Contingency Tables

Starter

1. **(Review of last lesson)**
The accountant of a company monitors the number of items produced each month together with the total cost of production. The data collected for a random sample of 12 months is:

Number of items (x) (1000s)	21	39	48	24	72	75	15	35	62	81	12	56
Production cost (y) (£1000)	40	58	67	45	89	96	37	53	83	102	35	75

(a) Use your calculator to find an equation for the regression line of y on x.
(b) The selling price of each item is $£2.20$. Find the level of output at which income and total costs are equal. Interpret this value.

2. The following table shows the grades achieved by random students in two schools.

Observed frequencies		Grade A	Grade B	Grade C	Totals
School	X	18	12	20	50
	Y	26	12	32	70
Totals		**44**	**24**	**52**	**120**

(a) From the table, write down:
(i) the probability of choosing a student from school X,
(ii) the probability of choosing a student who got a grade A.
(b) If the two probabilities from (a) are independent (i.e. the factors school and grade are independent), calculate the probability of choosing a students from school X who achieved an A grade.
(c) Using your answer to (b), calculate how many students we would expect to get a grade A in school X if 'school' and 'grade' were independent.
(d) Draw a new version of the table and calculate the expected values for the values in blue

Notes

The table above is an example of a contingency table. Two variables are being compared, in this case school vs. grade, but different to linear regression, the variables are classified according to several sets of attributes.

We start off with a table of observed frequencies and from this we generate a table of expected frequencies ***given that the variables are independent***.

Once we have the observed and expected frequencies, we need a mathematical way of deciding whether there is evidence to suggest the two variables are ***independent*** or ***not independent***. When the factors are ***not independent***, they are said to be associated. We do not write "there is evidence to suggest the two variables are dependent".

Obviously the greater difference between the observed and expected frequencies, the more likely the variables are not independent, but how great is too great?

Degrees of freedom, ν

Consider the table above. Given that the totals in the last row and column are fixed how many values can be entered into the blue cells before the cell values become constrained?

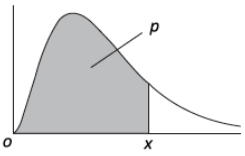
Once AX is filled, AY is constrained — so AX is one degree of freedom
Once BX is filled, BY is constrained — so BX is another degree of freedom
With AX and BX both filled, CX is constrained — no new degree of freedom
With CX filled by constraint, CY is constrained — no new degree of freedom
So in total the table has 2 degrees of freedom.

For an m by n contingency table, the degrees of freedom, ν ("nu") is given by $(m-1)(n-1)$.

$$\nu = (m-1)(n-1)$$

Tables and significance level

When we carry out a hypothesis test at the, say, $5\,\%$ level it means that the shaded area is $95\,\%$ and it is under that column where we find the critical value. If the test value is to the right of *x* in the diagram then it lies in the critical region and we reject the null hypothesis, H_0.



Carrying out a chi-squared, χ^2_ν, test

A chi-squared test is the method we use to consider whether two variables expressed in contingency tables are ***independent*** or ***not independent***.

1. State the null and alternative hypotheses:
   H_0 : The variables ____ and ____ are independent.
   H_1 : The variables ____ and ____ are not independent (i.e. there is an association)
2. Calculate expected frequencies by Expected frequency $= \dfrac{\text{row total} \times \text{column total}}{\text{grand total}}$
3. Calculate $\chi^2_{calc} = \sum \dfrac{(O_i - E_i)^2}{E_i}$ ***this is the test statistic***
4. Write down the number of degrees of freedom using $\nu = (m-1)(n-1)$
5. Decide on the level of significance of the test (e.g. $5\,\%$)
6. Find the the critical value for χ^2_ν for the given significance level from the tables
7. Compare χ^2_{calc} with χ^2_ν

$\chi^2_{calc} < \chi^2_\nu$: do not reject H_0, there is evidence to suggest the variables are independent
$\chi^2_{calc} > \chi^2_\nu$: reject H_0, there is evidence to suggest the variables are not independent

E.g. 1 Carry out a chi-squared test at the $5\,\%$ level to decide whether there is a connection between school and grades achieved. State your null and hypothesis clearly.

