

INTERNATIONAL ADVANCED LEVEL

MATHEMATICS/ FURTHER MATHEMATICS/ PURE MATHEMATICS

SCHEME OF WORK STATISTICS 3

Pearson Edexcel International Advanced Subsidiary in Mathematics (XMA01)

Pearson Edexcel International Advanced Subsidiary in Further Mathematics (XFM01)

Pearson Edexcel International Advanced Subsidiary in Pure Mathematics (XPM01)

Pearson Edexcel International Advanced Level in Mathematics (YMA01)

Pearson Edexcel International Advanced Level in Further Mathematics (YFM01)

Pearson Edexcel International Advanced Level in Pure Mathematics (YPM01)

First teaching September 2018

First examination from January 2019

First certification from August 2019 (International Advanced Subsidiary) and August 2020
(International Advanced Level)



Statistics 3

Unit	Title	Estimated hours
1	Combinations of random variables: Distribution of linear combinations of independent Normal random variables	7
2	Statistical sampling: Understand and use sampling techniques; Compare sampling techniques in context	4
3	Estimation, confidence intervals and tests	
<u>a</u>	Concepts of standard error, estimator and bias, including the quality of estimators	5
<u>b</u>	Concept of a confidence interval and its interpretation	2
<u>c</u>	Confidence interval for the mean of a Normal distribution with known variance	3
<u>d</u>	Statistical hypothesis testing for the mean of the Normal distribution	6
<u>e</u>	Use of the Central Limit Theorem	4
<u>f</u>	Hypothesis test for the difference between the means of two independent Normal distributions with known variances	3
<u>g</u>	Use of large sample results, hypothesis test for the difference between the means of two independent distributions with unknown variances	3
4	Goodness of fit and contingency tables: Chi-squared tests	12
5	Regression and correlation	
<u>a</u>	Spearman's rank correlation coefficient	5
<u>b</u>	Hypothesis testing for zero correlation	6
		60 hours

UNIT 1: Combinations of random variables
Distribution of linear combinations of independent Normal random variables (1.1)**Teaching time**
7 hours[Return to Overview](#)**SPECIFICATION REFERENCES**

1.1 Distribution of linear combinations of independent Normal random variables

PRIOR KNOWLEDGEIAL Mathematics – Statistics 1 content

- The Normal distribution (See S1 SoW Unit 5)
- Expectation and variance rules for linear functions of X (See S1 SoW Unit 4)

KEYWORDS

Normal distribution, expectation, mean, variance, standard deviation, independent random variables, sample, linear.

OBJECTIVES

By the end of the unit, students should:

- recall and be able to use the rule for the expectation of a linear combination of independent random variables;
- be able to use the rule from the formulae book for the variance of a linear combination of independent random variables;
- recognise the difference between the distributions of e.g. $X_1 + X_2$ and $2X$;
- be able to define a random variable to use that is appropriate to the context of a question.

TEACHING POINTS

It is worth starting with a reminder of Normal distribution problems, finding probabilities both on the calculator and by standardising first. Include examples in which students are required to find an unknown value for the variable X , the mean and/or variance having been given probability value/s.

Extend the ideas of $E(aX + b) = aE(X) + b$ and rules for $E(g(X))$ to $E(aX + bY) = aE(X) + bE(Y)$. Students need to be aware that X and Y are independent variables.

Introduce $\text{Var}(aX)$ and use $\text{Var}(aX) = E((aX)^2) - (E(aX))^2$ to prove $\text{Var}(aX) = a^2\text{Var}(X)$.

Extend this to the rule for variance, $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$, where X and Y are independent.

You should stress the importance of squaring the constants here, especially when a or b are negative. Remind students that a variance can never be negative.

Compare the distributions of e.g. $X_1 + X_2$ and $2X$, note that these have the same mean but different variances, whilst e.g. $X_1 + X_2$ and $X_1 - X_2$ have different means but the same variance.

Introduce the setting up of suitable random variables to model problems in context. For example, if the weight of a randomly selected egg can be modelled by $X \sim N(50, 49)$ and the weight of a randomly selected egg box by $Y \sim N(80, 36)$ then the weight of 4 eggs in a box can be modelled by

$T = X_1 + X_2 + X_3 + X_4 + Y$, etc.

