.005. Since the tables do not give us any lower bound, we use a lower bound of 0 (since all probabilities lie between 0 and 1). Therefore:

$$0 < P(X \geq 60) < 0.005.$$

Further Mathematics AS Unit 2: Further Statistics A

Goodness of Fit Tests

A chi-square goodness of fit test can be used to assess whether a proposed distribution is a good fit for some observed data. The theory of goodness of fit tests shall be discussed through the following example.

Worked Example 4

Suppose we wish to test whether a dice is fair or not. We throw the dice 120 times and record the number of times each score is obtained. The results are as follows:

Score	1	2	3	4	5	6
Frequency	12	16	15	23	24	30

We wish to investigate whether these results provide evidence that the dice is biased, or if the variation is just something that would be expected to occur naturally. We require a statistical hypothesis test to investigate this.

Our null hypothesis is that the dice is unbiased, therefore each score is equally likely to occur. This means that the scores follow a discrete uniform distribution with probability $p = \frac{1}{6}$. Therefore, the expected number of each score is $120 \times \frac{1}{6} = 20$. We can write the hypotheses as follows:

H_0 – the distribution of scores follows a discrete uniform distribution with probability $p = \frac{1}{6}$.
H_1 – the distribution of scores does not follow a discrete uniform distribution with probability $p = \frac{1}{6}$.

The expected frequencies can be added to the original table. The letter O shall be used to denote an observed value and E to denote an expected value.

Score	1	2	3	4	5	6
Observed Frequency, O	12	16	15	23	24	30
Expected Frequency, E	20	20	20	20	20	20

The total of the observed and expected frequencies is the same number, denoted by N. Therefore $\Sigma O = \Sigma E = N$, where Σ is used to denote 'sum of'. In this example, $N = 120$. The number of classes (i.e. categories) in the table, in this case 6, shall be denoted by n.

We require a method to assess the deviations between the observed and the expected frequencies. To do this, we introduce the test statistic:

$$X = \sum \frac{(O - E)^2}{E}.$$

This means that we sum the values of $\frac{(O-E)^2}{E}$ for each score in the table. The value of $\frac{(O-E)^2}{E}$ is referred to as the chi-square contribution for that score.

The test statistic X follows the chi-squared distribution approximately. Before being able to use the statistical tables to calculate critical values, we need to consider the significance level of the test and the degrees of freedom of the chi-square distribution that X approximately follows.

The significance level should be set before collecting the data. Recall that if a significance level is not stated in an exam question, a 5% level may be assumed. The degrees of freedom correspond to the number of classes in the table that must be filled, minus one degree of freedom for each restriction placed on the frequencies.

In this example, there are six classes (corresponding to each of the six scores). Since the total number of throws is pre-determined, namely 120, the frequency in the final class could be worked out if the values in the other five classes were known. Therefore, the degrees of freedom are $\nu = 6 - 1 = 5$.

The other restrictions that could be placed on the frequencies occur when parameters of the distribution specified in the null hypothesis are unknown. For this unit, values of the parameters will always be known. Therefore, the degrees of freedom will always be calculated as $\nu = n - 1$.

An important assumption when using the chi-square goodness of fit test is that the expected frequency in each class must be five or more. If a class has an expected frequency that is less than five, it should be combined with another class to give a resulting expected frequency of at least five.

Let us revisit the test statistic. By expanding the numerator and simplifying, we obtain:

$$\begin{aligned}\sum \frac{(O-E)^2}{E} &= \sum \frac{O^2 - 2OE + E^2}{E} \\ &= \sum \left(\frac{O^2}{E} - 2O + E\right) \\ &= \sum \frac{O^2}{E} - \sum 2O + \sum E \\ &= \sum \frac{O^2}{E} - 2\sum O + \sum E \\ &= \sum \frac{O^2}{E} - 2N + N \\ &= \sum \frac{O^2}{E} - N.\end{aligned}$$

Therefore, the test statistic X can be calculated using either of the following forms:

$$X = \sum \frac{(O-E)^2}{E} \qquad \text{OR} \qquad X = \sum \frac{O^2}{E} - N.$$

The second form is arguably easier for performing calculations. We shall use both equations to show how the observed value, x, of X for the dice data can be calculated.

Method 1: We calculate the chi-square contribution for each score:

Score	1	2	3	4	5	6
Obs Freq, O	12	16	15	23	24	30
Exp Freq, E	20	20	20	20	20	20
Chi-Square Contribution	$\frac{(12-20)^2}{20}$	$\frac{(16-20)^2}{20}$	$\frac{(15-20)^2}{20}$	$\frac{(23-20)^2}{20}$	$\frac{(24-20)^2}{20}$	$\frac{(30-20)^2}{20}$

Summing the chi-square contributions gives the value of the test statistic:

$$\begin{aligned} x &= \frac{(12-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(30-20)^2}{20} \\ &= \frac{64+16+25+9+16+100}{20} \\ &= \frac{230}{20} \\ &= 11.5. \end{aligned}$$

Method 2:

$$\begin{aligned} x &= \frac{12^2}{20} + \frac{16^2}{20} + \frac{15^2}{20} + \frac{23^2}{20} + \frac{24^2}{20} + \frac{30^2}{20} - 120 \\ &= \frac{2630}{20} - 120 \\ &= \frac{230}{20} \\ &= 11.5. \end{aligned}$$

Chi-square goodness of fit tests are always one-sided in the sense that only large values of the test statistic lead to a rejection of H_0. Therefore, for chi-square goodness of fit tests (and later chi-square tests of association), only upper tail critical values for a one-tail test need to be considered.

From tables, the critical value of the chi-square distribution with 5 degrees of freedom at the 5% significance level is $x_c = x_{0.95,5} = 11.070$. Therefore, the critical region is $x \geq 11.070$.

Since $x \geq x_c$, the test statistic lies in the critical region. Therefore, there is sufficient evidence to reject H_0 and to conclude, at the 5% significance level, that the dice is biased.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(5)$. From tables, $P(X \leq 11.070) = 0.95$ and $P(X \leq 12.833) = 0.975$. Therefore, $P(X \geq 11.070) = 0.05$ and $P(X \geq 12.833) = 0.025$. Thus $0.025 < p < 0.05$. We can be certain that the p-value is less than the significance level, which further supports the rejection of H_0.

