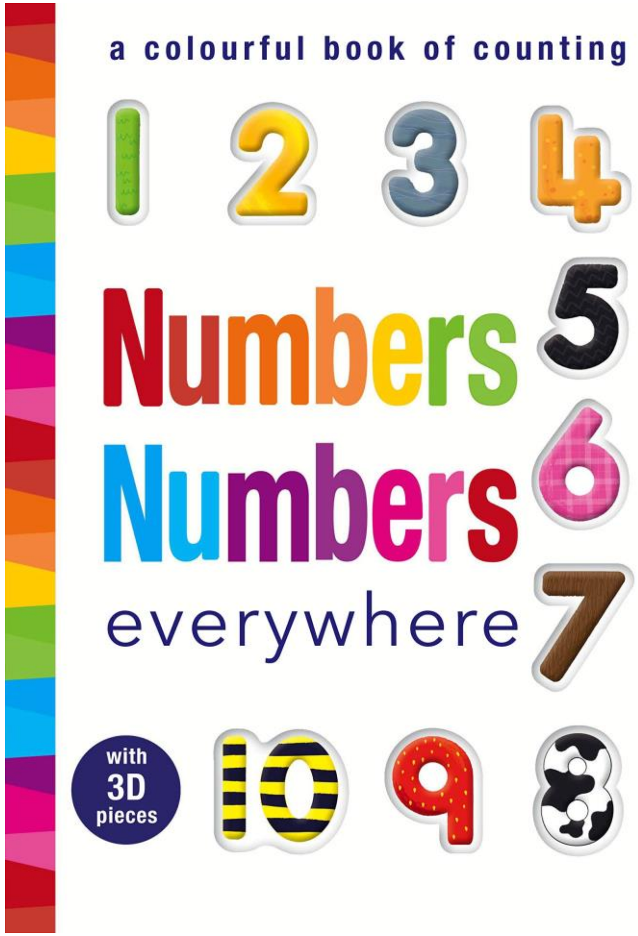
Data Analysis for the Social Sciences

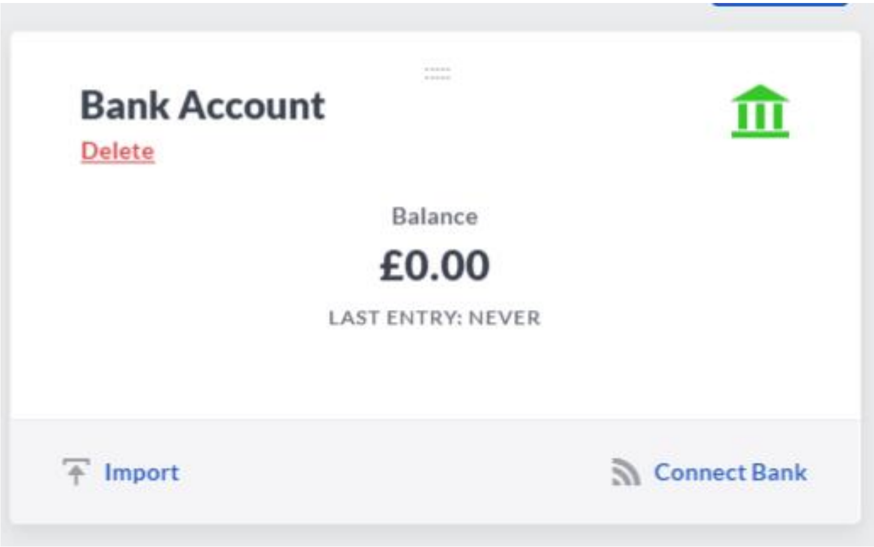
Quantitative Data Analysis I

2024/25

“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

Numbers, numbers everywhere



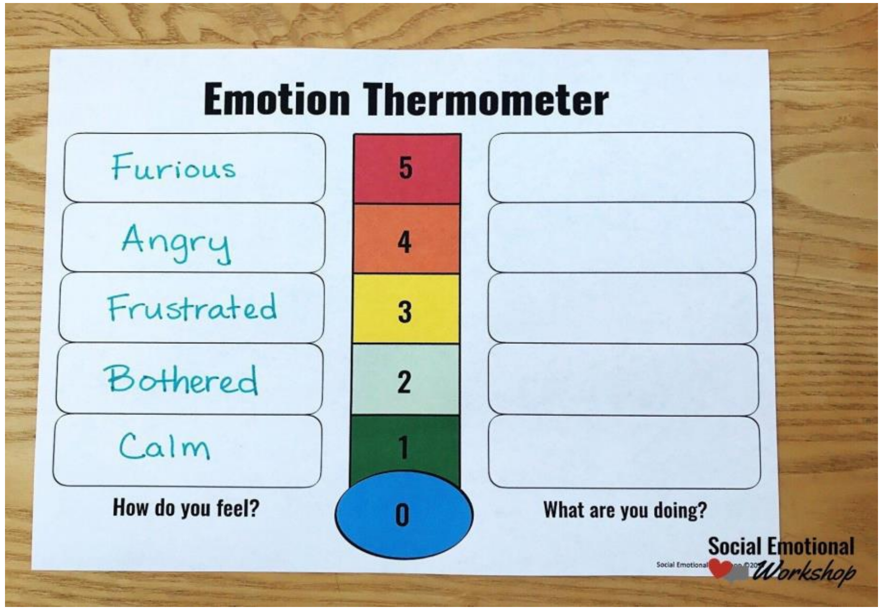

Bank Account

Delete

Balance

£0.00

LAST ENTRY: NEVER


Emotion Thermometer

Furious	5	
Angry	4	
Frustrated	3	
Bothered	2	
Calm	1	
How do you feel?	0	What are you doing?

Social Emotional Workshop

What are data?

A collection of observations (Agresti, 2018):

- **An observation** is a set of measurements for a case.
- **A measurement** is a description of some characteristic of a case.
- **A case** (or subject) is the entity we are observing e.g., individuals, countries, animals, companies, networks etc.

What are quantitative data?

A collection of observations in a particular format **[Variable-by-case matrix]**

- **Case** = the entity we are observing e.g., individuals, countries, animals, companies, networks etc.
- **Variable** = what we are measuring about a case e.g., a characteristic
- **Value** = the result of measuring a variable for a given case (MacInnes, 2017)

