Block 3: Analysing two variables (and sometimes three)

3.1.3 Contingency tables from SPSS

[last updated 5 March 2014]

Previous tutorial: 3.1.2 Analysing two variables
New SPSS command: **CROSSTABS**

In survey research, contingency tables are used in two ways. The first is **descriptive** by which we mean things like checking the sex and age-group structure of your sample or checking ownership of consumer goods by housing tenure, household income etc.

The second is **explanatory**. As well as **describing** combinations of two or more variables, social researchers often seek to **account** for **variation** in **dependent** variables (the ones being explained) from variation in **independent** variables (the ones being used to do the explaining). Between any two variables it's up to you to decide which way you think the effect flows!

To do this with contingency tables involves comparing percentages or calculating statistical measures of (non-) association. (I'll deal with the latter later in the course, but they are discussed at length in section 6 of Jim Ring's accompanying Statistical Notes and there are plenty of on-line tutorials for SPSS around, far better than anything I am qualified to write.)

For instance, "Is there a relationship between marital status and feeling happy" or "What effect does marital status have on self-reported happiness" ?

The following question has been asked in countless surveys in many countries over the years. This is an extract from the questionnaire for the 1973 Quality of Life in Britain survey: the question was:

Q.J6	Taking all things together, how would you say things are these days? Would you say you are:	READ OUT	VERY HAPPY FAIRLY HAPPY NOT TOO HAPPY	3 2 1	(37)

We'll be using data from this survey to explore the relationship:

Marital status → Feeling happy

. . . or is it the other way round?

Is this the true story, or are there any other variables, (related or unrelated to marital status) which might influence feeling happy? What might they be? How do they affect the relationship between marital status and feeling happy? Thus as well as **dependent** and **independent** variables, we also need to think of **test** variables to examine the initial relationship between marital status and feeling happy by **controlling** for the **test** variables. Does marital status affect feeling happy at all when other variables are taken into account? These are the kind of questions which make survey research so interesting.

When looking at contingency tables for only two variables (usually, but not always, dependent and independent) we use the term **zero order** tables (ie no control variables). When we introduce one test variable, we use the term **1st order** tables; two test variables, **2nd order** tables and so on. During this process we look at what happens to a zero order statistic (eg percent of married

people saying they feel "very happy" these days) when we control for a test variable (eg sex).

Here is the table of counts from the survey. We shall use SPSS to investigate the relationship (if any) between these two variables.

marital * happy Crosstabulation

Count

		happy			
		Not too happy	Fairly happy	Very happy	Total
marital	Single	10	105	38	153
	Married or cohabiting	22	321	342	685
	Widowed	11	59	24	94
	Separated or divorced	12	19	3	34
Total		55	504	407	966

[Source: Quality of Life in Britain, 1973 wave]

CROSSTABS

The SPSS command **CROSSTABS** creates 2-way contingency tables for two variables, and nested sub-tables when controlling for one or more test variables. It also provides percentages based on row, column and global totals, a range of measures of statistical (non-) association and various controls over table content and format.

General format:

```
CROSSTABS  [ TABLES =                          [optional only]
             <row varlist> BY <column varlist>  [ BY <control varlist>]
           / <row varlist> BY <column varlist>  [ BY <control varlist>]
                ~ ~ ~ ~ ~
           /CELLS <cell content options>
           /STATISTICS <statistics options>
           / <other options>
```

The variable(s) specified **before** the keyword **BY** will form the **rows** of the tables, and the variable(s) specified **after** keyword **BY** will form the **columns**, for example:

CROSSTABS sex BY agegroup .

sex * agegroup Crosstabulation

Count

		agegroup				
		18-29	30-44	45-59	60+	Total
sex	Men	113	112	100	115	440
	Women	126	118	131	148	523
Total		239	230	231	263	963

[Source: Quality of Life in Britain, 1973 wave]

CROSSTABS marital BY happy .

marital * happy Crosstabulation

Count

		happy			
		Not too happy	Fairly happy	Very happy	Total
marital	Single	10	105	38	153
	Married or cohabiting	22	321	342	685
	Widowed	11	59	24	94
	Separated or divorced	12	19	3	34
Total		55	504	407	966

[Source: Quality of Life in Britain, 1973 wave]

You can specify multiple tables in the same command, so both tables above can be produced by:

```
CROSSTABS       SEX BY AGEGROUP
                / MARITAL BY HAPPY  .
```

[NB] It is **crucial** to remember to separate multiple table specifications with a forward slash **/**. If you fail to do this, you will produce dozens of useless tables of variables cross-tabulated by themselves, get RSI[1] from scrolling and possibly also waste lots of paper and trees!

