

Block 3: Analysing two variables (and sometimes three)

3.1.1 Analysing two variables (Fifth form survey)

In a survey of fifth-formers at a comprehensive school in North London¹, one question consisted of a set of items measuring attitudes towards women. From this a crude scale of "Sexism" was derived ranging from 0 (No sexism) to 7 (Very high sexism). The results for all pupils looked like this:

All pupils	Code	Frequency	(%)
	0	6	5.3
	1	31	27.2
	2	26	22.8
	3	15	13.2
	4	10	8.8
	5	11	9.6
	6	10	8.8
	7	5	4.4
	-----		-----
	Total	114	100.0

Whilst this distribution is of interest in its own right, it isn't actually meaningful in research terms until we relate it to some other variable. For instance, one analysis we should like to see would be the distribution of "Sexism" separately for boys and girls. One way of doing this is to produce two separate frequency distributions which look like this:

Boys	Code	Frequency	(%)
	0	0	0.0
	1	4	8.2
	2	13	26.5
	3	9	18.4
	4	2	4.1
	5	8	16.3
	6	9	18.4
	7	4	8.2
	-----		-----
	Total	49	100.0

Girls	Code	Frequency	(%)
	0	5	10.0
	1	24	48.0
	2	10	20.0
	3	4	8.0
	4	5	10.0
	5	1	2.0
	6	1	2.0
	7	0	0.0
	-----		-----
	Total	50	100.0

The first thing to note is that there are no boys with a score of 0 and no girls with a score of 7. The

¹ Paul Ahmed, Harriet Cain and Alan Cook **Playground to Politics: a study of values and attitudes among fifth formers in a North London comprehensive school** Report on 2nd year project for BA Applied Social Studies (Social Research) Polytechnic of North London 1982

second is that the two distributions have a very different shape, and this is very easy to see when you compare them. The girls are all bunched at the low end, but the boys appear to have a bi-modal distribution with two bunches, one at the low and the other at the high end. These differences are disguised when the overall distribution is shown in the first example.

What we have just done above is to produce conditional frequencies as a transition from analysing one variable at a time to analysing two at a time. In fact if you put the two separate frequency tables side by side you would have what we call a **contingency table**.

For example:

		Sex of pupil		Row Total
		:Boys	Girls	
		: 1	: 2	:
Sexism	0	: 0	: 5	: 5
	1	: 4	: 24	: 28
	2	: 13	: 10	: 23
	3	: 9	: 4	: 13
	4	: 2	: 5	: 7
	5	: 8	: 1	: 9
	6	: 9	: 1	: 10
	7	: 4	: 0	: 4
Column Total		49	50	99

In this table, the **cells** are defined by the **sex** of the pupil (the two columns) and the score on the "**Sexism**" scale (the eight rows) which gives a total of 16 cells. In addition there is an extra **row total** column giving the total number of pupils obtaining each score, and an extra **column total** row giving the total number of pupils, together with a **grand total** for the number of pupils in the table as a whole. Although it is fairly easy to compare the distribution of scores in the table, because there are almost the same number of boys (49) as girls (50), it is normal practice to convert the figures to percentages in order to be able to compare more easily. This is particularly the case when the numbers in each group you want to compare are very different.

The table below has been converted to percentages and you should note that the base for percentaging has been given where necessary.

		Sex of pupil		Row Total
		:Boys	Girls	
		: %	: %	: %
Sexism	0	: 0.0	: 10.0	: 5.1
	1	: 8.2	: 48.0	: 28.3
	2	: 26.5	: 20.0	: 23.2
	3	: 18.4	: 8.0	: 13.1
	4	: 4.1	: 10.0	: 7.1
	5	: 16.3	: 2.0	: 9.1
	6	: 18.4	: 2.0	: 10.1
	7	: 8.2	: 0.0	: 4.0
N=100%		49	50	99

Although this table is now easier to read because the distributions have been standardised (by using percentages) it could be made even easier in two ways. First, the table can be condensed into fewer cells by grouping the scores on the "Sexism" scale; second, the percentages themselves can be simplified by getting rid of the decimal points. What is 0.1% of 50 cases anyway?

The very simplest table of all is one with two rows and two columns (the 2 x 2) so if we group the "Sexism" scores into "High" and "Low" (arbitrarily defining "Low" as 0-2 and "High" as 3-7) we get the following table. It also helps if we get rid of some of the lines.

		Sex of pupil		
		Boys	Girls	All
		----	-----	---
		%	%	%
Sexism	Low	35	78	57
	High	65	22	43
		====	====	==
N=100%		49	50	99

From this table we can see that boys are more sexist than girls in this particular fifth form (accepting for the moment our operational definitions and measuring instruments). However, by condensing the scores we have lost an important finding about the boys falling into two clearly defined groups, a finding which warrants further investigation. We can see that there is a difference, but how big is it, and how reliable are the data? What is the likelihood that this distribution could have arisen by chance? The latter question can be answered by using special techniques called significance tests, and these will be dealt with elsewhere in the course. One answer to the first question can be provided by looking at the difference between the percentage of boys and girls in each category. Thus in the "Low" group the difference is minus 43 percentage points (subtracting girls, 78 from boys, 35) and on the "high" group" it is plus 43 points (65 - 22). Conventionally, we choose one end of the scale as a **critierion** value (in this case, the "high" end) and state that boys are more likely to be sexist than girls.

If we want to attach a figure to this, we should use the **percentage point difference** of +43. This particular statistic is known formally as **EPSILON** (the Greek letter ϵ). It is important not only as a measure of the difference between groups, but also because, at a later stage of analysis using three or more variables, its value may change when additional variables are used as "**controls**", but this takes us out of the descriptive stage and into the explanatory.

We have already seen that conditional frequencies can be displayed for two or more groups in the form of contingency tables, and that for ease of comparison we convert raw data counts into a standardised form using percentages. We have also seen that if we have a dependent variable we must calculate the percentages to sum to 100% within the categories of the independent variable.

If there are a large number of categories to be tabulated, the tables can be reduced in size by grouping the values into fewer categories. Comparisons between groups can be made by subtracting the percentage in one group from the percentage in the other group having a specific value for the dependent variable, and then using this percentage difference (epsilon) as a summary measure.

In the fifth form survey, we calculated a measure of sexism, grouped it into "Low" and "High" and tabulated it by sex of pupil to give the following table:

		Sex of pupil		
		Boys	Girls	All
		----	-----	---
		%	%	%
Sexism	Low	35	78	57
	High	65	22	43
		====	====	====
N=100%		49	50	99

Taking the "High" end as the criterion value, we calculated an epsilon of +43 and concluded that boys were more sexist than girls.

[\[Back to Block 3 menu\]](#)