Worked Example 5

The number of telephone calls made to a counselling service is thought to be modelled by a Poisson distribution with mean 2.5. Data are collected on the number of calls received during one-hour periods, as shown in the table. Use the data to test, at the 10% significance level, whether a Poisson model is appropriate and find an interval for the p-value.

Number of calls per hour	0	1	2	3	4	5	6
Frequency	6	13	26	14	7	3	1

Solution:

The null and alternative hypotheses are:

H_0 – the number of calls can be modelled by the Poisson distribution with mean 2.5.

H_1 – the number of calls cannot be modelled by the Poisson distribution with mean 2.5.

Under H_0, a Poisson distribution can be used to describe the number of calls received during the one-hour periods. Whilst 6 was the greatest number of calls received per hour in this sample, the Poisson distribution can theoretically take any non-negative integer value (i.e. a value from 0, 1, 2,…). Therefore, we change the final category to 6 or more rather than 6 when calculating the expected frequencies.

Let X denote the number of calls received in a one-hour period. Under H_0, $X \sim \text{Po}(2.5)$. The probabilities and expected frequencies for the various values of X are calculated as follows:

$\boldsymbol{x}$	$\boldsymbol{P(X = x)}$	**Expected Frequency = 70 ×** $\boldsymbol{P(X = x)}$
0	$e^{-2.5} = 0.0821$	5.7459
1	$\frac{e^{-2.5} \times 2.5}{1!} = 0.2052$	14.3649
2	$\frac{e^{-2.5} \times 2.5^2}{2!} = 0.2565$	17.9561
3	$\frac{e^{-2.5} \times 2.5^3}{3!} = 0.2138$	14.9634
4	$\frac{e^{-2.5} \times 2.5^4}{4!} = 0.1336$	9.3521
5	$\frac{e^{-2.5} \times 2.5^5}{5!} = 0.0668$	4.6761
≥ 6	$1 - P(X \leq 5) = 0.0420$	2.9415

Since the expected frequencies for 5 and ≥ 6 are less than 5, we pool these classes together. This results in an expected frequency of 7.6175 which is greater than 5. Re-writing the table, including the observed and expected frequencies along with the chi-square contribution $\frac{(O-E)^2}{E}$ for each class, we have:

Number of calls per hour	0	1	2	3	4	5 or more
Observed Frequency, O	6	13	26	14	7	4
Expected Frequency, E	5.7459	14.3649	17.9561	14.9634	9.3521	7.6176
Chi-Square Contribution	0.0112	0.1297	3.6035	0.0620	0.5916	1.7180

Note that the expected frequencies should not be rounded to the nearest whole number, since expected frequencies do not necessarily need to be integers. A useful check is that $\Sigma O = \Sigma E = 70$. We calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{(O-E)^2}{E} = 0.0112 + \cdots + 1.7180 = 6.1160.$$

Alternatively, we can calculate x without using the chi-square contributions as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{6^2}{5.7459} + \frac{13^2}{14.3649} + \cdots + \frac{4^2}{7.6176} - 70 = 76.1160 - 70 = 6.1160.$$

The test statistic follows a chi-square distribution with $\nu = 6 - 1 = 5$ degrees of freedom. The critical value for a chi-square distribution with 5 degrees of freedom at the 10% significance level is $x_c = x_{0.9,5} = 9.236$. Therefore, the critical region is $x \geq 9.236$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, there is insufficient evidence to reject H_0, and so, at the 10% significance level, we conclude that a Poisson distribution with mean 2.5 can be used to model the number of calls received during a one-hour period to the counselling service.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(5)$. From tables, $P(X \leq 1.610) = 0.1$ and $P(X \leq 9.236) = 0.9$. Therefore, $P(X \geq 1.610) = 0.9$ and $P(X \geq 9.236) = 0.1$, and so we deduce that $0.1 < p < 0.9$. We can be certain that the p-value is greater than the significance level, which further supports there being insufficient evidence to reject H_0.

Worked Example 6
An egg packaging firm has introduced a new box for its eggs. Each box holds six eggs. Unfortunately, the firm finds that the new box tends to mark the eggs. It is believed that 57.5% of eggs are marked using the new boxes. Data on the number of eggs marked in 100 boxes are collected. Test, at the 0.5% level of significance, whether the data can be modelled using a binomial distribution, and find an interval for the p-value.

Number of marked eggs	0	1	2	3	4	5	6
Frequency	3	3	27	29	10	7	21

Solution:
The null and alternative hypotheses are:
H_0 – the number of marked eggs per box can be modelled by the binomial distribution. $B(6,0.575)$.
H_1 – the number of marked eggs per box cannot be modelled by the binomial distribution. $B(6,0.575)$.

Let X denote the number of marked eggs per box. Under H_0, $X \sim \mathrm{B}(6,0.575)$. This distribution can only take values between 0 and 6. The probabilities and expected frequencies for the various values of X are calculated as follows:

x	$P(X = x)$	Expected Frequency = $100 \times P(X = x)$
0	$0.425^6 = 0.0059$	0.5893
1	$\binom{6}{1} \times 0.575 \times 0.425^5 = 0.0478$	4.7837
2	$\binom{6}{2} \times 0.575^2 \times 0.425^4 = 0.1618$	16.1802
3	$\binom{6}{3} \times 0.575^3 \times 0.425^3 = 0.2919$	29.1877
4	$\binom{6}{4} \times 0.575^4 \times 0.425^2 = 0.2962$	29.6170
5	$\binom{6}{5} \times 0.575^5 \times 0.425 = 0.1603$	16.0280
6	$0.575^6 = 0.0361$	3.6142

There are three classes with expected frequency less than 5. We need to pool classes to alleviate this problem. We combine classes for $x = 0$ and $x = 1$ as well as the classes for $x = 5$ and $x = 6$. The new pooled classes have expected frequencies greater than 5. Re-writing the table to include the observed and expected frequencies as well as the values of the chi-square contribution $\frac{(O-E)^2}{E}$ for each class:

Number of marked eggs	0 and 1	2	3	4	5 and 6
Observed Frequency, O	6	27	29	10	28
Expected Frequency, E	5.3730	16.1802	29.1877	29.6170	19.6422
Chi-Square Contribution	0.0732	7.2353	0.0012	12.9934	3.5563

A useful check is $\Sigma O = \Sigma E = 100$. We calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{(O-E)^2}{E} = 0.0732 + \cdots + 3.5563 = 23.8595.$$

Alternatively, we can calculate x without using the chi-square contributions as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{6^2}{5.3730} + \frac{27^2}{16.1802} + \cdots + \frac{28^2}{19.6422} - 100 = 123.8594 - 100 = 23.8594.$$

The test statistic follows a chi-square distribution with $\nu = 5 - 1 = 4$ degrees of freedom. The critical value for a chi-square distribution with 4 degrees of freedom at the 0.5% significance level is $x_c = x_{0.995,4} = 14.860$. Therefore, the critical region is $x \geq 14.860$.