Working: H_0 : The variables school and grade are independent.
H_1 : The variables school and grade are not independent

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{\left(18 - 18\frac{1}{3}\right)^2}{18\frac{1}{3}} + \frac{\left(12 - 10\right)^2}{10} + \frac{\left(20 - 21\frac{2}{3}\right)^2}{21\frac{2}{3}} + \frac{\left(26 - 25\frac{2}{3}\right)^2}{25\frac{2}{3}} + \frac{\left(12 - 14\right)^2}{14} + \frac{\left(32 - 30\frac{1}{3}\right)^2}{30\frac{1}{3}}$$

$$\chi^2_{calc} = 0.9165$$

Degrees of freedom, $\nu = 2$

The critical value at the $5\,\%$ level is $\chi^2_2(5\%) = 5.991$

Since $\chi^2_{calc} = 0.9165 < 5.991 = \chi^2_2(5\%)$, we do not reject H_0 :
There is evidence to suggest the variables are independent
i.e. the grade achieved by students is independent of school attended

E.g. 2 Seven different types of locality were studied to see if ownership, or non-ownership, of a television was or was not related to locality. $\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$ was found to be 13.1.

(a) Carry out a hypothesis test at the $5\,\%$ level to see if locality and ownership of TV are related.

(b) Would your conclusion differ if the significance level was $2.5\,\%$? Explain your answer.

Limitations of the χ^2 test and combining rows/columns

For the χ^2 test to be reliable, the ***expected*** frequencies must all be ***greater than or equal to 5*** i.e. $E_i \geq 5$. This is known as Cochran's Rule (William Cochran, Scottish statistician, 1909-1980). In situations where this is not the case, columns or rows need to be combined.

E.g. 3 A university requires all science students to study non-science subject in their first year. Here are the choices they made:

	French	Poetry	Russian	Sculpture
Male	2	8	15	10
Female	10	17	21	37

Test at the $1\,\%$ level whether choice of subject is independent of sex.

Working: The expected frequency for males who choose French is $\frac{12 \times 35}{120} = 3.5$.

Since the expected frequency is below 5 we must combine columns.
It makes sense to combine French and Russian as they are both languages.
The new observed frequencies are:

New observed frequencies	French/Russian	Poetry	Sculpture
Male	17	8	10
Female	31	17	37

The expected frequencies are:

Expected frequencies	French/Russian	Poetry	Sculpture
Male	14.00	7.29	13.71
Female	34.00	17.71	33.29

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(17-14.00)^2}{14.00} + \frac{(8-7.29)^2}{7.29} + \frac{(10-13.71)^2}{13.71} + \frac{(31-34.00)^2}{34.00} + \frac{(17-17.71)^2}{17.71} + \frac{(37-33.29)^2}{33.29}$$
$$\chi^2_{calc} = 2.422$$

H_0 : sex and choice of subject are independent
H_1 : sex and choice of subject are not independent

Degrees of freedom, $\nu = 2 \times 1 = 2$
The critical value at the 1% level is $\chi^2_2(1\%) = 9.210$
Since $\chi^2_{calc} = 2.422 < 9.210 = \chi^2_2(1\%)$, we do not reject H_0.
There is no evidence of a relationship between sex and choice of subject

Video: Contingency tables
Video: Chi-squared hypothesis tests
Video: Contingency tables example

Solutions to Starter and E.g.s

Exercise
p99 6A Qu 1i, 2-5, (6-7 red)

Summary
Carrying out a chi-squared, χ^2_ν, test
A chi-squared test is the method we use to consider whether two variables expressed in contingency tables are ***independent*** or ***not independent***.

1. State the null and alternative hypotheses:
   H_0 : The variables ____ and ____ are independent.
   H_1 : The variables ____ and ____ are not independent (i.e. there is an association)
2. Calculate expected frequencies by Expected frequency $= \dfrac{\text{row total} \times \text{column total}}{\text{grand total}}$
3. Calculate $\chi^2_{calc} = \sum \dfrac{(O_i - E_i)^2}{E_i}$ ***this is the test statistic***
4. Write down the number of degrees of freedom using $\nu = (m-1)(n-1)$
5. Decide on the level of significance of the test (e.g. $5\,\%$)
6. Find the the critical value for χ^2_ν for the given significance level from the tables
7. Compare χ^2_{calc} with χ^2_ν

$\chi^2_{calc} < \chi^2_\nu$: do not reject H_0, there is evidence to suggest the variables are independent
$\chi^2_{calc} > \chi^2_\nu$: reject H_0, there is evidence to suggest the variables are not independent