As with the **FREQUENCIES** command there is a range of **sub-commands** and **keywords** to control the format and content of your tables and also to request appropriate statistics, and are used in the same way. The most commonly used sub-commands are:

/ CELLS = to specify the content of the table cells, and
/ STATISTICS = to specify the statistics to be applied,

It is also possible to ask for missing values to be tabulated (**/MISSING = INCLUDE**) for labels to be omitted, tables suppressed or values tabulated in descending rather than ascending order (**/ FORMAT** =) and for cells to be written to disk for later statistical processing without having to read the data all over again (**/WRITE =**)

The only ones you need to know about for now are:

Sub-command	Keyword	Meaning
/ CELLS	COUNT	Absolute cell frequency
	ROW	Row percent
	COLUMN	Column percent
	TOTAL	Total percent
/ STATISTICS	CHISQ	Chi-squared test (and I'm leaving that for later!)

As well as counts SPSS can produce three sets of percentages based on row totals, column totals and table totals.

marital * happy Crosstabulation

			happy			
			Not too happy	Fairly happy	Very happy	Total
marital	Single	Count	10	105	38	153
		% within marital	6.5%	68.6%	24.8%	100.0%
		% within happy	18.2%	20.8%	9.3%	15.8%
		% of Total	1.0%	10.9%	3.9%	15.8%
	Married or cohabiting	Count	22	321	342	685
		% within marital	3.2%	46.9%	49.9%	100.0%
		% within happy	40.0%	63.7%	84.0%	70.9%
		% of Total	2.3%	33.2%	35.4%	70.9%
	Widowed	Count	11	59	24	94
		% within marital	11.7%	62.8%	25.5%	100.0%
		% within happy	20.0%	11.7%	5.9%	9.7%
		% of Total	1.1%	6.1%	2.5%	9.7%
	Separated or divorced	Count	12	19	3	34
		% within marital	35.3%	55.9%	8.8%	100.0%
		% within happy	21.8%	3.8%	.7%	3.5%
		% of Total	1.2%	2.0%	.3%	3.5%
Total		Count	55	504	407	966
		% within marital	5.7%	52.2%	42.1%	100.0%
		% within happy	100.0%	100.0%	100.0%	100.0%
		% of Total	5.7%	52.2%	42.1%	100.0%

[Source: Quality of Life in Britain, 1973 wave]

[1] **R**epetitive **S**train **I**njury, common in typists' wrists and fingers

. . .but this makes for a very cluttered presentation, so we need to simplify it and keep only the figures that we really need.

Simple counts aren't much use in answering our research question either, so we need to standardise the distributions by converting the counts into percentages. This makes comparisons easier.

To answer our research question, the dependent variable is "feeling happy" and the independent variable is marital status. We therefore need to look at the distribution of percentages feeling very happy, fairly happy or not too happy for each marital status category. This means we want percentages to add up to 100% within each category of marital status.

Totalling to 100% in columns is the way accountants and statisticians do it,

CROSSTABS <depvar> **BY** <indvar1> .

CROSSTABS **HAPPY BY MARITAL**
/CELLS COUNT COL .

happy * marital Crosstabulation

			marital				
			Single	Married or cohabiting	Widowed	Separated or divorced	Total
happy	Very happy	Count	38	342	24	3	407
		% within marital	24.8%	49.9%	25.5%	8.8%	42.1%
	Fairly happy	Count	105	321	59	19	504
		% within marital	68.6%	46.9%	62.8%	55.9%	52.2%
	Not too happy	Count	10	22	11	12	55
		% within marital	6.5%	3.2%	11.7%	35.3%	5.7%
Total		Count	153	685	94	34	966
		% within marital	100.0%	100.0%	100.0%	100.0%	100.0%

. . . but it's much easier to compare figures visually down columns rather than across rows. I prefer to have the dependent variable across the top of the table and the independent variable(s) down the side, viz:

CROSSTABS <indvar1> **BY** <*d*epvar> .

To repeat the analysis for the tables above the **dependent** variable is **happy** and the **independent** variable is **marital** so we want **marital** to be the row variable and **happy** to be the column variable. We need to be able to compare marital groups so we ask for percent happy within the categories of marital status (along the rows) so we need to tabulate **marital** by **happy** and print **row** percent, thus for the zero order table:

CROSSTABS **MARITAL BY HAPPY**
/CELLS **COUNT ROW .**

marital * happy Crosstabulation

			happy			
			Not too happy	Fairly happy	Very happy	Total
marital	Single	Count	10	105	38	153
		% within marital	6.5%	68.6%	24.8%	100.0%
	Married or cohabiting	Count	22	321	342	685
		% within marital	3.2%	46.9%	49.9%	100.0%
	Widowed	Count	11	59	24	94
		% within marital	11.7%	62.8%	25.5%	100.0%
	Separated or divorced	Count	12	19	3	34
		% within marital	35.3%	55.9%	8.8%	100.0%
Total		Count	55	504	407	966
		% within marital	5.7%	52.2%	42.1%	100.0%

[Source: Quality of Life in Britain, 1973 wave]

This is still a bit cluttered, so a simpler table can be produced by:

CROSSTABS [2] **MARITAL BY HAPPY /CELLS = ROW .**

Row percentages are asked for because we are treating **happy** as the dependent variable and its marginal frequencies will total to 100% at the end of each row.

marital * happy Crosstabulation

% within marital

		happy			
		Not too happy	Fairly happy	Very happy	Total
marital	Single	6.5%	68.6%	24.8%	100.0%
	Married or cohabiting	3.2%	46.9%	49.9%	100.0%
	Widowed	11.7%	62.8%	25.5%	100.0%
	Separated or divorced	35.3%	55.9%	8.8%	100.0%
Total		5.7%	52.2%	42.1%	100.0%

[Source: Quality of Life in Britain, 1973 wave]

This table is easier to interpret, but we have now lost the base for percentaging at the end of each row. Without special programming beyond the scope of this tutorial, SPSS cannot produce a table with base n instead of 100%. The table is also a bit cluttered with superfluous % signs, but with judicious editing (in Word) can be made easier to read and interpreted (my emphasis in red)

Marital status of respondent * Q.J6 Would you say you are...?

% within Marital status of respondent

		happy J6 Would you say you are...?			Total
		Not too happy %	Fairly happy %	Very happy %	n = 100%
Marital status of respondent	Single	6.5	68.6	24.8	153
	Married or cohabiting	3.2	46.9	49.9	685
	Widowed	11.7	62.8	25.5	94
	Separated or divorced	35.3	55.9	8.8	34
Total		**5.7**	**52.2**	**42.1**	966

[Source: Quality of Life in Britain, 1973 wave]

Tables like this are known as **zero order** tables (ie no other variable apart from the two being tabulated). If you wish to introduce a single **control** or **test** variable to see if the distribution changes within each sex (known as a **1st order** table) there will be nested sub-tables of row-variable(s) by column-variable(s) for each value, or combination of values, of the control variable(s) after the second **BY**.

CROSSTABS MARITAL BY HAPPY BY SEX /CELL ROW.

marital * happy * sex Crosstabulation

% within marital

sex			happy			
			Not too happy	Fairly happy	Very happy	Total
Men	marital	Single	4.8%	76.2%	19.0%	100.0%
		Married or cohabiting	3.6%	53.5%	42.9%	100.0%
		Widowed	6.7%	66.7%	26.7%	100.0%
		Separated or divorced	50.0%	41.7%	8.3%	100.0%
	Total		5.2%	57.9%	36.9%	100.0%
Women	marital	Single	8.7%	59.4%	31.9%	100.0%
		Married or cohabiting	2.8%	40.7%	56.5%	100.0%
		Widowed	12.7%	62.0%	25.3%	100.0%
		Separated or divorced	27.3%	63.6%	9.1%	100.0%
	Total		6.1%	47.3%	46.6%	100.0%

[Source: Quality of Life in Britain, 1973 wave] **SPSS 19 output**

[2] SPSS reads only the first three or four characters of subcommands and keywords: you don't need the = sign either and lower case is easier to read so (less typing and easier to read): **cro marital by happy /cel row .**

marital Marital status of respondent * happy J6 Would you say you are...? * sex Sex of respondent Crosstabulation

% within marital Marital status of respondent

sex Sex of respondent			happy J6 Would you say you are...?			Total
			Not too happy	Fairly happy	Very happy	
1 Men	Marital status of respondent	Single	4.8%	76.2%	19.0%	100.0%
		Married or cohabiting	3.6%	53.5%	42.9%	100.0%
		Widowed	6.7%	66.7%	26.7%	100.0%
		eparated or divorced	50.0%	41.7%	8.3%	100.0%
	Total		5.2%	57.9%	36.9%	100.0%
2 Women	Marital status of respondent	Single	8.7%	59.4%	31.9%	100.0%
		Married or cohabiting	2.8%	40.7%	56.5%	100.0%
		Widowed	12.7%	62.0%	25.3%	100.0%
		Separated or divorced	27.3%	63.6%	9.1%	100.0%
	Total		6.1%	47.3%	46.6%	100.0%
Total	Marital status of respondent	Single	6.5%	68.6%	24.8%	100.0%
		Married or cohabiting	3.2%	46.9%	49.9%	100.0%
		Widowed	11.7%	62.8%	25.5%	100.0%
		Separated or divorced	35.3%	55.9%	8.8%	100.0%
	Total		5.7%	52.2%	42.1%	100.0%

