9.4 Goodness of fit and contingency tables

Name: _____ Class: _____ Date: _____ **Total: 27 marks**

KEY WORDS expected frequency • degrees of freedom

Goodness of fit and contingency tables

Objective

Compute expected frequencies for fitted models and contingency tables, combine low cells, set the degrees of freedom (subtracting estimated parameters), and conclude in context (objectives 4.1–4.2).

1 Key knowledge and worked examples

Two shapes. Fit: $E_i = np_i$ from the model, $X^2 = \sum \frac{(O-E)^2}{E}$, df = cells − 1 − (parameters estimated). Contingency: $E = \frac{\text{row}\times\text{col}}{\text{total}}$, df $= (r-1)(c-1)$. Combine cells with $E < 5$ first.

Worked example (WST03 June 2025 Q3, Q5). (i) The diesel/month 1–6 cell of the 3×3 engine table: $E = \frac{35\times 36}{250} = 5.04$, contributing $\frac{(10-5.04)^2}{5.04} = 4.88$; with 7.444 given for the rest, $X^2 = 12.325 > 9.488$ on 4 df – association. (ii) Litters: $\hat{p} = 0.488$; the 0-male class has $E = 8.79$ and contributes $\frac{(2-8.79)^2}{8.79} = 5.25$, so $X^2 = 10.95$ on $6 - 1 - 1 = 4$ df > 9.488 – the binomial is not suitable.

2 Practice

2.1 For a 4×3 contingency table with grand total 240: state the degrees of freedom and write the formula for one expected frequency. [2]

2.2 A fair-die model over 6 cells with $n = 120$: write down all six expected frequencies. [2]

2.3 A fit loses one parameter to estimation and has 7 cells; three tail cells have expected frequencies 1.2, 2.8, 4.6. What combining is needed, and what is the resulting df? [3]

3 Exam-style questions

3.1 *(Edexcel IAL WST03 June 2025 Q3.)* A garage sold 250 new cars with petrol, diesel or electric engines and recorded when each car first experienced an engine problem in year 1 (table as in the paper: petrol 12/25/95, diesel 10/8/17, electric 14/10/59). (a) Calculate the expected frequency for a diesel car first experiencing a problem in months 1–6. (b) Given that the value of $\sum \frac{(O-E)^2}{E}$ for the remaining 8 cells is 7.444 (3 dp), test at the 5% level whether there is an association, stating hypotheses, the test statistic, the degrees of freedom and the critical value. [9]

3.2 *(Edexcel IAL WST03 June 2025 Q5.)* The numbers of male puppies in 250 litters of size 5 are 2, 40, 90, 85, 30, 3 for 0 to 5 males. Jeff believes a binomial distribution is suitable; the expected frequencies for 1, 2, 3, 4, 5 males are 41.92, 79.91, 76.16, 36.30, 6.92 (2 dp). (a) Estimate the proportion of male puppies. (b) Find the expected frequency r for 0 males. (c) Given $\sum \frac{(O-E)^2}{E}$ for the other five cells is 5.70, test at the 5% level whether the binomial is suitable. [11]

Solutions

2.1 df $= (4-1)(3-1) = 6$; $E = \frac{\text{row total} \times \text{column total}}{240}$.

2.2 All six $E = 120 \times \frac{1}{6} = 20$.

2.3 Combine the three tail cells (all $E < 5$) into one ($E = 8.6$): 5 cells, df $= 5-1-1 = 3$.

3.1 (a) $E = \dfrac{35 \times 36}{250} = 5.04$. (b) H_0: no association between the time of the first problem and the engine type; H_1: there is an association. The diesel cell contributes $\frac{(10-5.04)^2}{5.04} = 4.881$, so $X^2 = 7.444 + 4.881 = 12.325$ on df $= (3-1)(3-1) = 4$; the 5% critical value is 9.488. Since $12.325 > 9.488$, reject H_0 – evidence of an association between when a car first experiences an engine problem and the type of engine.

3.2 (a) $\hat{p} = \dfrac{0(2) + 1(40) + 2(90) + 3(85) + 4(30) + 5(3)}{250 \times 5} = \dfrac{610}{1250} = 0.488$. (b) $r = 250 - (41.92+79.91+76.16+36.30+6.92) = 8.79$. (c) H_0: the binomial distribution is a suitable model; H_1: it is not. The r-class contributes $\frac{(2-8.79)^2}{8.79} = 5.25$, so $X^2 = 5.70+5.25 = 10.95$. No cells need combining here (every $E \geq 5$); the degrees of freedom are $6-1-1 = 4$, losing one for the parameter $\hat{p}$ estimated from the data. Critical value 9.488: $10.95 > 9.488$, reject – significant evidence that the binomial distribution is NOT a suitable model.

Local review guidance: expected frequencies from the correct rule (model probability or row×col/total); df with the estimated parameter subtracted; rejection decided against the table value and concluded in context. This is authored classroom guidance.