SO5041 Unit 4: Bivariate analyses

Brendan Halpin, Sociology

Autumn 2023/4

SO5041

Reading

Background Reading

- Agresti, *Statistical Methods for the Social Sciences*, chapter 1 (background) and chapter 3 (descriptive statistics)
- To follow: Agresti, chapter 2, Sampling and Measurement

SO5041

Bivariate analysis

Types of variables

- So far i have been using several terms for types of variable:
  - Categorical
  - Grouped
  - Continuous
  - Quantitative
- A more formal classification that corresponds with the sorts of analysis possible, goes:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

NOIR – categorical

- Nominal variables have categories where each category is "just a name", i.e., nominal: example: religion, region of birth, political allegience
- With nominal variables we can't do much more than present frequencies
- Ordinal variables have categories where each category is something different, but where it is possible to put the categories in a meaningful high–low order. Examples include highest maths qualification, exam letter grades, attitudes (agree strongly, agree, neutral, disagree, disagree strongly) and so on. with ordinal variables, frequencies are meaningful, and the "cumulative percent" column that Stata provides is also meaningful. Additionally, with ordinal variables we can calculate the median

NOIR – quantitative

- Quantitative variables can be interval or ratio
- Interval variables are variables like temperature, where the difference between, say, 35° and 40° is the same as between 50° and 55°. That is, the meaning of an *interval* is the same no matter where it is. However, with temperature (centigrade or fahrenheit) it is not true that 40° is twice as hot as 20°, because 0° is not a true starting point. With interval variables it makes sense to calculate the mean. That is, if we have five days with temperatures of 11°, 9°, 12°, 11° and 10°, it makes sense to say the average is $\frac{11+9+12+11+10}{5} = 10.6$

NOIR – ratio variables

- Ratio variables are like interval variables, and are probably more commonly encountered. these are quantitative variables, where the zero makes sense. distance, duration, money, age are all ratio variables, because 0 kilometres, 0 seconds, €0 and 0 years are all things that make sense. ratio variables can also use mean, but unlike interval variables we can also work with proportions (twice as old, twice as rich, 10% farther and so on)

SO5041

Bivariate Summaries

Recap: Univariate analysis

- Frequencies: `tab foreign`
- Pie chart: `graph pie, over(foreign)`
- Bar chart: `graph bar, over(foreign)`
- Medians, etc: `centile mpg, centile(25 50 75)`
- Mean, standard deviation: `su mpg`
- Histogram: `histogram mpg`

Bivariate Analysis

- So far we have dealt with describing one variable at a time: important but limited
- Bi-variate (two-variable) summaries allow us to look at *relationships* between variables as well
- Where both variables are categorical (nominal, ordinal or grouped) we can use two-way tables, cross-tabulations

Cross-tabulation

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta

. tab empstat sex
```

usual employment situation	school leavers sex 1	2	Total
working for payment	388	380	768
unemployed	67	46	113
looking for 1st job	170	151	321
student	471	490	961
engaged in home dutie	0	19	19
permanent disability/ other	4	0	4
	4	7	11
Total	1,104	1,093	2,197

Column Percentages

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta
. tab empstat sex, col
```

Key
frequency *column percentage*

usual employment situation	school leavers sex 1	2	Total
working for payment	388 35.14	380 34.77	768 34.96
unemployed	67 6.07	46 4.21	113 5.14
looking for 1st job	170 15.40	151 13.82	321 14.61
student	471 42.66	490 44.83	961 43.74
engaged in home dutie	0 0.00	19 1.74	19 0.86
permanent disability/	4 0.36	0 0.00	4 0.18
other	4 0.36	7 0.64	11 0.50
Total	1,104 100.00	1,093 100.00	2,197 100.00

Too many percents!

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta
. tab empstat sex, row col
```

Key
frequency *row percentage* *column percentage*

usual employment situation	school leavers sex 1	2	Total
working for payment	388	380	768
	50.52	49.48	100.00
	35.14	34.77	34.96
unemployed	67	46	113
	59.29	40.71	100.00
	6.07	4.21	5.14
looking for 1st job	170	151	321
	52.96	47.04	100.00
	15.40	13.82	14.61
student	471	490	961
	49.01	50.99	100.00
	42.66	44.83	43.74
engaged in home dutie	0	19	19
	0.00	100.00	100.00
	0.00	1.74	0.86
permanent disability/	4	0	4
	100.00	0.00	100.00
	0.36	0.00	0.18
other	4	7	11
	36.36	63.64	100.00
	0.36	0.64	0.50
Total	1,104	1,093	2,197
	50.25	49.75	100.00

Ordinal variable