A level Further Mathematics: Statistics 3

Questions comparing two random variables are also important. For example, given the distributions of independent Normal random variables X and Y , find the probability that a randomly selected value from Y is more than double a randomly selected value from X . This can be found by using $D = Y - 2X$ and finding $P(D > 0)$.

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

Answers should always be given in the context of the question. Questions may be posed as straight Normal distribution calculations but are much more likely to be contextualised. Finding an unknown parameter given probabilities, or combining a Normal distribution problem with one using a discrete distribution would be a good way to develop problem-solving skills. For example, finding the sample size needed if a quality control officer randomly selecting boxes of 4 eggs needs to be sure that his sample has at least one box with a total weight of less than 250g.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

The most common problem is for students to successfully distinguish between the variables

$\sum_{i=1}^n X_i$ and nX .

Another common error is forgetting to square negative values when finding the variance of a distribution, e.g. when considering the distribution of $X - Y$.

NOTES

Proofs of formulae are not required but using $\text{Var}(aX) = E((aX)^2) - (E(aX))^2$ to prove $\text{Var}(aX) = a^2\text{Var}(X)$ is accessible and provides better understanding.

A level Further Mathematics: Statistics 3

UNIT 2: Statistical sampling Understand and use sampling techniques; Compare sampling techniques in context (2.1) (2.2)

Teaching time
4 hours

[Return to overview](#)

SPECIFICATION REFERENCES

- 2.1 Methods for collecting data. Simple random sampling. Use of random numbers for sampling
- 2.2 Other methods of sampling: stratified, systematic, quota

PRIOR KNOWLEDGE

International GCSE/GCSE (9–1) in Mathematics at Higher Tier

- Infer properties of populations or distributions from a sample, while knowing the limitations of sampling (GCSE only)
- Apply statistics to describe a population

IAL Mathematics – Statistics 2 content

- Sampling terminology; advantages and disadvantages of sampling (See S2 SoW Unit 4)

KEYWORDS

Population, census, sample, sampling unit, sampling frame, simple random sampling, stratified sampling, systematic sampling, quota sampling.

OBJECTIVES

By the end of the unit, students should:

- understand and be able to use sampling techniques;
- be able to describe advantages and disadvantages of sampling techniques;
- be able to select or critique sampling techniques in the context of solving a statistical problem;
- understand that different samples can lead to different conclusions about the population.

TEACHING POINTS

Students will also be expected to be familiar with different types of sampling including simple random, stratified, systematic and quota sampling, and be able to use random numbers.

Students will gain a more thorough understanding of the types of sampling if the advantages and disadvantages alongside the method used for each type are understood. They will then be more able to select an appropriate technique for a given statistical problem and be able to critique a technique which has been used.

Give students the opportunity to use the techniques they learn about on real-world data sets.

OPPORTUNITIES FOR PROBLEM SOLVING/MODELLING

This is a perfect opportunity to use real-world data sets and discuss how different samples from the same data set could lead to different conclusions.

A level Further Mathematics: Statistics 3

COMMON MISCONCEPTIONS/ EXAMINER REPORT QUOTES

Students need to be able to describe the sampling techniques clearly and will lose marks if they are not sufficiently precise.

As always, answers must be given using the context of the question and not simply be quoted from text books in a general form.

Some students confuse sample sizes and population sizes. Candidates need to be clear about the difference between sample sizes and population sizes.

UNIT 3: Estimation, confidence intervals and tests[Return to Overview](#)**SPECIFICATION REFERENCES**

- 3.1 Concepts of standard error, estimator, bias
- 3.2 The distribution of the sample mean $\bar{X}$
- 3.3 Concept of a confidence interval and its interpretation
- 3.4 Confidence limits for a Normal mean, with variance known
- 3.5 Hypothesis test for the means of a Normal distribution with variance known
- 3.6 Use of Central Limit theorem to extend hypothesis tests and confidence intervals to samples from non-Normal distributions. Use of large sample results to extend to the case in which the variance is unknown
- 3.7 Hypothesis test for the difference between the means of two Normal distributions with variances known
- 3.8 Use of large sample results to extend to the case in which the population variances are unknown