Since $x \geq x_c$, the test statistic lies in the critical region. Therefore, there is sufficient evidence to reject H_0, and so, at the 0.5% significance level, we conclude that a binomial distribution $B(6,0.575)$ is not an appropriate model for the number of marked eggs per box.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(4)$. From tables, $P(X \leq 14.860) = 0.995$ and $P(X \geq 14.860) = 0.005$. Therefore, $0 < p < 0.005$. We can be certain that the p-value is less than the significance level, which further supports the rejection of H_0.

Summary of Approach for the Chi-Square Goodness of Fit Test

1. State the null and alternative hypotheses in terms of the distribution being investigated.
2. Calculate the expected frequency for each class, assuming the distribution specified in the null hypothesis is true.
3. Find the value of the test statistic by either calculating chi-square contributions $\frac{(O-E)^2}{E}$ for each class and then $x = \sum \frac{(O-E)^2}{E}$, or calculating $x = \sum \frac{O^2}{E} - N$.
4. The degrees of freedom are calculated as $\nu = n - 1$, where n is the number of classes.
5. Find the critical region using the upper one-tail critical value from the percentage points of the chi-square distribution table.
6. State the conclusion, in terms of the hypotheses and in context:
    a. If the test statistic does not lie in the critical region, there is insufficient evidence to reject H_0, which suggests the proposed distribution is a suitable fit to the data.
    b. If the test statistic does lie in the critical region, there is sufficient evidence to reject H_0, suggesting that the proposed distribution is not a suitable fit to the data.
7. The p-value is calculated as $P(X \geq x)$, where $X \sim \chi^2(\nu)$. Very often, it is only possible to obtain intervals for the p-value.

Test for Association in Contingency Tables

Before considering the chi-square test for association, we introduce the concept of contingency tables. Suppose we have two categorical variables (e.g. gender and smoking status) and have a sample of 100 students. Each student could be in one of four categories - a male smoker, a male non-smoker, a female smoker, and a female non-smoker.

A contingency table lists the numbers of students that fall into each of these categories. One of these variables is listed in the rows, whilst another is listed in the columns. Each entry in the table is called a cell.

Assuming that the row variable and the column variable are independent, then the expected count of a cell, E, is calculated as follows:

$$\text{Expected Cell Count, } E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total Number of People, } N}.$$

This result holds since, assuming the variables are independent:

$$P(\text{belonging to a row and column}) = P(\text{row}) \times P(\text{column}).$$

Multiplying this by the total number of people, N, gives the expected cell count.

Worked Example 7

Consider a group of 100 students. Of these students, 40 are males and 60 are females. 10 of the males like Marmite, and 40 of the females do not like Marmite. Represent this information in a contingency table and find the expected counts in each cell.

Solution:

There are four categories:

- Males who like Marmite – we know there are 10 of these.
- Males who dislike Marmite – there are 40 males, 10 like Marmite, therefore 30 dislike Marmite.
- Females who like Marmite – there are 60 females, 40 dislike Marmite, 20 like Marmite.
- Females who dislike Marmite – we know there are 40 of these.

The contingency table is as follows:

	Like Marmite	*Dislike Marmite*	*Total*
Male	10	30	40
Female	20	40	60
Total	30	70	100

We can now compute the expected counts for each cell:

- Males who like Marmite - $E = \frac{40\times30}{100} = 12.$
- Males who dislike Marmite - $E = \frac{40\times70}{100} = 28.$
- Females who like Marmite - $E = \frac{60\times30}{100} = 18.$
- Females who dislike Marmite - $E = \frac{60\times70}{100} = 42.$

In the test of association for a contingency table, we test whether there is an association between two variables. In the next example, we explore whether there is an association between gender and Marmite preferences for the data in example 7.

In order to use the chi-square distribution, the degrees of freedom need to be considered. Suppose we have a contingency table with r rows and c columns. In total, there are $r \times c$ cells in the table. There are r row totals and c column totals that are known.

However, since the sum of the row and column totals gives the same value, one of these values can be calculated by knowing the others. Therefore, there are $r + c - 1$ totals that need to be pre-determined. Hence, the degrees of freedom are:

$$\nu = rc - (r + c - 1) = rc - r - c + 1 = (r - 1)(c - 1).$$

Worked Example 8

A group of 100 students, 40 males and 60 females, were asked to taste some Marmite. The students were then asked whether they liked the Marmite. Ten of the males liked the Marmite, and 40 of the females disliked the Marmite. Test, at the 5% significance level, whether there is an association between gender and Marmite preference. Calculate an interval for the p-value. Does this support the conclusion of your test?

Solution:

The null and alternative hypotheses are:

H_0 – there is no association between gender and Marmite preference.

H_1 – there is an association between gender and Marmite preference.

