SO5041 Unit 3: Numbers as Information – Data

Brendan Halpin, Sociology

Autumn 2019/0

Number as information

- In the Lab we will begin by coding data in numerical form, entering it on the computer and using Stata to make simple descriptive summaries of it
- This activity is an important stage of the quantitative research process: turning information into numbers and numbers into information

Steps

- There are several important steps in this process
  - Determine a clear set of information items to collect (i.e., what questions you want to ask)
  - Determine a simple way of representing this information
    - Represent it as a single "quantitative" number, e.g., Euros per annum for income, or
    - Determine a fixed set of categories grouping every possible answer, and attach a number to each category
  - Collect it (easier said than done) and enter on a computer
  - But for the analysis to make sense we need to bring in information about that the numbers mean: variable labels and value labels restore some part of the meaningful information hidden by coding as numbers

Descriptive summaries

- When we enter data on a computer we can easily present descriptive univariate summaries (i.e., summarising one variable at a time)
- For categorical or grouped variables (i.e., where there is a "small" number of different values) we can use a frequency table: how many of each sort are there? what proportion of each sort are there?
- For variables with a "quantitative" interpretation (i.e., where the number measures something like age, income, duration, distance, or where it is a count like number of children, number of cigarettes per day) we can explore the mean or average, or perhaps the median

Frequency Table

```
. tab rjbstat
```

current economic activity	Freq.	Percent	Cum.
self-employed	929	6.82	6.82
employed	6,749	49.58	56.41
unemployed	468	3.44	59.84
retired	3,034	22.29	82.13
maternity leave	75	0.55	82.68
family care	786	5.77	88.46
ft studt, school	878	6.45	94.91
lt sick, disabld	608	4.47	99.38
gvt trng scheme	22	0.16	99.54
other	63	0.46	100.00
Total	13,612	100.00	

Summarising a "quantitative" variable

```
. su rfimn
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rfimn	13,612	1454.687	1459.462	0	56916.67

Codebook

```
. codebook rfimn
```

---

rfimn — total income: last month

---

type:	numeric (double)		
label:	rfimn, but 9658 nonmissing values are not labeled		
range:	[0,56916.668]	units:	1.000e-08
unique values:	9,658	missing .:	0/13,612
examples:	470.71835		
	925.07501		
	1401.9929		
	2175.7588		

Mean, Median

- The mean is defined as the sum total of the variable, divided by the number of cases:

$$\bar{x} = \frac{\sum x_i}{n}$$

- The median is defined as the value such that 50% of cases are lower and 50% of cases are higher

Calculating the median

- To calculate the median "by hand":
  - sort the values in order
  - if there is an odd number of cases (e.g., 101), find the middle one (e.g., 51st) – its value is the median
  - if there is an even number of cases (e.g., 100), find the middle pair (e.g., 50th and 51st) – the median is half way between their values

Median via Stata