PRIOR KNOWLEDGE

- Expectation and variance rules (See Unit 1)

IAL Mathematics – Statistics 1 content

- Mean and variance (See S1 SoW Unit 1)
- The Normal distribution (See S1 SoW Unit 5)

IAL Mathematics - Statistics 2 content

- Hypothesis tests (See S2 SoW Unit 4)

KEYWORDS

Random variable, mean, expectation, expected value, standard deviation, variance, independent, sample, population, parameter, sample mean, sample size, sampling distribution, statistic, standard error, estimator, bias, unbiased estimators, Central Limit Theorem, confidence interval, confidence limits, confidence level, hypothesis test, null hypothesis, alternate hypothesis, test statistic, critical region, critical value, p -value, significance level, significant evidence, Normal distribution, approximation.

3a. Concepts of standard error, estimator and bias including the quality of estimators (3.1) (3.2)

Teaching time
5 hours

OBJECTIVES

By the end of the sub-unit, students should:

- be able to explain standard error and its relationship to the distribution of the sample mean, $\bar{X}$;
- understand how a sample statistic may be used to estimate a population parameter;
- be able to explain what is meant by an unbiased estimator;
- be able to explain what is meant by bias in an estimator.

TEACHING POINTS

Start by deriving the mean and variance of $\bar{X}$ where $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ and $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ to establish that the sampling distribution of the mean has mean μ and variance $\frac{\sigma^2}{n}$.

Define the standard error as the standard deviation of this statistic.

Explain that an estimator is a statistic used to find an estimate for a population parameter.

Stress that not all such statistics are good estimators.

Define an unbiased estimator as an estimator T for a parameter θ such that $E(T) = \theta$.

Define the bias of an estimator T for a parameter θ as $E(T) - \theta$, so that for an unbiased estimator the bias is 0.

Use 'hat' notation for estimators, i.e. $\hat{\mu}$ as an estimator for μ and $\hat{\sigma}^2$ as an estimator for σ^2 .

Students need to know that the sample mean $\bar{x}$ is used as an unbiased estimator for μ and that the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is used as an unbiased estimator for σ^2 .

It can be helpful if students know the alternative means of calculating s^2 , i.e. $\frac{S_{xx}}{n-1}$ and $\frac{n}{n-1} \left(\frac{\sum x^2}{n} - \bar{x}^2 \right)$ as well as being able to generate the value of s^2 on their calculators.

Explain that if you have two unbiased estimators then the one with the smallest variance is the best one to use.

A nice piece of extension work is to prove that s^2 is an unbiased estimator for σ^2 .

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

Questions may use summary statistics rather than raw data so students should be able to use, and undo, various versions of the formulae, e.g. for s^2 .

Questions are likely to tie together several subsections of this topic, e.g. using a confidence interval for μ to find the value of the standard error.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Examiners have highlighted that students are not always clear about what the standard error represents: 'However, very few candidates had any idea about what the comparison of standard errors was all about.'

3b. Concept of a confidence interval and its interpretation (3.3)**Teaching time**
2 hours**OBJECTIVES**

By the end of the sub-unit, students should

- be able to explain the meaning of the term confidence interval;
- be able to explain a confidence interval in the context of the question.

TEACHING POINTS

Students need to realise that given any sample, the value of $\bar{x}$ is only an estimate for the population mean μ and that there is no way of establishing from the sample data alone how far that estimate is from the real value of μ .

Students should know that a 95% confidence interval for a population parameter θ is one where there is a probability of 0.95 that the interval found contains θ , it is not true that the probability that θ lies in the interval is 0.95. It can also be useful to explain that if 100 samples are taken from a population and a 95% confidence interval for a population parameter θ is obtained from each sample then approximately 95 of these will contain θ . Mathematical software can also be used to demonstrate this. Stress that the parameter has a precise, though unknown, value; it is not a variable. It is the interval which is dependent on the values in the sample and therefore if the process were repeated the interval would change, i.e. the interval varies.

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

Consider problems tying in with the binomial distribution. For example, given 5 independent 95% confidence intervals for μ , find the probability that exactly 3 of them contain the true value of μ .