As per example 7, the contingency table is as follows:

	Like Marmite	*Dislike Marmite*	*Total*
Male	10	30	40
Female	20	40	60
Total	30	70	100

Assuming H_0 is true, we can calculate the expected frequencies by multiplying the row and column totals and dividing by the overall total. This gives the table of expected values:

	Like Marmite	*Dislike Marmite*
Male	12	28
Female	18	42

The table of chi-square contributions is found by calculating $\frac{(O-E)^2}{E}$ for each cell:

	Like Marmite	*Dislike Marmite*
Male	0.3333	0.1429
Female	0.2222	0.0952

Summing the values in the table we obtain the value of test statistic, x, i.e.,

$$x = \sum \frac{(O-E)^2}{E} = 0.3333 + 0.1429 + 0.2222 + 0.0952 = 0.7937.$$

Alternatively, we calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{10^2}{12} + \frac{30^2}{28} + \frac{20^2}{18} + \frac{40^2}{42} - 100 = 100.7937 - 100 = 0.7937.$$

The test statistic follows a chi-square distribution with $\nu = (r-1)(c-1) = 1 \times 1 = 1$ degrees of freedom. The critical value for a chi-square distribution with 1 degrees of freedom at the 5% significance level is $x_c = x_{0.95,1} = 3.841$. Therefore, the critical region is $x \geq 3.841$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, there is insufficient evidence to reject H_0 and so, at the 5% significance level, there is insufficient evidence of an association between gender and Marmite preference.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(1)$. From tables, $P(X \leq 0.016) = 0.1$ and $P(X \leq 2.706) = 0.9$. Therefore, $P(X \geq 0.016) = 0.9$ and $P(X \geq 2.706) = 0.1$ and so we deduce that $0.1 < p < 0.9$. We can be certain that the p-value is greater than the significance level, which further supports there being insufficient evidence to reject H_0.

Worked Example 9

A random sample of 400 social media users were classified according to their gender and age group (under 30, 30-59, 60 and over). A summary of the results is given in the table below.

a) Use an appropriate chi-squared test using a 5% significance level to examine whether there is an association between age group and gender. State the null and alternative hypotheses clearly and include a table showing the contributions of each cell to the test statistic.
b) Comment on the chi-squared contributions for each of the groups.
c) A website claims that men are more likely to use social media as they get older. Is there evidence to support the website's claim?

	Male	*Female*	*Total*
Under 30	78	76	154
30-59	83	59	142
60 and over	75	29	104
Total	236	164	400

Solution:
(a) The null and alternative hypotheses are:
H_0: there is no association between age group and gender of social media users
H_1: there is an association between age group and gender of social media users

Assuming H_0 is true, we can calculate the expected frequencies by multiplying the row and column totals and dividing by the overall total. This gives the table of expected values:

	Male	*Female*
Under 30	$\frac{154\times236}{400} = 90.86$	63.14
30-59	83.78	58.22
60 and over	61.36	42.64

The table of chi-squared contributions is found by calculating $\frac{(O-E)^2}{E}$ for each cell:

	Male	*Female*
Under 30	$\frac{(78-90.86)^2}{90.86}$ =1.820158	2.619252
30-59	0.007262	0.01045
60 and over	3.032099	4.363265

Summing the values in the table we obtain the value of test statistic, x, i.e.,

$$x = \sum \frac{(O-E)^2}{E} = 1.820158 + 2.619252 + \cdots + 4.363265 = 11.852 \text{ (to 3 d.p.)}.$$

The test statistic follows a chi-squared distribution with $\nu = (r-1)(c-1) = 2 \times 1 = 2$ degrees of freedom. The critical value for a chi-squared distribution with 2 degrees of freedom at the 5% significance level is $x_c = x_{0.95,2} = 5.991$. Hence, the critical region is $x \geq 5.991$.

Since $x \geq x_c$, the test statistic lies in the critical region. Therefore, there is sufficient evidence to reject H_0 and so, at the 5% significance level, there is sufficient evidence to conclude that there is an association between age group and gender of social media users.

(b) The chi-squared contributions for the 30-59 group are very small for both males and females. This means the observed values in this group are close to what would be expected if there were no association. The chi-squared contributions for the other 4 cells are above 1, with the two cells for 60 and overs having the largest contributions. The data in these cells differs the most from what would be expected if there was no association between age and

gender. For the 60 and over age group, there are fewer females and more males than would be expected if there were no association.

(c) For this sample, there are more females in the under 30 group and fewer in the 60 and over age group than would be expected if there were no association. The reverse is true for males. Hence, the data does support the website's claim.

Summary of Approach for the Chi-Square Goodness of Fit Test

1. State the null and alternative hypotheses in terms of association between the variables.
2. Calculate the expected frequency for each cell, assuming the null hypothesis is true.
3. Find the value of the test statistic by either creating a table of chi-square contributions $\frac{(O-E)^2}{E}$ for each class and then calculating $X = \sum \frac{(O-E)^2}{E}$, or calculating $X = \sum \frac{O^2}{E} - N$.
4. The degrees of freedom are calculated as $\nu = (r-1)(c-1)$.
5. Find the critical region using the upper one-tail critical value from the percentage points of the chi-square distribution table.
6. State the conclusion in terms of the hypotheses and in context:
    a. If the test statistic does not lie in the critical region, there is insufficient evidence to reject H_0, which suggests there is not an association between the variables.
    b. If the test statistic does lie in the critical region, there is sufficient evidence to reject H_0, suggesting that there is an association between the variables.
7. The p-value is calculated as $P(X \geq x)$, where $X \sim \chi^2(\nu)$. Very often, it is only possible to obtain intervals for the p-value.

Further Mathematics AS Unit 2: Further Statistics A

Questions

1. Using the percentage points of the χ^2-distribution table, find the values of x:
   a) $P(X \leq x) = 0.95$ where $X \sim \chi^2(24)$
   b) $P(X \leq x) = 0.01$ where $X \sim \chi^2(45)$
   c) $P(X \geq x) = 0.01$ where $X \sim \chi^2(4)$
   d) $P(X \geq x) = 0.975$ where $X \sim \chi^2(13)$

2. Find the following values, and write them as probability statements of the form $P(X \leq x) = p$, stating clearly the distribution of X:
   a) $x_{0.95,15}$
   b) $x_{0.025,26}$
   c) $x_{0.1,95}$
   d) $x_{0.99,23}$

3. Let $X \sim \chi^2(22)$. Find intervals for the following probabilities. In each case, give your answer in the form $a < p < b$, where a,b are values to be determined.
   a) $p = P(X \leq 32)$
   b) $p = P(X \geq 10)$
   c) $p = P(X \geq 7)$
   d) $p = P(X \geq 41)$
   e) $p = P(X \leq 50)$
   f) $p = P(X \leq 13)$

4. Applicants for a national teacher training course are required to pass a mathematics test. Each year, the applicants are tested in groups of 6, and the number of successful applicants in each group is recorded. The overall proportion of successful applicants has remained constant over the years and is equal to 60% of the applicants. The results from 150 randomly chosen groups are shown in the following table. Test, at the 1% significance level, the goodness of fit of the distribution $B(6,0.6)$ for the number of successful applicants in a group and find an interval for the p-value.