```
. centile rfimn
```

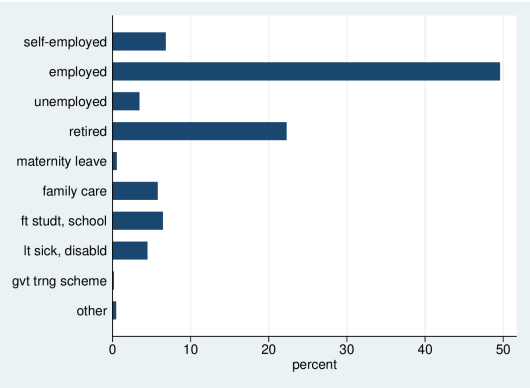
Variable	Obs	Percentile	Centile	Binom. Interp. [95% Conf.	Interval]
rfimn	13,612	50	1141.668	1119.325	1166.667

Graphs

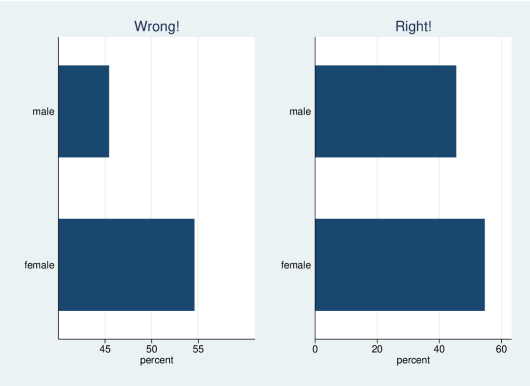
- We can also make graphical summaries
- For categorical variables the bar-chart is very good: this is like a frequency table where the height of the bars is proportional to the number/proportion in that category
- We can also use Pie-charts: these are circles with segments (pie-slices) whose angle is proportional to the size of the category (there are some arguments that bar-charts are better)
- For quantitative variables we can use histograms: these are very like bar-charts, but break the "continuous" variable into groups. The bars in a histogram touch, to show that the variable is really continuous. In a histogram, the area under the "curve" (or stepped line) for a given range is proportional to the numbers in that range. It is sometimes interesting to vary the group size (the "bin size") to get more or less detail

Bar chart

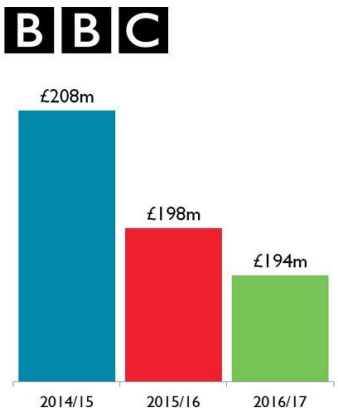
```
graph hbar, over(rjbstat)
```



Zero is important



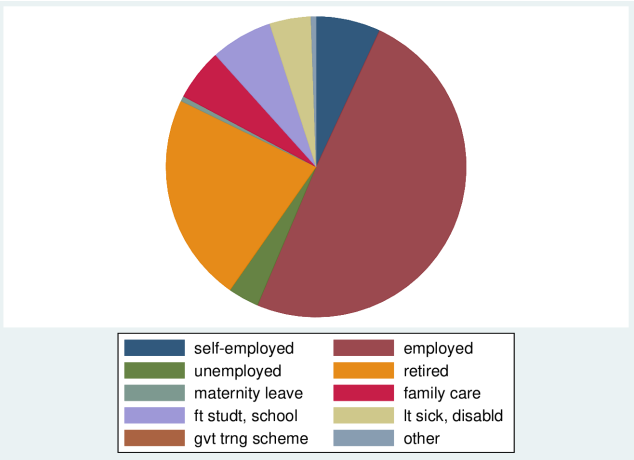
Beware: People lie with barcharts



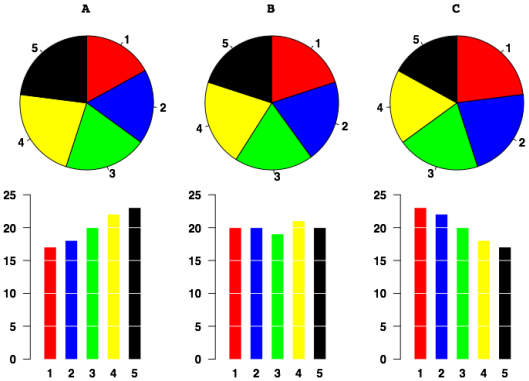
Talent pay

We have significantly reduced the total bill spent on paying talent, down again this year.

Pie chart


self-employed
employed
unemployed
retired
maternity leave
family care
ft studt, school
lt sick, disabld
gvt trng scheme
other

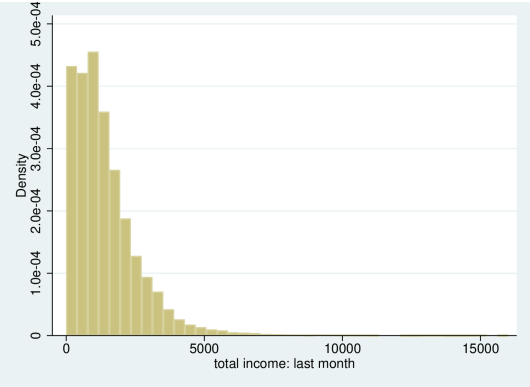
Pie is bad for you



Source: http://junkcharts.typepad.com/junk_charts/2009/08/community-outreach-piemaking.html

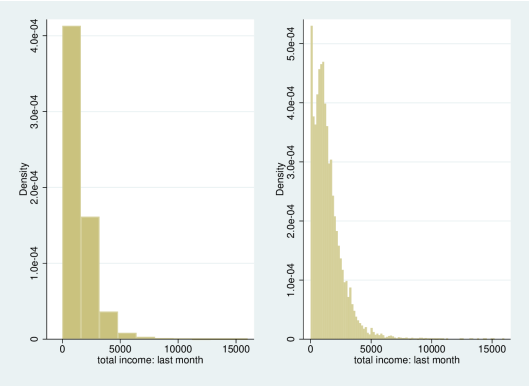
Histograms

```
hist rfimn if rfimn<=16000
```



Different bin widths

```
hist rfimn if rfimn<=16000, bin(10)
hist rfimn if rfimn<=16000, bin(100)
```



Some interesting reading on graphics

- William S Cleveland *The Elements of Graphing Data*. Rev. ed. Murray Hill, N.J.: AT&T Bell Laboratories, 1994
- Edward Tufte, *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press, 1983
- Kieran Healy, *Data Visualization: A Practical Introduction* 2018. Focused on R, but Chapter 1 is a general discussion. Online at `https://socviz.co/`