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

The definition of a confidence interval is very subtle and students often have difficulty remembering this correctly. In exam questions they may be asked to use this definition in the context of the question which can also be problematic.

Examiners made the comment: ‘Candidates need to understand that we do not talk about population means in terms of probability, but rather that the population mean is fixed and it is the confidence interval which does the moving.’

3c. Confidence interval for the mean of a Normal distribution with known variance (3.4)**Teaching time**
3 hours**OBJECTIVES**

By the end of the sub-unit, students should be able to:

- find a confidence interval for the mean of a normal distribution with known variance given the sample mean and sample size or the sample data;
- use a given confidence interval for the mean of a normal distribution to find an unknown such as the sample mean, sample size or population variance;
- comment on a claim made in the context of the question based on whether or not a new sample mean lies within a calculated confidence interval;
- explain the relationship between the width of the confidence interval and each of: the standard deviation of the population, the sample size and the size of confidence level used.

TEACHING POINTS

In order for students to understand the structure of the calculation for the confidence interval, $\left(\bar{x} - z \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z \times \frac{\sigma}{\sqrt{n}}\right)$ it is worth starting by going back to practising finding Normal distribution probabilities by standardisation.

Students should be able to find confidence intervals given raw data and σ or a summary statistic such as $\sum x$ along with n and σ . It is also worth working on the reverse process: using the confidence interval to find an unknown such as σ or n .

Students should be aware that these confidence intervals are symmetric about $\bar{x}$ and therefore $\bar{x}$ is the average of the confidence limits.

Explore the impact of varying each of n , σ and the confidence level used on the width of the confidence interval. Mathematical software can also be used to visualise this.

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

A key problem which students often find difficult is: when given a known confidence level and standard deviation of the population, to find the minimum sample size required to ensure that the width of the confidence interval is less than a stated value.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

An examiner comments on the difficulties encountered in interpreting a value compared to a confidence interval: 'Part (d) was completely misunderstood by a lot of candidates. There were many who stated that 20 000 was not in the interval, but missed the important point that it was above the interval's upper limit. Many did not understand what this meant in terms of the claim made.'

NOTES

Students will be expected to know how to apply the Normal distribution and use the standard error and obtain confidence intervals for the mean, rather than be concerned with any theoretical derivations.

3d. Statistical hypothesis testing for the mean of the Normal distribution (3.5)
Teaching time
6 hours

OBJECTIVES

By the end of the sub-unit, students should:

- be able to conduct a statistical hypothesis test for the mean of the Normal distribution;
- be able to interpret the results in context;
- know how confidence intervals link with hypothesis tests.

TEACHING POINTS

Remind students of the properties of the Normal distribution and the parameters it uses. Questions could involve a known, given or assumed variance and students should be aware of this.

Hypothesis tests need to be carried out for the mean of the Normal distribution. For $X \sim N(\mu, \sigma^2)$, students need to understand that for a sample, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Refer back to the formula used to translate X into Z and make sure students know they can test μ using $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1^2)$.

This is the third type of hypothesis test that students are expected to be able to carry out, so the importance of using the correct parameter in the hypotheses should be emphasised here. Hypotheses for the Normal distribution should be stated in terms of μ .

As in all cases, conclusions need to be written clearly and in the context of the question.

Highlight the use of the standardisation of $\bar{X}$, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ in both hypothesis tests for the mean and in the construction of the confidence interval for μ , $\left(\bar{x} - z \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z \times \frac{\sigma}{\sqrt{n}}\right)$.

OPPORTUNITIES FOR PROBLEM SOLVING/MODELLING

Students should be given a variety of questions in context. They can also be given a variety of mixed problems of all types of hypothesis tests in order to practise distinguishing between tests. Ensure all hypotheses are written in terms of the correct parameter.

COMMON MISCONCEPTIONS/ EXAMINER REPORT QUOTES

Common errors in exam situations include: not expressing hypotheses precisely enough; using an incorrect parameter or not using a parameter at all; incorrectly applying the continuity correction; and not giving a conclusion or answer to the question using the given context.

NOTES

Knowledge of the central limit theorem is not required, neither are proofs of the sample formulae used.