Number of successful applicants	0	1	2	3	4	5	6
Frequency	1	3	25	51	38	30	2

5. The number of customers entering a shop during a random sample of 100 one-minute intervals was recorded. The results are shown in the table below. Use a goodness of fit test with a 10% level of significance to determine whether a Poisson distribution with mean 2.7 would be a suitable fit to the data. Calculate an interval for the corresponding p-value and explain whether this supports the conclusion of the test.

Number of customers	0	1	2	3	4	5	6	7
Frequency	7	17	21	19	17	14	4	1

6. A spinner has ten sections numbered from 0 to 9. The spinner should be designed in such a way that it is equally likely to land in any of the ten sections. The spinner was spun 200 times and the number of times it landed in each section is listed in the table below. Investigate, using a 1% level of significance and an appropriate goodness of fit test, whether the spinner has been designed correctly.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	14	18	22	26	22	16	30	10	26	16

7. Customers were asked which of three brands of coffee, A, B and C, they prefer. For a random sample of 80 male customers and 60 female customers, the numbers preferring each brand are shown in the following table. Test, at the 5% significance level, whether there is a difference between coffee preferences of male and female customers.

	A	*B*	*C*
Male	32	36	12
Female	18	30	12

A larger random sample is now taken. It contains $80n$ male customers and $60n$ female customers, where n is a positive integer. It is found that the proportions choosing each brand are identical to those in the smaller sample. Find the least value of n that would lead to a different conclusion for the 5% significance level hypothesis test.

8. A researcher is investigating the relationship between education level (school, college, university) and home ownership status (own, rent, living with family/friends, other). She asks a random sample of 1000 people their education level and home ownership status. The results are summarised in the contingency table below. Test, at the 0.5% significance level, whether there is a relationship between home ownership and education level. Calculate an interval for the corresponding p-value.

	School	*College*	*University*	*Total*
Own a Home	150	185	207	542
Rent a Home	80	81	66	227
Living with Family/Friends	47	56	15	118
Other	32	47	34	113
Total	309	369	322	1000

9. A group of pupils are investigating the relationship between the school level (primary, secondary and sixth form) and preferred time of day (mornings, afternoons or evenings) of children in their county. A random sample of 400 pupils were selected and their responses summarised in the contingency table below.

	Primary	*Secondary*	*Sixth Form*
Mornings	59	68	36
Afternoons	57	30	18
Evenings	19	48	65

a) Explain why a chi-squared test is appropriate.
b) State the null and alternative hypotheses for the test.
c) Provide a table of chi-squared contributions. Without computing the test statistic, explain how it is possible to deduce that the result of the test will be significant at the 5% level of significance.
d) State the conclusion of the test in context.

e) If 3 of the 400 students are selected at random, one from each of the 3 school levels, find the probability that all 3 of them prefer mornings. Give your answer correct to 4 significant figures.

Question 1 Worked Solution

Using the percentage points of the χ^2-distribution table, find the values of x such that:

a) $P(X \leq x) = 0.95$ where $X \sim \chi^2(24)$, $\boldsymbol{x = 36.415}$
b) $P(X \leq x) = 0.01$ where $X \sim \chi^2(45)$, $\boldsymbol{x = 25.901}$
c) $P(X \geq x) = 0.01$ where $X \sim \chi^2(4)$, $\boldsymbol{x = 13.277}$ **(since** $\boldsymbol{P(X \leq x) = 0.99}$)
d) $P(X \geq x) = 0.975$ where $X \sim \chi^2(13)$, $\boldsymbol{x = 5.009}$ **(since** $\boldsymbol{P(X \leq x) = 0.025}$)

Question 2 Worked Solution

Find the following values, and write them as probability statements of the form $P(X \leq x) = p$, stating clearly the distribution of X:

a) $x_{0.95,15} = 24.996$, $P(X \leq 24.996) = 0.95$ where $X \sim \chi^2(15)$
b) $x_{0.025,26} = 13.844$, $P(X \leq 13.844) = 0.025$ where $X \sim \chi^2(26)$
c) $x_{0.1,95} = 77.818$, $P(X \leq 77.818) = 0.1$ where $X \sim \chi^2(95)$
d) $x_{0.99,23} = 41.638$, $P(X \leq 41.638) = 0.99$ where $X \sim \chi^2(23)$

Question 3 Worked Solution

Let $X \sim \chi^2(22)$. Find intervals for the following probabilities. In each case, give your answer in the form $a < p < b$, where a, b are values to be determined.

a) $p = P(X \leq 32)$
Since $P(X \leq 30.813) = 0.9$ and $P(X \leq 33.924) = 0.95$, then $0.9 < p < 0.95$.

b) $p = P(X \geq 10)$
Since $P(X \leq 9.542) = 0.01$ and $P(X \leq 10.982) = 0.025$, then $P(X \geq 9.542) = 0.99$ and $P(X \geq 10.982) = 0.975$. Therefore, $0.975 < p < 0.99$.

c) $p = P(X \geq 7)$
Since $P(X \leq 8.643) = 0.005$ then $P(X \geq 8.643) = 0.995$. Therefore, $0.995 < p < 1$.

d) $p = P(X \geq 41)$
Since $P(X \leq 40.289) = 0.99$ and $P(X \leq 42.796) = 0.995$, then $P(X \geq 40.289) = 0.01$ and $P(X \geq 42.796) = 0.005$. Therefore, $0.005 < p < 0.01$.

e) $p = P(X \leq 50)$
Since $P(X \leq 42.796) = 0.995$, then $0.995 < p < 1$.

f) $p = P(X \leq 13)$
Since $P(X \leq 12.338) = 0.05$ and $P(X \leq 14.041) = 0.1$, then $0.05 < p < 0.1$.

Question 4 Worked Solution

Applicants for a national teacher training course are required to pass a mathematics test. Each year, the applicants are tested in groups of 6, and the number of successful applicants in each group is recorded. The overall proportion of successful applicants has remained constant over the years and is equal to 60% of the applicants. The results from 150 randomly chosen groups are shown in the following table. Test, at the 1% significance level,

the goodness of fit of the distribution $B(6,0.6)$ for the number of successful applicants in a group and find an interval for the p-value.