```
. tab mopfama mopfamb, row
```

Key
frequency *row percentage*

pre-school child suffers if mother works	family suffers if mother works full-time					
	strongly	agree	neithr ag	disagree	strongly	Total
strongly agree	572	328	79	32	21	1,032
	55.43	31.78	7.66	3.10	2.03	100.00
agree	267	2,507	876	354	37	4,041
	6.61	62.04	21.68	8.76	0.92	100.00
neithr agree, disagre	55	813	3,070	1,010	95	5,043
	1.09	16.12	60.88	20.03	1.88	100.00
disagree	32	295	437	2,777	344	3,885
	0.82	7.59	11.25	71.48	8.85	100.00
strongly disagree	6	23	40	164	680	913
	0.66	2.52	4.38	17.96	74.48	100.00
Total	932	3,966	4,502	4,337	1,177	14,914
	6.25	26.59	30.19	29.08	7.89	100.00

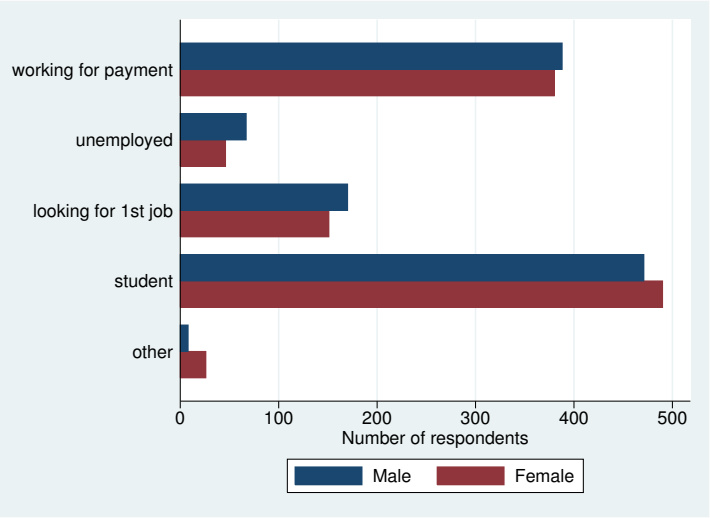
SO5041

Bivariate graphs

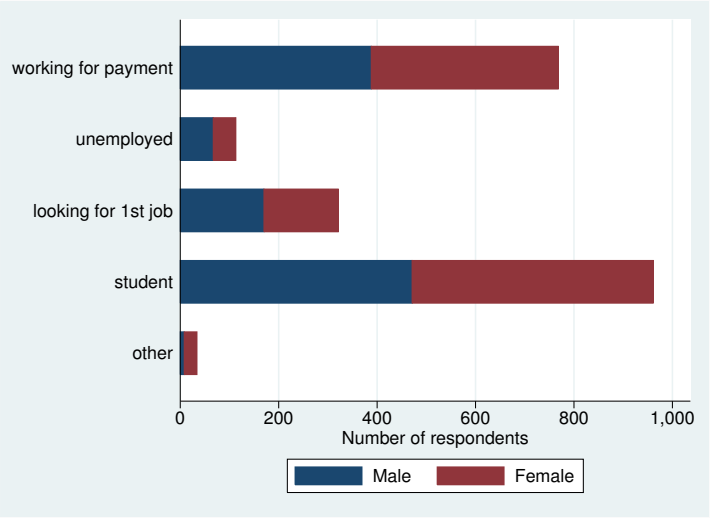
Bivariate graphs

- Graphical ways of presenting the same information as a cross-tabulation include clustered and stacked bar charts:
- Clusters are good for looking at distributions within groups
- Stacked bars are good where it is important to emphasise the sizes of the groups

Clustered bar chart



Stacked bar chart



Compare means – continuous within categorical

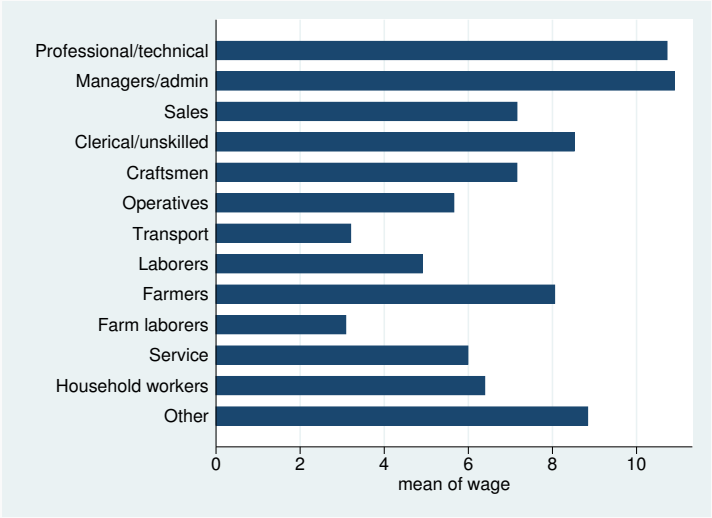
- Where one of the variables is categorical and the other is continuous (interval or ratio) we can “compare means”. That is, instead of calculating mean income, we can calculate mean income for different groups

Wage by occupational group

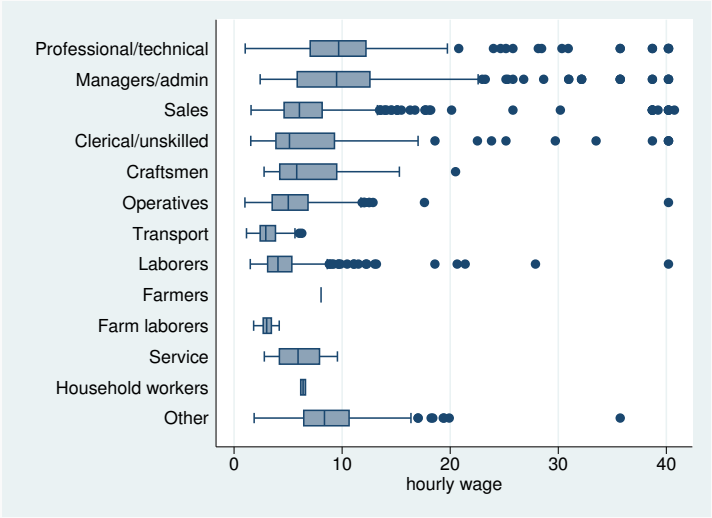
```
. sysuse nlsw88, clear
(NLSW, 1988 extract)