[Source: Quality of Life in Britain, 1973 wave] **SPSS 21 output** (inserts sub-totals)

However, it's easier to compare men and women if the figures are displayed by marital within sex:

```
CROSSTABS    SEX BY HAPPY BY MARITAL /CELL ROW.
```

SPSS 21 displays sub-totals with gridlines:

Sex of respondent * J6 Would you say you are...? * Marital status of respondent Crosstabulation

% within Sex of respondent

Marital status of respondent			J6 Would you say you are...?			Total
			Not too happy	Fairly happy	Very happy	
Single	Sex of respondent	Men	4.8%	76.2%	19.0%	100.0%
		Women	8.7%	59.4%	31.9%	100.0%
	Total		6.5%	68.6%	24.8%	100.0%
Married or cohabiting	Sex of respondent	Men	3.6%	53.5%	42.9%	100.0%
		Women	2.8%	40.7%	56.5%	100.0%
	Total		3.2%	46.9%	49.9%	100.0%
Widowed	Sex of respondent	Men	6.7%	66.7%	26.7%	100.0%
		Women	12.7%	62.0%	25.3%	100.0%
	Total		11.7%	62.8%	25.5%	100.0%
Separated or divorced	Sex of respondent	Men	50.0%	41.7%	8.3%	100.0%
		Women	27.3%	63.6%	9.1%	100.0%
	Total		35.3%	55.9%	8.8%	100.0%
Total	Sex of respondent	Men	5.2%	57.9%	36.9%	100.0%
		Women	6.1%	47.3%	46.6%	100.0%
	Total		5.7%	52.2%	42.1%	100.0%

[Source: Quality of Life in Britain, 1973 wave] **SPSS 21 output** (inserts sub-totals and grid-lines)

The default output table is again cluttered, but can be made easier to read and interpret if you change the SPSS settings to print labels only, get rid of the % signs in the body of the table and replace the 100% totals with the actual base n. This takes a long time!

S Sex by Happy controlling for Marital

% within Sex of respondent

Marital status of respondent			J6 Would you say you are...? Not too happy %	Fairly happy %	Very happy %	Total (n = 100%)
Single	Sex of respondent	Men	4.8	76.2	19.0	84
		Women	8.7	59.4	31.9	69
	Sub-total		**6.5**	**68.6**	**24.8**	**153**
Married or cohabiting	Sex of respondent	Men	3.6	53.5	42.9	331
		Women	2.8	40.7	56.5	354
	Sub-total		**3.2**	**46.9**	**49.9**	685
Widowed	Sex of respondent	Men	6.7	66.7	26.7	15
		Women	12.7	62.0	25.3	79
	Sub-total		**11.7**	**62.8**	**25.5**	**94**
Separated or divorced	Sex of respondent	Men	50.0	41.7	8.3	12
		Women	27.3	63.6	9.1	22
	Sub-total		**35.3**	**55.9**	**8.8**	**34**
Total	Sex of respondent	Men	5.2	57.9	36.9	442
		Women	6.1	47.3	46.6	524
	Total		**5.7**	**52.2**	**42.1**	**966**

[Source: Quality of Life in Britain, 1973 wave]

From these tables (my emphasis in red) we can see that, for the 153 people who are single, the figure of **24.8%** answering "very happy" in the zero-order table is actually composed (a weighted average) of **19%** for 84 men and **31.9%** for 69 women in the 1st order table. Of the 34 separated or divorced respondents, **35%** report feeling "not too happy", but men (**50%** of 12) are almost twice as likely as women (**27%** of 22) to report this. Now why might that be?

Is it the same for the 1975 wave (different sample from the same primary sampling units) or for more recent UK surveys, or for people in other countries? Are the sub-samples large enough to be confident, within certain limits, that these are "real" differences? Probably not, as a general rule is that 40 should be the minimum cell count to use as a base for %%. Even then moving a single case form one category to another makes a net difference of five percentage points. The number of widowed men (15) and separated or divorced respondents (34) are below this threshold.

This is what **analysis** (from the Greek for "break up") is all about. Once you start hunting for the truth in this way, hopefully guided by reasonable theory or previously formulated hypotheses, you will find it as addictive as it is challenging and also a lot of fun, especially once you get the hang of SPSS syntax.

End of tutorial

Next: **Exercises for Contingency tables**

Next session: 3.1.4.1 Income differences work-through

[Back to Block 3 menu]