3e. Use of the central limit theorem (3.6)

Teaching time

4 hours

OBJECTIVES

By the end of the sub-unit, students should:

- be able to quote the central limit theorem;
- know the conditions for the use of the central limit theorem;
- recognise, from problems in context, where the central limit theorem is required;
- be able to use large sample results to extend the central limit theorem to the case in which the variance is unknown.

TEACHING POINTS

Students should know that for a population with mean μ and variance σ^2 , for large n , $\bar{X} \approx \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ regardless of the distribution of X .

Emphasise that this is an approximation and that the approximation improves as n increases.

Students should also understand that if the random variable X is itself normally distributed then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ in any case and the central limit theorem is not required, so there is no constraint on sample size in this case.

Compare the formulae for the population variance, $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ and the unbiased estimate for the variance, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, as $n \rightarrow \infty$, and point out that $s^2 \approx \sigma^2$ when n is large. The central limit can then be extended to the case in which the population variance, σ^2 , is unknown.

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

It sometimes takes a little time to convince students that this works! It can be useful to carry out spreadsheet exercises to generate largish samples of fixed size (perhaps $n = 50$) of, for example, random numbers and then study the distribution of their means. Other mathematical software such as Autograph can also be useful.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

One examiner made this comment about a question involving the use of the central limit theorem in context: 'The final part revealed a worrying lack of knowledge about the central limit theorem and its importance in testing. Far too many candidates are still regurgitating the text book and making no effort to answer in the context of the question. There was also some confusion between sample sizes and population sizes.'

NOTES

No proofs are required. Knowledge of the t -distribution is not required, but it is useful to point out that $\bar{X}$ has a t -distribution when σ^2 is unknown, and that $\bar{X}$ can only be approximated by a normal distribution when n is large.

3f. Hypothesis test for the difference between the means of two independent Normal distributions with known variances (3.7)
Teaching time
3 hours

OBJECTIVES

By the end of the sub-unit, students should:

- understand the necessary conditions for this test to be valid, and be able to state the underlying assumptions;
- be able to apply this test given either raw data or summary statistics for the two samples;
- be able to distinguish between one and two tail tests and state the appropriate null and alternative hypotheses;
- be able to use either critical values or p-values to determine the result of the test;
- be able to state the result of the hypothesis test in the context of the question;
- be able to explain the relevance of the significance level to the result of the test.

TEACHING POINTS

Emphasise the necessary conditions of independence and any assumptions for this hypothesis test.

Null and alternative hypotheses should always be stated in terms of the population means, not the sample means, so e.g. $H_0: \mu_x = \mu_y$ and $H_1: \mu_x \neq \mu_y$

Students need to think specifically about the context of the question in deciding whether the test is one or two-tailed and therefore the form that H_1 should take. They should be aware that in a real-life situation the null and alternative hypotheses would be set up before the data was collected and therefore their decision should be based on what is to be tested not on the values of the data obtained.

Remind students of the steps expected in any hypothesis test: statement of hypotheses, calculation of test statistic and critical value/s or p-value, comparison and statement of the result of the test, bearing in mind the significance level, and a conclusion in the context of the question.

In two tail tests the significance level is split between the two parts of the critical region so that e.g. for a 5% significance level the critical region in each tail is $\leq 2.5\%$

Students should be aware that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0, 1)$$

where $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, as given in the formulae book. It can be helpful to show where this comes from using the usual standardisation process and expectation and variance rules.

Students should ensure that they get in the habit of showing working for their test statistic, $\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$

The most common questions ask for a test of equality between the population means but the comparatively simple extension to asking for a test that the population mean μ_x is e.g. at least 3 more than μ_y is worth careful work as this can easily cause confusion.

A level Further Mathematics: Statistics 3

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

Setting up null and alternative hypotheses correctly is a key part of the question, giving students a mixed batch of questions and asking them to state H_0 and H_1 for each helps raise awareness of this key step.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Students frequently fail to give their conclusions in the context of the question, there is often a mark available specifically for this part of the hypothesis test.