. tab occupation, su(wage)
```

occupation	Summary of hourly wage Mean	Std. Dev.	Freq.
Professio	10.723624	6.3510736	317
Managers/	10.899784	7.5215875	264
Sales	7.1544893	5.0427568	726
Clerical/	8.5166856	8.5653995	102
Craftsmen	7.152988	3.7629626	53
Operative	5.6538061	3.3861932	246
Transport	3.2004937	1.3209668	28
Laborers	4.9058895	3.6083499	286
Farmers	8.0515299	0	1
Farm labo	3.0837354	.76638443	9
Service	5.9887432	2.1399333	16
Household	6.3888881	.31313052	2
Other	8.8362936	4.128914	187
Total	7.7779283	5.7613599	2,237

Another use of bar charts



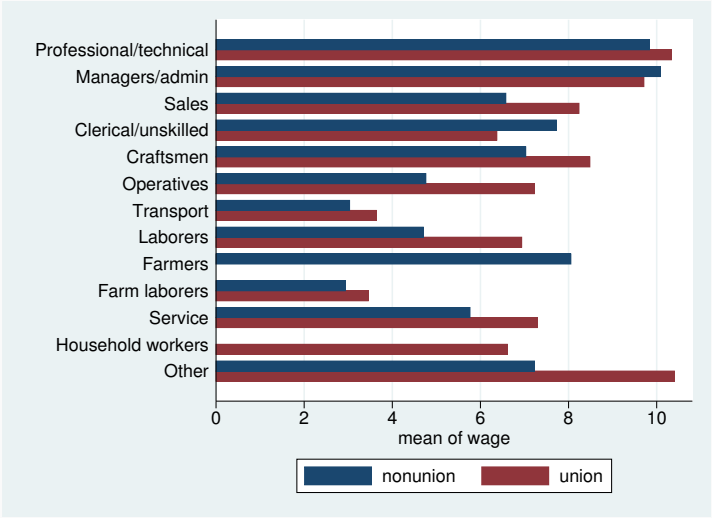
Boxplot


Professional/technical
Managers/admin
Sales
Clerical/unskilled
Craftsmen
Operatives
Transport
Laborers
Farmers
Farm laborers
Service
Household workers
Other
0
10
20
30
40
hourly wage

Boxplot

- The boxplot summarises the distribution of a variable:
  - 50% of the distribution is within the box
  - The bar in the middle is the median
  - The “whiskers” extend to the minimum and maximum, excluding “outliers”
  - Outliers are cases more than 1.5 IQR above or below the corresponding quartile (these are often marked separately)

Three-way analysis



Scatterplot

- Where both variables are continuous, a very useful device is the scatterplot: