2.6 Further Statistics: Hypothesis Testing and Chi-Squared Depth: Test Pack

EDEXCEL 9FM0 · Calculator allowed · about 65 minutes

Total Marks

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

Key facts: chi-squared tests, in more depth

Original content written for Revision Library.

A chi-squared test compares observed frequencies O with the expected frequencies E predicted by a model, using X^2 = sum of (O-E)^2/E over all classes or cells, after any pooling. For a goodness-of-fit test to a fully specified distribution, df = (number of classes) - 1; if a parameter is estimated from the same sample (for example a Poisson mean, a binomial p, or a geometric p), one further degree of freedom is lost for each parameter estimated, because the estimation process artificially improves the apparent fit. For continuous data grouped into classes, expected frequencies are found from probabilities calculated using the standard Normal distribution, via z = (x - mu)/sigma and the function Phi. For an r by c contingency table testing independence between two categorical variables, E = (row total * column total)/grand total (this follows directly from the multiplication rule P(row and column) = P(row) * P(column) applied to the sample), and df = (r-1)(c-1). Classes or cells with an expected frequency below 5 should be pooled with a neighbouring class or column before the test statistic is calculated, and the degrees of freedom adjusted accordingly. Yates' continuity correction, X^2 = sum of (|O-E| - 0.5)^2/E, is applied only when df = 1 (every 2 by 2 contingency table), since this is the only case where the discrete test statistic is poorly approximated by the continuous chi-squared distribution. The signed standardised residual for a cell, e = (O-E)/sqrt(E), shows the direction as well as the size of a departure from the model, and the sum of its squares reproduces X^2. Because X^2 is never negative, only large (upper-tail) values ever provide evidence against H0; the calculated statistic is compared with the critical value from chi-squared tables at the chosen significance level and the stated degrees of freedom, and H0 is rejected if the statistic exceeds this value. As with any hypothesis test, the significance level equals P(Type I error), reducing it increases P(Type II error) for a fixed sample size, and increasing the sample size generally increases the power of the test; a chi-squared statistic also scales with sample size, so a very large sample can make even a small, unimportant deviation from a model 'statistically significant'. Combining ('pooling') data from distinct groups before testing can also mask or even reverse a genuine association found within each group, an effect known as Simpson's paradox.

© Revision Library | revisionlibrary.co.uk

1 For each scenario below, state (i) whether a chi-squared goodness-of-fit test or a chi-squared test for association is required, and (ii) the number of degrees of freedom that should be used, giving a brief reason. You are not required to state hypotheses or carry out any calculations.

(a) A tutor at a sixth-form college in Preston rolls what is claimed to be a fair four-sided die 100 times, to test whether each of the four scores is equally likely. **(2)**

(b) A biologist counts the number of ladybirds, X, found on each of 60 randomly chosen leaves in a garden, and wants to test whether X can be modelled by a Poisson distribution, with the mean estimated from the sample. After combining categories with a small expected frequency, 5 categories remain in total. **(2)**

(c) A researcher cross-tabulates 150 sixth-formers by whether they are currently studying a language A-level (yes or no) and whether they intend to take a gap year before university (yes or no). **(2)**

(d) A supermarket records X, the number of attempts a customer needs before their loyalty card scans successfully at a new self-checkout scanner, for a random sample of 90 customers, in order to test whether X can be modelled by a geometric distribution, with p estimated from the sample. After combining the two largest categories, 5 categories remain in total. **(2)**

(Total for Question 1 is 8 marks)

© Revision Library | revisionlibrary.co.uk

2 This question is about the general properties of chi-squared tests.

(a) Explain why a chi-squared test (whether for goodness of fit or for association) is always carried out using only the upper tail of the chi-squared distribution, even though the underlying question of whether a model fits well could be thought of as two-sided. **(2)**

(b) A shopkeeper wants to test specifically whether a six-sided die is biased towards rolling a six more often than expected, rather than just 'biased' in some unspecified way. Explain why a chi-squared goodness-of-fit test would not be the most suitable test for this specific question, and suggest a more suitable alternative approach. **(2)**

(c) Explain, with reference to the chi-squared formula, why every expected frequency used in the test must be comfortably greater than zero. **(2)**

(Total for Question 2 is 6 marks)

---

3 The random variable $X \sim B(10, p)$. A test of H0: $p = 0.3$ against H1: $p > 0.3$ is carried out at the 5% significance level.