The statement of the result of a hypothesis test is often misunderstood by students. Students need to understand that if the test statistic lies in the critical region then there is significant evidence at the level of the test to reject H_0 ; this does not however guarantee that H_1 is true. Similarly if the test statistic does not lie in the critical region then there is insufficient evidence at the level of the test to reject H_0 , therefore H_0 is accepted, but this does not mean that it must be correct!

3g. Use of large sample results, hypothesis test for the difference between the means of two independent distributions with unknown variances (3.8)

Teaching time
3 hours

OBJECTIVES

By the end of the sub-unit, students should:

- understand the necessary conditions for this test to be valid, and be able to state the underlying assumptions;
- understand how and when the central limit theorem might apply to this test;
- be able to apply this test given either raw data or summary statistics for the two samples;
- be able to distinguish between one and two tail tests and state the appropriate null and alternative hypotheses;
- be able to use either critical values or p-values to determine the result of the test;
- be able to state the result of the hypothesis test in the context of the question;
- be able to explain the relevance of the significance level to the result of the test;

TEACHING POINTS

This sub-unit should follow relatively easily from the previous one as the process and working are much the same and the teaching points made above are still valid here. However, you should make students aware of the differences in terms of the necessary conditions for the test and that when population variances are unknown they must use the sample variances s_x^2 and s_y^2 .

Stress that for unknown variances this test is only valid when sample sizes are large and that if the underlying distributions are not Normal then the central limit theorem is required to justify the use of a Z distribution in this test.

Students should be aware that they can use
$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} \sim N(0, 1).$$

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

As the range of tests available to students increases it is worth giving students a mixed batch of questions and asking them to state which hypothesis test is appropriate and why. They could also make a summary of the test statistics in each case, with any assumptions.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Not testing for equality of means: ‘The different style for the hypotheses caused some problems to candidates who liked to have the null hypothesis as an equality and did not realise the problem which arises from non-exhaustive sets of values in the hypotheses.’

UNIT 4: Goodness of fit and contingency tables Chi-squared tests (4.1) (4.2)

Teaching time
12 hours

[Return to Overview](#)

SPECIFICATION REFERENCES

4.1 The null and alternative hypotheses. The use of $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ as an approximate χ^2 statistic

4.2 Degrees of freedom

PRIOR KNOWLEDGE

- Hypothesis testing for the mean of a Normal distribution (See Unit 3)

IAL Mathematics – Statistics 1 content

- Discrete uniform distributions (See S1 SoW Unit 4)
- Normal distribution (See S2 SoW Unit 5)

IAL Mathematics – Statistics 2 content

- Binomial and Poisson distributions (See S2 SoW Unit 1)
- Continuous uniform distribution (See S2 SoW Unit 3)
- Statistical hypothesis testing (See S2 SoW Unit 4)

KEYWORDS

Unbiased, biased, hypothesis, null hypothesis, alternative hypothesis, significance level, critical value, observed frequencies, expected frequencies, goodness of fit, contingency tables, rows, columns, degrees of freedom, constraint, parameter, cell, p-value, discrete uniform distribution, binomial distribution, Poisson distribution, Normal distribution.

OBJECTIVES

By the end of the sub-unit, students should:

- know the process of and be able to use a goodness of fit test;
- know how to find the number of degrees of freedom of the expected values, including when one or more parameters are estimated from the data;
- be able to use $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ as an approximate χ^2 statistic;
- be able to apply goodness of fit tests to include the discrete uniform, binomial, Poisson, Normal and continuous uniform distributions;
- be able to show frequencies by means of a contingency table;
- know how to obtain p-values from calculators;
- be able to use tables to find critical values.

TEACHING POINTS

There are two main types of chi-squared tests here: a ‘goodness of fit’ test and a ‘contingency table’ test.

Both types use $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ as an approximate χ^2 statistic.

A level Further Mathematics: Statistics 3

In a ‘goodness of fit’ test, we are effectively carrying out a hypothesis test to determine whether a particular distribution is a suitable model (‘good fit’) for some observed data. The distributions we are testing for include the discrete uniform, binomial, Poisson, Normal and continuous uniform distributions.

The null hypothesis is that the observed distribution *does not differ* from a theoretical one (assumption of no difference), i.e. the theoretical distribution is a suitable model, and the alternative hypothesis is that the observed distribution *does differ* from the theoretical one, i.e. the theoretical distribution is not a suitable model.