Number of successful applicants	0	1	2	3	4	5	6
Frequency	1	3	25	51	38	30	2

The null and alternative hypotheses are:
H_0 – the number of successful applicants in a group can be modelled by $B(6,0.6)$.
H_1 – the number of successful applicants in a group cannot be modelled by $B(6,0.6)$.

Let X denote the number of successful applicants in a group. Under H_0, $X \sim \mathrm{B}(6,0.6)$. This distribution can only take values between 0 and 6. The probabilities and expected frequencies for the various values of X are calculated as follows:

$\boldsymbol{x}$	$\boldsymbol{P(X=x)}$	**Expected Frequency =** $\boldsymbol{150 \times P(X=x)}$
0	$0.4^6 = 0.0041$	0.6144
1	$\binom{6}{1} \times 0.6 \times 0.4^5 = 0.0369$	5.5296
2	$\binom{6}{2} \times 0.6^2 \times 0.4^4 = 0.1382$	20.736
3	$\binom{6}{3} \times 0.6^3 \times 0.4^3 = 0.2765$	41.472
4	$\binom{6}{4} \times 0.6^4 \times 0.4^2 = 0.3110$	46.656
5	$\binom{6}{5} \times 0.6^5 \times 0.4 = 0.1866$	27.9936
6	$0.6^6 = 0.0467$	6.9984

The class for $x = 0$ has expected frequency less than 5. We need to pool classes to alleviate this problem. We combine classes for $x = 0$ and $x = 1$. Re-writing the table to include the observed and expected frequencies as well as the values of the chi-square contribution $\frac{(O-E)^2}{E}$ for each class:

No. of successful applicants	0 and 1	2	3	4	5	6
Observed Frequency, O	4	25	51	38	30	2
Expected Frequency, E	6.144	20.736	41.472	46.656	27.9936	6.9984
Chi-Square Contribution	0.7482	0.8768	2.1890	1.6059	0.1438	3.5700

A useful check is $\Sigma O = \Sigma E = 150$. We calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{(O-E)^2}{E} = 0.7482 + \cdots + 3.5700 = 9.1337.$$

Alternatively, we can calculate x without using the chi-square contributions as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{4^2}{6.144} + \frac{25^2}{20.736} + \cdots + \frac{2^2}{6.9984} - 150 = 159.1337 - 150 = 9.1337.$$

The test statistic follows a chi-square distribution with $\nu = 6 - 1 = 5$ degrees of freedom. The critical value for a chi-square distribution with 5 degrees of freedom at the 1% significance level is $x_c = x_{0.99,5} = 15.086$. Therefore, the critical region is $x \geq 15.086$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, at the 1% significance level, there is insufficient evidence to reject H_0, and so we conclude that a binomial distribution $B(6,0.6)$ is an appropriate model for the number of successful applicants in a group.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(5)$. From tables, $P(X \geq 1.610) = 0.9$ and $P(X \geq 9.236) = 0.1$. Therefore, $0.1 < p < 0.9$. We can be certain that the p-value is greater than the significance level, further supporting the assertion that H_0 should not be rejected.

Question 5 Worked Solution

The number of customers entering a shop during a random sample of 100 one-minute intervals was recorded. The results are shown in the table below. Use a goodness of fit test with a 10% level of significance to determine whether a Poisson distribution with mean 2.7 would be a suitable fit to the data. Calculate an interval for the corresponding p-value and explain whether this supports the conclusion of the test.

Number of customers	0	1	2	3	4	5	6	7
Frequency	7	17	21	19	17	14	4	1

The null and alternative hypotheses are:
H_0: the number of customers arriving each minute can be modelled by $\mathrm{Po}(2.7)$.
H_1: the number of customers arriving each minute cannot be modelled by $\mathrm{Po}(2.7)$.

Under H_0, a Poisson distribution can be used to describe the number of customers arriving each minute. Whilst 7 was the greatest number of customers recorded arriving per minute in this sample, the Poisson distribution can theoretically take any non-negative integer value (i.e. a value from 0, 1, 2,…). Therefore, we change the final category to 7 or more rather than 7 when calculating the expected frequencies.

Let X denote the number of customers arriving during a one-minute interval. Under H_0, $X \sim \mathrm{Po}(2.7)$. The probabilities and expected frequencies for the various values of X are calculated as follows:

x	$P(X = x)$	Expected Frequency = 100 × $P(X = x)$
0	$e^{-2.7} = 0.0672$	6.7206
1	$\frac{e^{-2.7} \times 2.7}{1!} = 0.1815$	18.1455
2	$\frac{e^{-2.7} \times 2.7^2}{2!} = 0.2450$	24.4964
3	$\frac{e^{-2.7} \times 2.7^3}{3!} = 0.2205$	22.0468
4	$\frac{e^{-2.7} \times 2.7^4}{4!} = 0.1488$	14.8816
5	$\frac{e^{-2.7} \times 2.7^5}{5!} = 0.0804$	8.0360
6	$\frac{e^{-2.7} \times 2.7^5}{5!} = 0.0362$	3.6162
≥ 7	$1 - P(X \leq 6) = 0.0206$	2.0569

Since the expected frequencies for 6 and ≥ 7 are less than 5, we pool classes. The classes for $x = 6$ and $x \geq 7$ are combined. This results in all classes having expected frequencies that are 5 or more.