(a) Find the critical region for the test. **(2)**

(b) State the probability of a Type I error, to 3 significant figures. **(2)**

(c) Find the probability of a Type II error when the true value of p is 0.5, to 3 significant figures. **(2)**

(Total for Question 3 is 6 marks)

© Revision Library | revisionlibrary.co.uk

4 A supermarket's self-checkout has a loyalty card scanner that sometimes needs more than one attempt before it reads a customer's card successfully. For a random sample of 100 customers, X is the number of attempts needed up to and including the first successful scan. It is suggested that X can be modelled by a geometric distribution, $X \sim Geo(p)$, where p is unknown and is to be estimated from the sample. The results are shown below.

Number of attempts (X): 1, 2, 3, 4, 5 or more.
Frequency: 42, 26, 16, 9, 7.

Attempts	Frequency
1	42
2	26
3	16
4	9
5 or more	7

(a) Show that the mean number of attempts for this sample is 2.13, treating the open category '5 or more' as if it were exactly 5. **(2)**

(b) Hence show that an estimate for p is 0.4695 (to 4 decimal places), using the fact that $E(X) = 1/p$ for a geometric distribution. **(1)**

(c) State suitable null and alternative hypotheses for this goodness-of-fit test. **(1)**

(d) Using $p = 0.4695$, calculate $P(X=1)$, $P(X=2)$, $P(X=3)$, $P(X=4)$ and $P(X \geq 5)$, giving each to 4 decimal places. **(3)**

(e) Calculate the expected frequency for each category, to 2 decimal places. **(2)**

(f) State the number of degrees of freedom for this test, with a reason. **(1)**

(g) Calculate the value of the chi-squared test statistic. **(2)**

(h) Using a 5% significance level and the critical value $\text{chi-squared}_3(0.05) = 7.815$, test whether Geo(0.4695) is a suitable model for the number of scan attempts, stating your conclusion in context. **(2)**

(Total for Question 4 is 14 marks)

© Revision Library | revisionlibrary.co.uk

5 A chain of driving instructors wants to see whether taking additional theory-based hazard perception sessions (alongside standard lessons) is associated with passing the driving test first time. Because overall pass rates differ a lot between branches for reasons unconnected with the hazard sessions (for example, differing local test centres), data are collected and considered separately at two branches.

Branch 1: took sessions - passed 54, failed 66, row total 120. Did not take sessions - passed 10, failed 20, row total 30.
Branch 2: took sessions - passed 24, failed 6, row total 30. Did not take sessions - passed 72, failed 48, row total 120.

(a) Calculate the pass rate (as a percentage, to 1 decimal place) for students who took the hazard sessions and for those who did not, separately for each branch. Comment on the direction of the association in each branch. **(3)**

(b) Calculate the corresponding pass rates for the COMBINED data (both branches together), and comment on how this compares with your answer to part (a). **(3)**

(c) Carry out a chi-squared test for association at the 5% significance level on the COMBINED data, stating your conclusion in context. (Critical value: chi-squared$_1$(0.05) = 3.841.) **(4)**

(d) Carry out a chi-squared test for association at the 5% significance level on Branch 2's data alone, stating your conclusion in context. (Critical value: chi-squared$_1$(0.05) = 3.841.) **(3)**

(e) Explain, with reference to your answers above, why simply combining ('pooling') data from different branches before testing for association can be misleading, and suggest what the company should do instead when analysing this kind of data. **(3)**

(Total for Question 5 is 16 marks)

© Revision Library | revisionlibrary.co.uk

Mark scheme · 2.6 Further Statistics: Hypothesis Testing and Chi-Squared Depth: Test Pack

Question 1

(a) a chi-squared goodness-of-fit test for a uniform distribution is needed; this is fully specified, with no parameters estimated from the data **B1**

(a) df = 4 - 1 = 3, cao **B1**

(a) Answer: Goodness-of-fit test (uniform); df = 3.