Students should be taught to use the distribution they are testing for to work out the expected frequencies, and compare them with the observed frequencies using
$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}.$$

If any of the expected frequencies are less than 5, then combine frequencies in the data table until the expected frequency is at least 5.

The degrees of freedom refers to how many free choices can be made in the allocation of expected frequencies and is calculated using:

number of degrees of freedom = number (n) of cells (after any combining) – number of constraints.

This is usually equal to $n - 1$, unless testing a binomial model in which p has to be estimated from the observed data, or testing a Poisson model in which λ is estimated from the observed data. In these cases v is equal to $n - 2$. The Normal distribution has two parameters, μ and σ^2 . If only one of these has to be estimated from the observed data then v is equal to $n - 2$, but if both have to be estimated from the observed data then v is equal to $n - 3$.

The test is always one-tailed. The value of X^2 is the test statistic, and the value of χ^2 is a distribution.

The value of X^2 is compared to the value of χ^2 found by looking at the table of percentage points of the χ^2 distribution with v equal to the number of degrees of freedom of the expected values.

To be significant, the value of X^2 must be greater than the value of χ^2 . In this case we would reject the null hypothesis. The conclusion must be put in the context of the problem.

Students need to know the conditions under which each of the distributions being tested for are likely to arise. These must be stated in the context of the question.

We carry out a ‘contingency table’ test, to find out whether two particular attributes (or factors) of a population are associated. We display the frequency distribution of the two variables using a *contingency table*. The distribution of one variable is shown in rows and the other in columns.

For contingency tables, Expected frequency = $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$ and the degrees of freedom for an $h \times k$ table is $v = (h - 1)(k - 1)$.

For a contingency test, the null hypothesis is that there is *no* association between the variables (or that the variables are independent) and the alternative hypothesis is that there *is* an association (or the variables are dependent).

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Hypotheses need to be stated in the context of the question and simply stating ‘no association’ for the null hypothesis is not sufficient as the variables under consideration must be mentioned. For a contingency test, the null hypothesis is that there is no association, but in context, it would read, for example ‘Cholesterol is independent of intake of saturated fats’. Hypotheses can be phrased in terms of ‘association’ or ‘independence’.

The conditions under which a distribution is valid must also be stated in context; specific assumptions are required and simply listing all the requirements for a binomial distribution will not earn the marks.

A level Further Mathematics: Statistics 3

Although the degrees of freedom for an $h \times k$ contingency table is given by $\nu = (h - 1)(k - 1)$, combining cells still needs to be taken into account: ‘Most students were able to explain the calculation of the degrees of freedom in part (f), but too many thought that the pooling of two columns would simply lead to the loss of 1 degree of freedom from the original number.’

Encourage students to use non-rounded values in all calculations; premature rounding can result in a costly loss of marks in exam situations.

NOTES

Lengthy calculations of $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ will not be required.

Students are encouraged to obtain p-values from their calculator and use tables to find the critical values.

UNIT 5: Regression and correlation

[Return to Overview](#)

SPECIFICATION REFERENCES

- 5.1 Spearman's rank correlation coefficient, its use, interpretation and limitations
- 5.2 Testing the hypothesis that a correlation is zero

PRIOR KNOWLEDGE

- [Hypothesis testing \(See Units 3 & 4\)](#)
[IAL Mathematics - Statistics 1 content](#)
- Correlation, coding (See S1 SoW Unit 3)
[IAL Mathematics - Statistics 2 content](#)
- Hypothesis testing (See S2 SoW Unit 4)

KEYWORDS

Bivariate distribution, association, linear, scatter diagram, correlation, positive correlation, negative correlation, no correlation, mean, quadrant, product moment correlation coefficient (PMCC), summation, outlier, coding, rank, tie, Spearman's rank correlation coefficient, population, sample, hypothesis test, population parameter, sample statistic, test statistic, inference, null hypothesis, alternative hypothesis, significance level, one-tailed test, two-tailed test, critical region, boundary value, critical value, acceptance region, significant.