Re-writing the table including the observed and expected frequencies along with the chi-squared contributions $\frac{(O-E)^2}{E}$ for each class we have:

Number of customers	0	1	2	3	4	5	6 or more
Observed Frequency, O	7	17	21	19	17	14	5
Expected Frequency, E	6.7206	18.1455	24.4964	22.0468	14.8816	8.0360	5.6732
Chi-Squared Contribution	0.011616	0.072314	0.499045	0.421058	0.301555	4.426244	0.079884

Note that the expected frequencies should not be rounded to the nearest whole number, since expected frequencies do not necessarily need to be integers. A useful check is that $\Sigma O = \Sigma E = 100$. We calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{(O-E)^2}{E} = 0.011616 + \cdots + 0.079884 = 5.8117.$$

Alternatively, we can calculate x without using the chi-square contributions as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{7^2}{6.7206} + \cdots + \frac{5^2}{5.6732} - 100 = 105.8117 - 100 = 5.8117.$$

The test statistic follows a chi-square distribution with $\nu = 7 - 1 = 6$ degrees of freedom. The critical value for a chi-square distribution with 6 degrees of freedom at the 10% significance level is $x_c = x_{0.9,6} = 10.645$. Therefore, the critical region is $x \geq 10.645$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, there is insufficient evidence to reject H_0, and so, at the 10% significance level, we conclude that a Poisson distribution with mean 2.7 can be used to model the number of customers arriving each minute.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(6)$. From tables, $P(X \leq 2.204) = 0.1$ and $P(X \leq 10.645) = 0.9$. Therefore, $P(X \geq 2.204) = 0.9$ and $P(X \geq 10.645) = 0.1$, and so we deduce that $0.1 < p < 0.9$. We can be certain that the p-value is greater than the significance level, which further supports there being insufficient evidence to reject H_0.

Question 6 Worked Solution

A spinner has ten sections numbered from 0 to 9. The spinner should be designed in such a way that it is equally likely to land in any of the ten sections. The spinner was spun 200 times and the number of times it landed in each section is listed in the table below. Investigate, using a 1% level of significance and an appropriate goodness of fit test, whether the spinner has been designed correctly.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	14	18	22	26	22	16	30	10	26	16

The null and alternative hypotheses are:
H_0: the distribution of sections follows a discrete uniform distribution with $p = 0.1$.
H_1: the distribution of sections does not follow a discrete uniform distribution with $p = 0.1$.

Under H_0, a discrete uniform distribution can be used to describe the distribution of the sections. For a sample of 200 spins, we would expect the spinner to land on each section twenty times.

The expected frequencies are all greater than 5 so no pooling is required. We create a table including the observed and expected frequencies along with the chi-squared contribution $\frac{(O-E)^2}{E}$ for each class:

Digit	0	1	2	3	4	5	6	7	8	9
Observed Frequency, O	14	18	22	26	22	16	30	10	26	16
Expected Frequency, E	20	20	20	20	20	20	20	20	20	20
Chi-Squared Contribution	1.8	0.2	0.2	1.8	0.2	0.8	5	5	1.8	0.8

We calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{(O-E)^2}{E} = 1.8 + 0.2 + \cdots 1.8 + 0.8 = 17.6.$$

Alternatively, we can calculate x without using the chi-squared contributions as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{14^2}{20} + \frac{18^2}{20} + \cdots + \frac{16^2}{20} - 200 = 217.6 - 200 = 17.6.$$

The test statistic follows a chi-squared distribution with $\nu = 10 - 1 = 9$ degrees of freedom. The critical value for a chi-squared distribution with 9 degrees of freedom at the 1% significance level is $x_c = x_{0.99,9} = 21.666$. Therefore, the critical region is $x \geq 21.666$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, there is insufficient evidence to reject H_0 and so, at the 1% significance level, we conclude a discrete uniform distribution can be used to model the distribution of sections the spinner lands on. Therefore, there is insufficient evidence to suggest that the spinner is not designed correctly.

Question 7 Worked Solution

Customers were asked which of three brands of coffee, A, B and C, they prefer. For a random sample of 80 male customers and 60 female customers, the numbers preferring each brand are shown in the following table. Test, at the 5% significance level, whether there is a difference between coffee preferences of male and female customers.

	A	*B*	*C*
Male	32	36	12
Female	18	30	12

A larger random sample is now taken. It contains $80n$ male customers and $60n$ female customers, where n is a positive integer. It is found that the proportions choosing each brand are identical to those in the smaller sample. Find the least value of n that would lead to a different conclusion for the 5% significance level hypothesis test.

The null and alternative hypotheses are:
H_0 – there is no association between gender and coffee brand preference.
H_1 – there is an association between gender and coffee brand preference.

The contingency table with totals is as follows:

	A	*B*	*C*	*Total*
Male	32	36	12	80
Female	18	30	12	60
Total	50	66	24	140

Assuming H_0 is true, we can calculate the expected frequencies by multiplying the row and column totals and dividing by the overall total. This gives the table of expected values:

	A	*B*	*C*
Male	28.5714	37.7143	13.7143
Female	21.4286	28.2857	10.2857

The table of chi-square contributions is found by calculating $\frac{(O-E)^2}{E}$ for each cell:

	A	B	C
Male	0.4114	0.0779	0.2143
Female	0.5486	0.1039	0.2857

Summing the values in the table, we obtain the value of test statistic, x, i.e.,

$$x = \sum \frac{(O-E)^2}{E} = 0.4114 + \cdots + 0.2857 = 1.6418.$$

Alternatively, we calculate the value of the test statistic, x, as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{32^2}{28.5714} + \cdots + \frac{12^2}{10.2857} - 140 = 141.6418 - 140 = 1.6418.$$

The test statistic follows a chi-square distribution with $\nu = (r-1)(c-1) = 1 \times 2 = 2$ degrees of freedom. The critical value for a chi-square distribution with 2 degrees of freedom at the 5% significance level is $x_c = x_{0.95,2} = 5.991$. Therefore, the critical region is $x \geq 5.991$.

Since $x < x_c$, the test statistic does not lie in the critical region. Therefore, there is insufficient evidence to reject H_0, and so, at the 5% significance level, there is insufficient evidence of an association between gender and coffee brand preference.

The contingency table for the larger random sample is as follows:

	A	B	C	Total
Male	$32n$	$36n$	$12n$	$80n$
Female	$18n$	$30n$	$12n$	$60n$
Total	$50n$	$66n$	$24n$	$140n$

The table of expected values is:

	A	B	C
Male	$28.5714n$	$37.7143n$	$13.7143n$
Female	$21.4286n$	$28.2857n$	$10.2857n$

The table of chi-square contributions is:

	A	B	C
Male	$0.4114n$	$0.0779n$	$0.2143n$
Female	$0.5486n$	$0.1039n$	$0.2857n$

Summing the values in the table, we obtain the value of test statistic, x, i.e.,

$$x = \ 0.4114n + \cdots + 0.2857n = 1.6418n.$$

We reject H_0 when x lies in the critical region, i.e. $x \geq 5.991$. We therefore require:

$$1.6418n \geq 5.991 \qquad \Rightarrow \qquad n \geq 3.649.$$

Therefore, the least value of n that would lead to a different conclusion for the 5% significance level is 4.

