Data Analysis for the Social Sciences

Quantitative Data Analysis II

2023-09-26

Welcome everyone to the second of two lectures on quantitative data analysis.

Our focus today is on how to conduct a sensible, robust piece of quantitative data analysis.

"The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." John Tukey

Research Aims and Variables

Our first task is to explicitly state what our aim is and what variables we have.

Research aims

We can use quantitative methods for (Agresti, 2018):

1. **Designing research studies** to investigate questions of interest (including the process of obtaining data)
2. **Description** – summarising the data appropriately
3. **Inference -** making predictions using the data, in a way that deals with uncertainty of our analysis

We are not interested in 1 in this module (covered last year in Foundations of Quantitative Research Methods), as we will use data that have already been collected. But quantitative methods can help us design sampling strategies e.g., who to sample and how (+ how many).

Research aims

Description and inference are two ways of analysing data.

"*Descriptive statistics summarize the information in a collection of data. Inferential statistics provide predictions about a population, based on data from a sample of that population*." (Agresti, 2018: 17)

Identifying variables

Are climate change beliefs associated with sex and age among British people?

Say we have a research question. How can we use it to structure our analysis? Put another way, which concept are we interested in explaining?

Identifying variables

Dependent Variable (Y) = outcome we are interested in explaining / predicting.

Independent Variable (X) = factor we think explains / predicts the outcome.

$$Y = X_1 + \in$$

$$Y = X_1 + X_2 + \cdots + X_K + \in$$

Y = X + e is a simple model for when we have one independent variable.

We usually focus on more than one independent variable in social science analyses.

Identifying variables

Are climate change beliefs associated with sex and age among British people?

Y = Climate change beliefs

X1 = Sex at birth

X2 = Age

Returning to our research question, is it now apparent what our dependent and independent variables are?

In practice there may be more than one dependent variable e.g., many questions asking about climate change beliefs.

Implications

Research Aims	Affects choice of analytical technique	Mean, median, standard deviation, correlation statistics = descriptive statistics Chi-squared, confidence intervals, p-values = inferential statistics
Identifying Variables	Affects what variables are included in analysis and to what degree	1 Y and 5 X = six variables needing to described, and five relationships needing to be explored

Structuring your analysis

Order	Type of analysis	Purpose	Techniques
1	**Univariate**	Analyse each variable individually	Frequency table Mean, median and mode
2	**Bivariate**	Examine the relationships between the dependent variable and each independent variable	Scatterplots Cross-tabulations
3	**Multivariate**	Examine whether the bivariate relationships vary across values of other variables	Cross-tabulations by groups Statistical model

These are in order

[Break]

Univariate

Univariate analysis

Univariate analysis is concerned with summarising a single variable, specifically:

1. The **central tendency** of the values
2. The **variability** (distribution) of the values

Measures of central tendency

1. **Mean** = typical value
2. **Median** = typical observation / case
3. **Mode** = most common value

"*The mean, median, and mode are complementary measures. They describe different aspects of the data. In any particular example, some or all their values may be useful*." (Agresti, 2018: 53)

Measures of central tendency try to summarise variables using a single number. Is there a representative / typical value for a variable?

Measures of central tendency

Properties of these measures (Agresti, 2018):

Mean	Median	Mode
Influenced by outliers	Not influenced by outliers	Not influenced by outliers
Not necessarily an actual value	Actual value	Actual value
Applicable to numeric variables	Applicable to numeric and ordinal variables	Applicable to all variables

Measures of variability

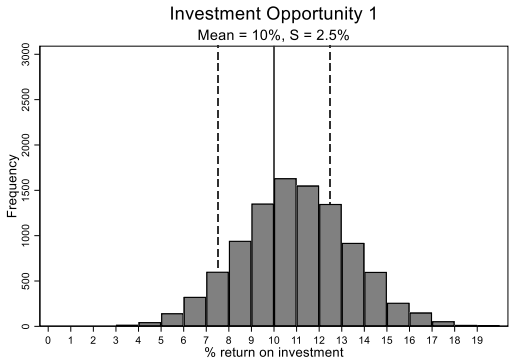
Range = difference between maximum and minimum values

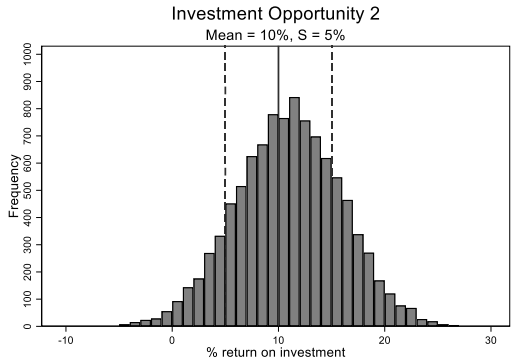
= 88 – 16 = 72

Standard Deviation (s) = typical difference between a value and the mean

The larger the standard deviation is, the more spread out the observations are (Agresti, 2018).