5a. Spearman's rank correlation coefficient (5.1)

Teaching time

5 hours

OBJECTIVES

By the end of the sub-unit, students should:

- be able to calculate Spearman's rank correlation coefficient using $r_s = 1 - \frac{6 \sum d^2}{n(n^2-1)}$;
- know how to use and interpret Spearman's rank correlation coefficient, and know its limitations;
- understand how to deal with ties.

TEACHING POINTS

Recap scatter diagrams and the terminology to describe correlation. Also recap the product moment correlation coefficient (PMCC), denoted by r , and that $|r| \leq 1$. If $r = \pm 1$ then the data points lie on a perfect straight line on a graph.

The PMCC is used when it is reasonable to assume that the data comes from an underlying bivariate Normal distribution. An elliptically shaped distribution on a scatter diagram is one possible indication. Spearman's rank correlation coefficient is denoted by r_s and is used when the data is ranked in some sort of order or where there is reason to believe that the conditions necessary for the PMCC may not be met. Spearman's rank correlation does not assume the data is linear.

Spearman's rank correlation coefficient is derived from the product moment correlation coefficient and is interpreted as follows:

- $r_s = +1$ means that the rankings are in perfect agreement
- $r_s = -1$ means that the rankings are in exact reverse order
- $r_s = 0$ means that there is no correlation between the rankings

The closer r_s is to zero, the weaker the association between the ranks.

OPPORTUNITIES FOR REASONING/PROBLEM SOLVING

Numerical questions involving ties will not be set. However, some understanding of how to deal with ties will be expected. When there is a tie between rankings (i.e. two or more individuals have the same rank), each individual is assigned a rank equal to the mean of the ranks for the tied values. No change is made in the ranks of the other values. We then proceed by finding the PMCC of the ranks.

Although a proof is not required, a nice piece of extension work is to derive the formula for the Spearman's rank correlation coefficient from the formula for the PMCC. This provides an understanding of why the formula $r_s = 1 - \frac{6 \sum d^2}{n(n^2-1)}$ is not valid when the ranks are tied.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Care must be taken in ranking the data and students must not forget the '1 - ' in the formula.

NOTES

When the raw data is provided students should be encouraged to use their calculators to find the correlation coefficient. They will also need to know how to apply the formulae when only summary data is given.

5b. Hypothesis test for zero correlation (5.2)**Teaching time**
6 hours**OBJECTIVES**

By the end of the sub-unit, students should:

- be able to test the hypothesis that a correlation is zero using Spearman's rank correlation;
- be able to test the hypothesis that a correlation is zero using the product moment correlation coefficient.

TEACHING POINTS

We are testing for zero correlation, so the null hypothesis will always be $H_0: \rho = 0$ or $H_0: \rho_s = 0$.

The alternative hypothesis could be

$H_1: \rho \neq 0$ (the correlation is either positive or negative; two-tailed test)

$H_1: \rho > 0$ (positive correlation; one-tailed test)

$H_1: \rho < 0$ (negative correlation; one-tailed test).

We calculate the value of r or r_s , and then look up critical values for the level of significance. Compare the absolute value of r to the positive critical value. If $|r|$ is greater than the critical value, then the result is significant and we reject H_0 . For the PMCC, in order to have significant positive linear correlation, the original value of r must be positive, and in order to have significant negative correlation, the original value of r must be negative.

OPPORTUNITIES FOR PROBLEM SOLVING/MODELLING

This is a good opportunity here to bring together the S1 and S3 content relating to regression and correlation.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Notation and stating a conclusion are the most common errors. Students need to be aware that hypotheses should always be stated in terms of ρ_s or ρ and test a null hypothesis that ρ_s or $\rho = 0$.

Take care with critical values and conclusions: 'A number of students did not ensure that their critical value matched their alternative hypothesis and of course there were the usual crop of conclusions that did not include a reference to the context.'

Some students muddle up one and two-tailed tests. The test can be one-tailed (for a claim of positive/negative correlation), or two-tailed (for a claim of *either* positive or negative correlation).

Students must take care to use the correct table when looking up critical values for the PMCC or Spearman's rank.

NOTES

Students need to be able to use tables for Spearman's and product moment correlation coefficients.