Question 8 Worked Solution

A researcher is investigating the relationship between education level (school, college, university) and home ownership status (own, rent, living with family/friends, other). She asks a random sample of 1000 people their education level and home ownership status. The results are summarised in the contingency table below. Test, at the 0.5% significance level, whether there is a relationship between home ownership and education level. Calculate an interval for the corresponding p-value.

	School	*College*	*University*	*Total*
Own a Home	150	185	207	542
Rent a Home	80	81	66	227
Living with Family/Friends	47	56	15	118
Other	32	47	34	113
Total	309	369	322	1000

The null and alternative hypotheses are:
H_0: there is no association between education level and home ownership status
H_1: there is an association between education level and home ownership status

Assuming H_0 is true, we can calculate the expected frequencies by multiplying the row and column totals and dividing by the overall total. This gives the table of expected values:

	School	*College*	*University*
Own a Home	167.478	199.998	174.524
Rent a Home	70.143	83.763	73.094
Living with Family/Friends	36.462	43.542	37.996
Other	34.917	41.697	36.386

The table of chi-squared contributions is found by calculating $\frac{(O-E)^2}{E}$ for each cell:

	School	*College*	*University*
Own a Home	1.8240	1.1247	6.0432
Rent a Home	1.3852	0.0911	0.6885
Living with Family/Friends	3.0456	3.5644	13.9177
Other	0.2437	0.6744	0.1565

Summing the values in the table we obtain the value of test statistic, x, i.e.,

$$x = \sum \frac{(O-E)^2}{E} = 1.8240 + 1.1247 + \cdots + 0.1565 = 32.7591.$$

Alternatively, we can calculate the value of the test statistic as follows:

$$x = \sum \frac{O^2}{E} - N = \frac{150^2}{167.478} + \frac{185^2}{199.998} + \cdots + \frac{34^2}{36.386} - 1000 = 32.7591.$$

The test statistic follows a chi-squared distribution with $\nu = (r-1)(c-1) = 3 \times 2 = 6$ degrees of freedom. The critical value for a chi-squared distribution with 6 degrees of freedom at the 0.5% significance level is $x_c = x_{0.995,6} = 18.548$. Hence, the critical region is $x \geq 18.548$.

Since $x \geq x_c$, the test statistic lies in the critical region. Therefore, there is sufficient evidence to reject H_0 and so, at the 0.5% significance level, there is sufficient evidence to conclude that there is an association between education level and the home ownership status.

The p-value is $p = P(X \geq x)$, where $X \sim \chi^2(6)$. From tables, $P(X \leq 18.548) = 0.995$ and so $P(X \geq 18.548) = 0.005$. Therefore, $0 < p < 0.005$. We can be certain that the p-value is smaller than the significance level, which further supports there being sufficient evidence to reject H_0 and conclude that there is a relationship between home ownership status and education level.

Question 9 Worked Solution

A group of pupils are investigating the relationship between the school level (primary, secondary and sixth form) and preferred time of day (mornings, afternoons or evenings) of children in their county. A random sample of 400 pupils were selected and their responses summarised in the contingency table below.

	Primary	*Secondary*	*Sixth Form*
Mornings	59	68	36
Afternoons	57	30	18
Evenings	19	48	65

i) Explain why a chi-squared test is appropriate.
ii) State the null and alternative hypotheses for the test.
iii) Provide a table of chi-squared contributions. Without computing the test statistic, explain how it is possible to deduce that the result of the test will be significant at the 5% level of significance.
iv) State the conclusion of the test in context.
v) If 3 of the 400 students are selected at random, one from each of the 3 school levels, find the probability that all 3 of them prefer mornings. Give your answer correct to 4 significant figures.

(i) Since the variables 'time of day' and 'school level' are both categorical variables and we are looking for an association between them, a chi-squared test is appropriate.

(ii) The null and alternative hypotheses are:
H_0: there is no association between the school level of pupils and their preferred time of day
H_1: there is an association between the school level of pupils and their preferred time of day

(iii) The contingency table with totals is as follows:

	Primary	*Secondary*	*Sixth Form*	*Total*
Mornings	59	68	36	163
Afternoons	57	30	18	105
Evenings	19	48	65	132
Total	135	146	119	400

Assuming H_0 is true, we can calculate the expected frequencies by multiplying the row and column totals and dividing by the overall total. This gives the table of expected values:

	Primary	*Secondary*	*Sixth Form*
Mornings	55.0125	59.495	48.4925
Afternoons	35.4375	38.325	31.2375
Evenings	44.55	48.18	39.27

The table of chi-squared contributions is found by calculating $\frac{(O-E)^2}{E}$ for each cell:

	Primary	*Secondary*	*Sixth Form*
Mornings	0.2890	1.2158	3.2183
Afternoons	13.1200	1.8084	5.6096
Evenings	14.6533	0.0007	16.8585

The test statistic follows a chi-squared distribution with $\nu = (r-1)(c-1) = 2 \times 2 = 4$ degrees of freedom. The critical value for a chi-squared distribution with 4 degrees of freedom at the 5% significance level is $x_c = x_{0.95,4} = 9.488$. Therefore, the critical region is given by $x \geq 9.488$.

However, three of the chi-squared contributions are larger than the critical value. Therefore, it is guaranteed that the test statistic, which is the sum of the chi-squared contributions, will be larger than the critical value.

This means the test statistic lies in the critical region. Therefore, there is sufficient evidence to reject H_0 and so, at the 5% significance level, there is evidence to suggest that there is an association between the school level of pupils and their preferred time of day.

vi) The probability of choosing a primary pupil who prefers mornings is $\frac{59}{135}$. The probability of choosing a secondary pupil who prefers mornings is $\frac{68}{146}$. The probability of choosing a sixth form pupil who prefers mornings is $\frac{36}{119}$.

Therefore, the probability that 3 pupils are selected who prefer mornings when choosing a pupil from each school level is:

$$\frac{59}{135} \times \frac{68}{146} \times \frac{36}{119} = \frac{472}{7665} = 0.06158 \text{ (to 4 s. f.)}.$$