(b) a chi-squared goodness-of-fit test for a Poisson distribution is needed, with 1 parameter (the mean) estimated from the same data **B1**

(b) df = 5 - 1 - 1 = 3, cao **B1**

(b) Answer: Goodness-of-fit test (Poisson, estimated mean); df = 3.

(c) a chi-squared test for association (independence), using a 2 by 2 contingency table, is needed **B1**

(c) df = (2-1)(2-1) = 1, cao **B1**

(c) Answer: Test for association (2 by 2 contingency table); df = 1.

(d) a chi-squared goodness-of-fit test for a geometric distribution is needed, with 1 parameter (p) estimated from the same data **B1**

(d) df = 5 - 1 - 1 = 3, cao **B1**

(d) Answer: Goodness-of-fit test (geometric, estimated p); df = 3.

Question 2

(a) the test statistic sum of $(O-E)^2/E$ is always non-negative, and only large values give evidence of a poor fit or an association, arising from big discrepancies between observed and expected frequencies **B1**

(a) small values simply indicate the data fit the model even better than typically expected, which is not evidence against H0, so only the upper (right-hand) tail is used to define the rejection region **B1**

(a) Answer: Only large values of the non-negative test statistic can indicate a poor fit or an association, so only the upper tail can ever provide evidence against H0.

(b) a chi-squared goodness-of-fit test only tests whether the overall pattern of frequencies differs from that expected under a fair die; it cannot test a specific one-sided (directional) alternative hypothesis such as 'P(six) is greater than 1/6' **B1**

(b) a more suitable approach is a hypothesis test on the single probability P(rolling a six) directly, e.g. modelling the number of sixes in n rolls as B(n, 1/6) and carrying out a one-tailed test for an increase in p **B1**

(b) Answer: Chi-squared GOF cannot test a specific direction of bias; a one-tailed binomial test on P(six) directly would be more suitable.

(c) each category contributes a term $(O-E)^2/E$ to the test statistic, which is undefined if E = 0 (division by zero) and becomes extremely large and unstable as E approaches 0 **B1**

(c) this is why categories with a small expected frequency (conventionally E < 5) must be combined with a neighbouring category before the test statistic is calculated **B1**

(c) Answer: A term $(O-E)^2/E$ is undefined at E=0 and unstable for small E; this is why low-E categories are pooled before testing.

Question 3

(a) uses B(10, 0.3) cumulative probabilities **M1**

(a) critical region $X \geq 6$ **A1**

(a) Answer: critical region $X \geq 6$

(b) P(X in critical region | p = 0.3) **M1**

(b) 0.0473 **A1**

(b) Answer: 0.0473

(c) P(X not in critical region | p = 0.5) **M1**

(c) 0.623 **A1**

(c) Answer: 0.623

© Revision Library | revisionlibrary.co.uk

Question 4

(a) sum of x*f = 1*42 + 2*26 + 3*16 + 4*9 + 5*7 = 213 **M1**

(a) cso: mean = 213/100 = 2.13 **A1**

(a) Answer: Mean = 2.13

(b) cao: p_{hat} = 1/2.13 = 0.4695 (4 dp) **B1**

(b) Answer: p_{hat} = 0.4695

(c) H0: X can be modelled by Geo(0.4695); H1: X cannot be modelled by Geo(0.4695) **B1**

(c) Answer: H0: the number of attempts follows Geo(0.4695). H1: it does not.