Measures of variability





Consider two investment opportunities:
Which one would you prefer to take? [Class Poll]

Bivariate

Bivariate analysis

Bivariate analysis is concerned with making comparisons using two variables.

The purpose of comparing two or more variables is to uncover *relationships*.

Relationships can be strong, moderate or weak; positive, negative or non-existent (Huntington-Klein, 2021).

In quantitative data analysis: is a dependent variable related to one or more independent variables?

Is academic performance related to attendance at workshops?

Bivariate analysis of categorical variables

	Achieved a 2:1 in module (%)	
Attended at least 50% of workshops	Yes	No
Yes	38	62
No	26	74
	32	**68**

Question: is there a relationship between these two variables? [SHOW POLL]

Representing the joint distribution of two categorical variables is called a crosstabulation or contingency table.

For example, we can ask the question: is your likelihood of achieving at least a 2:1 contingent on attending workshops?

Dependent variable is usually in the columns, independent variable in the rows, and we use row percentages to make comparisons between categories of the independent variable.

Correlations

Examining the joint distribution of two variables is informative but leaves one outstanding question:

- How can we quantify the pattern in the joint distribution? (De Mesquita and Fowler, 2021)

Correlations tell us about the extent to which two features of the world tend to occur together. (De Mesquita and Fowler, 2021)

Thinking of the pattern shown in the previous table, it would be good to quantify how strong or weak a relationship is.

Therefore a correlation is simply a single number that summarises a relationship between two variables.

There are lots of different correlation statistics and the choice of the most appropriate one is determined by the level of measurement of your variables.

We'll see examples in the lab today but some you may have heard of: Pearson's r coefficient, Kendall's Tau, Gamma, Cramer's V.

Multivariate

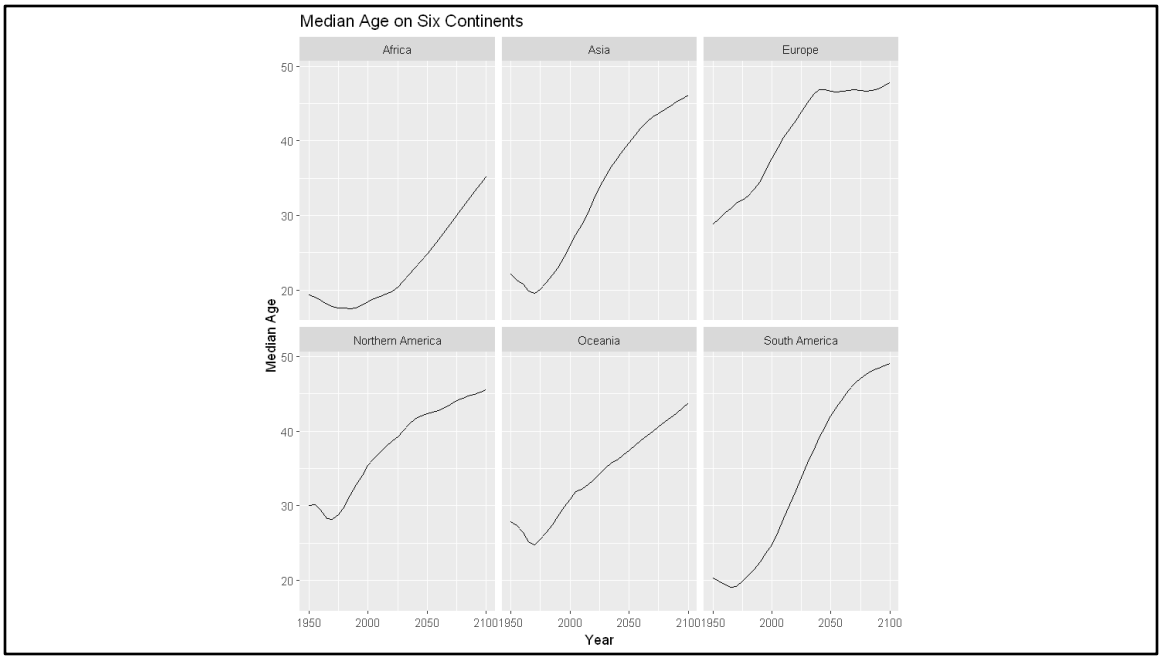
Multivariate analysis

Multivariate analysis is concerned with testing whether the bivariate analyses vary across values of a third / fourth / fifth etc variable.

The social world is complex and there many relevant factors for a single outcome (or many independent variables affecting a dependent variable).

Is there a difference in the earnings of men and women?

Is this the case for all age groups? Or is it really only older men who earn more than older women?



We covered loads today, so please take the time to revise the lecture, engage with the core reading etc before the first assessment.

If you get through all of the content associated with this week and last, you will have an excellent grasp of what quantitative data analysis involves, regardless of whether you conduct some QDA for assessment 2.