Block 3: Analysing two variables (and sometimes three)

4.2.1 Income differences – Statistical significance

[Draft only: 20 August 2013]

In exercise 3.1.4.1 Income differences work-through we produced the following table:

Q901a Sex of respondent * Q918b Gross inc of R (if working) [3 groups] Crosstabulation

		Q918b Gross income of R (if working) [3 groups]			Total
		<£6000 %	<£12000 %	£12000+ %	(n = 100%)
	Total	**30.1**	**38.0**	**31.9**	1560
Q901a Sex of respondent	Men	9.8	41.4	48.7	874
	Women	55.8	33.7	10.5	686
	Epsilon	**-46.0**	**+7.7**	**+38.3**	

. . in which the epsilons have a marked and very large shift from -46.0 through +7.7 to +38.3

The next phase of analysis is to see what happens to these percentages and epsilons when controlling for one or more test variables, but before that we need to ask ourselves whether the apparent differences are really there or whether they could have arisen by chance.

This is where statistical testing is needed. We start with the hypothesis that there is no relationship between sex and earnings, the null hypothesis. For this to be true, we need to know to what extent the observed distribution of differs from the expected distribution if sex and earnings were completely unrelated. So we need to compare the observed distribution with the expected distribution if sex and earnings were completely unrelated. This would be:

Q901a Sex of respondent * Q918b Gross income of R (if working) [3 groups] Crosstabulation

		Q918b Gross income of R (if working) [3 groups]			Total
		<£6000 %	<£12000 %	£12000+ %	(n = 100%)
	Total	**30.1**	**38.0**	**31.9**	1560
Q901a Sex of respondent	Men	**30.1**	**38.0**	**31.9**	874
	Women	**30.1**	**38.0**	**31.9**	686
	Epsilon	**0**	**0**	**0**	

. . in which the percentages for the men and women are the same as those for the whole sample and the epsilons are all 0.

To check the extent to which our observed distribution is significantly different from the expected distribution, we need to calculate the probability that the observed distribution has occurred purely by chance. To do this we need to work with the observed cell counts:

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	86	362	426
	Women	686	383	231	72

. . and compare them to the expected cell counts. How do we get these expected counts?

We get the expected count for men earning under £6000 by taking the column total 469 (number in sample earning under £6000), multiplying it by the row total 874 (number of men in the sample) and dividing it by the global total 1560 (number of cases in the whole sample):

469 x 874 / 1560 = 262.8

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	262.8		
	Women	686			

If the null hypothesis is true, we expect to get 262.8 men earning under £6000.

We get the expected figure for women earning under £6000 by taking the column total 469 (number in sample earning under £6000), multiplying it by the row total 686 (number of women in the sample) and dividing it by the global total 1560 (number of cases in the whole sample):

469 x 686 / 1560 = 206.2

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	262.8		
	Women	686	206.2		

and so on until we get to women earning £12000 or more:

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	262.8	332.2	279.0
	Women	686	206.2	260.8	219.0

Try it yourself by filling in cells in the table below from the original table of counts.

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874			
	Women	686			

We can now compare the expected counts: (E)

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	262.8	332.2	279.0
	Women	686	206.2	260.8	219.0

with the observed counts: (O)

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	86	362	426
	Women	686	383	231	72

. . calculate differences by subtracting the expected counts from the observed counts: (O – E)

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	-176.8	29.8	147.0
	Women	686	176.8	-29.8	-147.0

. . square all the differences (to get rid of negative numbers): $(O - E)^2$

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	31244.19	886.1308	21606.74
	Women	686	31244.19	886.1308	21606.74

. . divide each squared difference by the expected count: $\dfrac{(O - E)^2}{E}$

		Total	Q918b Gross income of R (if working) [3 groups]		
			<£6000	<£12000	£12000+
Total		1560	469	593	498
Q901a Sex of respondent	Men	874	118.91	2.67	77.44
	Women	686	118.91	2.67	77.44

. . and finally sum the resulting figures (Σ: Greek capital S which means add them all up)

$$\sum \frac{(O - E)^2}{E} = 452.57$$

We have just calculated a statistic called chi-square which can be used to test the null hypothesis that there is no relationship between sex and earnings.

By itself the chi-square statistic doesn't tell us anything: it depends on something called degrees of freedom and has to be checked against special tables. Degrees of freedom means the number of cells for which counts don't need to be calculated once other cells have been filled. Basically they are fixed once the remaining counts can be determined from the row or column totals and the already filled cells. In this table, once the counts in any 4 cells are known, the counts for the remaining two cells are automatically determined by the existing cells and the row and column totals.

Degrees of freedom for any table can be expressed as the number rows (r) minus 1 (r -1) multiplied by the number of columns (c) minus 1 (c -1) which in this table is:

$$(r - 1)(c - 1) = (2-1)(3-1) = 1 \times 2 = 2.$$

Chi-square is then checked against a table of critical values for various degrees of freedom to give you the probability (p) that the observed counts could have arisen by chance. Depending on the value of p, you can then decide whether to accept or reject the null hypothesis. Much social research rejects the null hypothesis when the p value falls below 0.05 (1 in 20) but some research uses the more rigorous criterion value of 0.001 (1 in 1000).

You don't have to do all the calculations for chi-square yourself: SPSS does it all for you, including calculating the p values.

crosstabs sex by incr3 /statistics chisq.

Q901a Sex of respondent * Q918b Gross income of R (if working) [3 groups] Crosstabulation

Count

		Q918b Gross income of R (if working) [3 groups]			Total
		<£6000	<£12000	<£12000+	
Q901a Sex of respondent	Men	86	362	426	874
	Women	383	231	72	686
Total		469	593	498	1560

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	**452.573[a]**	**2**	**.000**

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 206.24.

A chi-square value of 452.573 (the exact figure we calculated step-by-step above) with **2** degrees of freedom indicates a 0.000 probability of the observed distribution arising by chance The actual probability is less than 0.0005 (5 in 10,000) otherwise it would have displayed as 0.001 in the table.

We can therefore reject the null hypothesis.

For what it's worth, the original table from 3.1.4.1 Income differences work-through:

sex Q901a Sex of respondent * v1727 Q.918b Income group of respondent (if working) Crosstabulation

% within sex Q901a Sex of respondent

	Q.918b Income group of respondent (if working)														Total
	Under £2000	£2000 < £2999	£3000 < £3999	£4000 < £4999	£5000 < £5999	£6000 < £6999	£7000 < £7999	£8000 < £9999	£10000 < £11999	£12000 < £1999	£15000 < £17999	£18000 < £19999	£20000 < £24000	£24000 or more	
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	**n = 100%**
Total	**5.2**	**5.7**	**5.8**	**6.0**	**7.4**	**7.2**	**8.1**	**11.6**	**11.2**	**12.2**	**7.1**	**3.7**	**1.9**	**7.0**	**1560**
Men	0.3	0.8	0.9	2.4	5.4	5.3	8.7	13.4	14.1	17.4	10.9	5.7	3.2	11.6	**874**
Women	11.4	12.0	12.1	10.5	9.9	9.6	7.3	9.3	7.4	5.7	2.3	1.2	0.1	1.2	**686**
Epsilon	**-11.1**	**-11.2**	**-11.2**	**-8.1**	**-4.5**	**-4.3**	**+1.4**	**+4.1**	**+6.7**	**+11.7**	**+8.6**	**+4.5**	**+3.1**	**+10.4**	

. . has 13 degrees of freedom and chi-square is calculated as 523.436:

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	523.436[a]	13	.000

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 206.24.