SO5041 Unit 5: More on types of variables

Brendan Halpin, Sociology

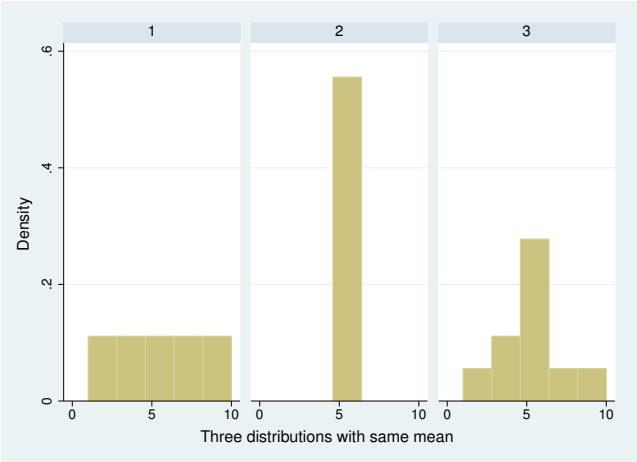
Autumn 2019/0

Spread in quantitative variables

- When we have quantitative variables like age, income, we use *measures of central tendency* like mean and median to describe the "middle"
- However, the dispersion about the middle is also important: completely different distributions can have the same mean:

	A	B	C
1	5.5	1.0
2	5.5	4.0
3	5.5	4.5
4	5.5	5.2
5	5.5	5.3
6	5.5	5.6
7	5.5	5.9
8	5.5	6.0
9	5.5	7.5
10	5.5	10.0
Mean	5.5	5.5	5.5

Spread in quantitative variables



Summarising spread

- As well as summarising the middle, we often want to summarise the spread:
  - Range
  - Inter-quartile range
  - Standard Deviation
- The range is important, but depends too much on just the two extreme cases
- The IQR is much more stable
- The standard deviation has useful statistical properties

The "Standard Deviation"

- The standard deviation is calculated from the "deviations" from the mean, $\bar{X}$:

$$\text{Deviation} = X_i - \bar{X}$$

- The deviations add to zero so we can't just add them up: instead square them and then add them up:

$$\sum (X_i - \bar{X})^2$$

- To get a sort of average, we divide by sample size minus 1, and take the square root:

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Standard Deviation: good measure of spread

- The standard deviation indicates how spread out the data is: the more spread out, the bigger the StDev
- It's a good measure because it depends on every single case, not just the extreme pairs (range) or the quartiles
- It has useful statistical properties which help when we come to statistical inference

More on types of variables

- We have already divided variables into nominal/ordinal/interval/ratio
- We can also differentiate between "qualitative" and "quantitative":
  - Nominal variables are clearly qualitative: observations with different values have different qualities
  - Interval/ratio variables directly represent quantities: amount of money, number of children, distance

What about ordinal variables?

- Where do ordinal variables (*e.g.*, level of education) fit in?
  - In between?
  - Often treated as qualitative – categories are clearly "different"
  - Sometimes treated as quantitative: *e.g.*, letter-grades are given scores for calculating QCA; "Likert" scale answers (strongly agree to strongly disagree) are given scores of -2, -1, 0, 1, 2
  - Applying scores to an ordinal variable implies there is an underlying interval/ratio variable which we do not measure
  - In the QCA example this is true: the individual exam marks are on a ratio scale
  - In the likert-scale example, we may feel there is a continuum of agreement–disagreement which is approximately measured by the five options

Sensitivity

- There, however, we may feel that the steps between neutral and (dis)agree are bigger than between (dis)agree and strongly (dis)agree so we might want the scores to go -3, -2, 0, 2, 3 (or -3, -1, 0, 1, 3)
- When we base an analysis on applied scores like these, it is good to do a "sensitivity analysis": compare the results with different scoring systems to see how sensitive your analysis is to the scoring method

Summary of 5-point ordinal scale

Question	VP	P	Neutral	G	VG	N	Score
Overall satisfaction with Department of Sociology	1.4	1.7	23.2	52.6	21.1	289	0.781
National reputation of University of Limerick	0.7	2.5	16.5	36.6	43.7	279	0.840
Overall quality of academic programme within UL	0.0	5.1	13.6	54.2	27.1	295	0.807
Overall quality of services provided by UL	1.0	7.7	20.1	47.8	23.4	299	0.770
Overall quality of the social aspects at UL services	0.7	3.2	23.6	36.6	35.9	284	0.808
Quality of Student Academic Administration office	3.9	12.5	33.9	37.1	12.5	280	0.684
Quality of Fees office	2.9	7.5	47.3	33.5	8.8	239	0.675

Yet another distinction!

- A new distinction: discrete versus continuous
- Discrete variables have a finite number of possible values
  - Nominal and ordinal variables are by definition discrete
  - Count variables are discrete: 0, 1, 2, 3 . . . – you can't actually have 2.4 children
  - Grouped interval variables are discrete
- Continuous variables can have an infinite number of values in principle: 2, 2.4, 2.43, 2.435, 2.4358 and so on.
- Ungrouped interval/ratio variables are continuous in principle: an infinite number of values are possible
- In practice, we treat income as continuous though it can't vary in amounts less than 1 cent (and age, though we often measure it in integer years, or to the nearest month, etc.)