(d) uses $P(X=x) = q^{x-1} * p$ with q = 1 - p = 0.5305 **M1**

(d) uses $P(X \geq 5) = 1 - [P(X=1) + P(X=2) + P(X=3) + P(X=4)]$ **M1**

(d) all five values correct: 0.4695, 0.2491, 0.1321, 0.0701, 0.0792 **A1**

(d) Answer: 0.4695, 0.2491, 0.1321, 0.0701, 0.0792

(e) multiplies each probability by 100 **M1**

(e) all correct: 46.95, 24.91, 13.21, 7.01, 7.92 **A1**

(e) Answer: 46.95, 24.91, 13.21, 7.01, 7.92 (sum = 100.00)

(f) df = 5 - 1 - 1 = 3 (5 categories, minus 1 for the total, minus 1 because p was estimated from the sample) **B1**

(f) Answer: df = 3

(g) uses X^2 = sum of $(O-E)^2/E$ with the correct O and E values **M1**

(g) awrt 1.83 **A1**

(g) Answer: X^2 = awrt 1.83

(h) 1.83 < 7.815 **A1**

(h) ft: do not reject H0; no significant evidence against modelling the number of scan attempts by a geometric distribution with p = 0.4695 **A1**

(h) Answer: 1.83 < 7.815, do not reject H0; the geometric model is a suitable fit.

Question 5

(a) correct method for at least one percentage **M1**

(a) all four percentages correct: Branch 1: 45.0% and 33.3%; Branch 2: 80.0% and 60.0% **A1**

(a) comment: in both branches, students who took the extra hazard sessions had a higher pass rate than those who did not **A1**

(a) Answer: Branch 1: 45.0% (sessions) vs 33.3% (no sessions). Branch 2: 80.0% (sessions) vs 60.0% (no sessions). Sessions associated with a higher pass rate in both branches.

(b) combines the two branches' figures first: sessions 78 passed out of 150; no sessions 82 passed out of 150 **M1**

(b) combined percentages correct: 52.0% (sessions) and 54.7% (no sessions) **A1**

(b) comment: remarkably, when the branches are combined, students who did NOT take the extra sessions now have the higher pass rate; the direction of the association has reversed compared with each individual branch **A1**

(b) Answer: Combined: 52.0% (sessions) vs 54.7% (no sessions) - the direction of association reverses compared with each branch separately.

(c) expected frequencies: 80, 70, 80, 70 (e.g. E(sessions, pass) = 150*160/300 = 80) **M1**

(c) uses X^2 = sum of $(O-E)^2/E$ **M1**

(c) cao: X^2 = awrt 0.214 **A1**

(c) 0.214 < 3.841, do not reject H0: no significant evidence of an association between taking the sessions and passing, in the combined data **A1**

(c) Answer: X^2 = 0.214 < 3.841; not significant, no evidence of association in the combined data.

(d) expected frequencies: 19.2, 10.8, 76.8, 43.2 **M1**

(d) cao: X^2 = awrt 4.17 **A1**

(d) 4.17 > 3.841, reject H0: significant evidence of an association between taking the sessions and passing, at Branch 2 **A1**

(d) Answer: X^2 = 4.17 > 3.841; significant evidence of association at Branch 2.

(e) pooling can hide, distort, or even reverse a genuine association that is present within each branch, because the branches differ substantially both in their overall pass rates and in how many students take the extra sessions (branch acts as a lurking/confounding variable). This effect is known as Simpson's paradox **B1**

(e) different branches can also show different strengths (or even different conclusions) from the same test, as seen here: Branch 2 shows a significant association, while the combined data does not **B1**

(e) the company should analyse (and test) each branch separately rather than relying only on the combined totals; if a single overall comparison is genuinely needed, a method that accounts for branch (such as testing within each branch and combining the evidence, or matching/stratifying by branch) should be used rather than a naive pooled total **B1**

(e) Answer: Pooling across branches can mask or reverse a real association (Simpson's paradox) because branch is a confounding variable; the company should test each branch separately, or use a method that adjusts for branch, rather than pooling the raw totals.

© Revision Library | revisionlibrary.co.uk