Measurement

Measurement Scale	Level of Measurement	Description	Example
Categorical	Nominal	Presence of some attribute	Sex at birth, Ethnicity
Categorical	Ordinal	More or less of some attribute	Social class, Degree classification
Numeric	Interval	More or less of some attribute (and by how much)	Income, Number of deaths, Age

Research Aims and Variables

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

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 influenced by a person’s sex and age?

Identifying variables

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

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

$$Y = X_1 + \in$$

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

Identifying variables

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

Y = Climate change beliefs

X1 = Sex at birth

X2 = Age

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

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

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

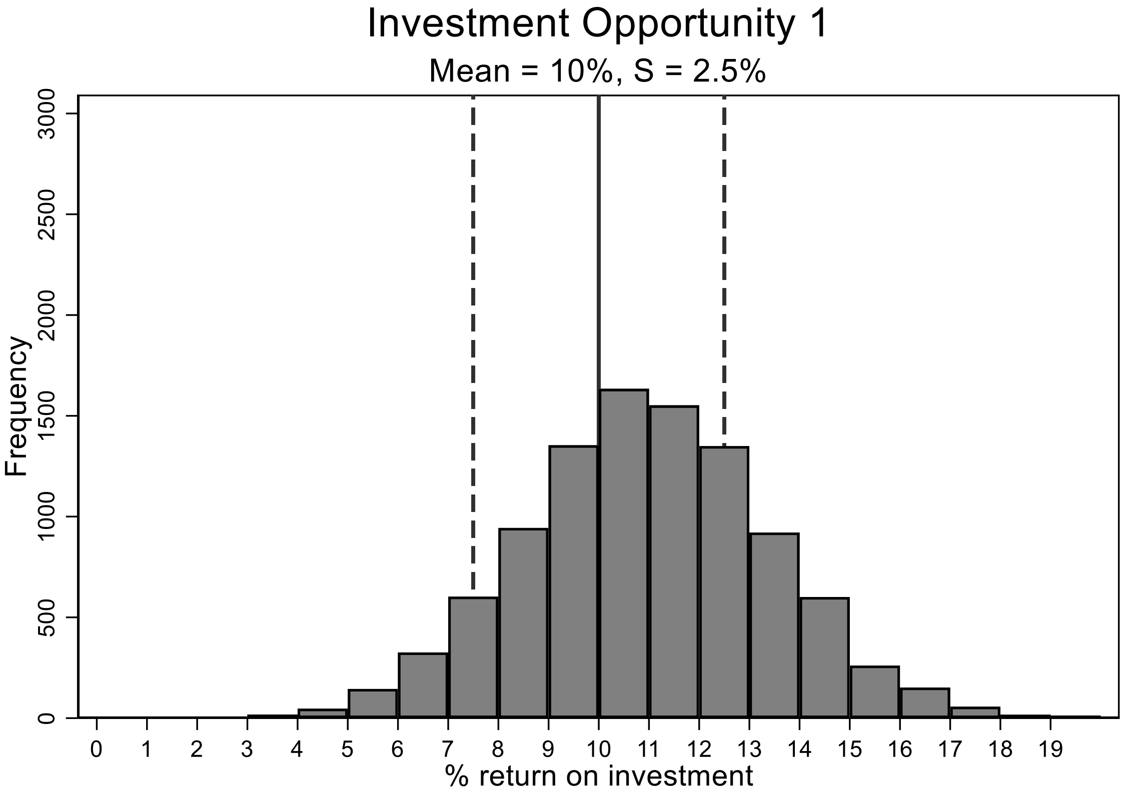
Range = difference between maximum and minimum values

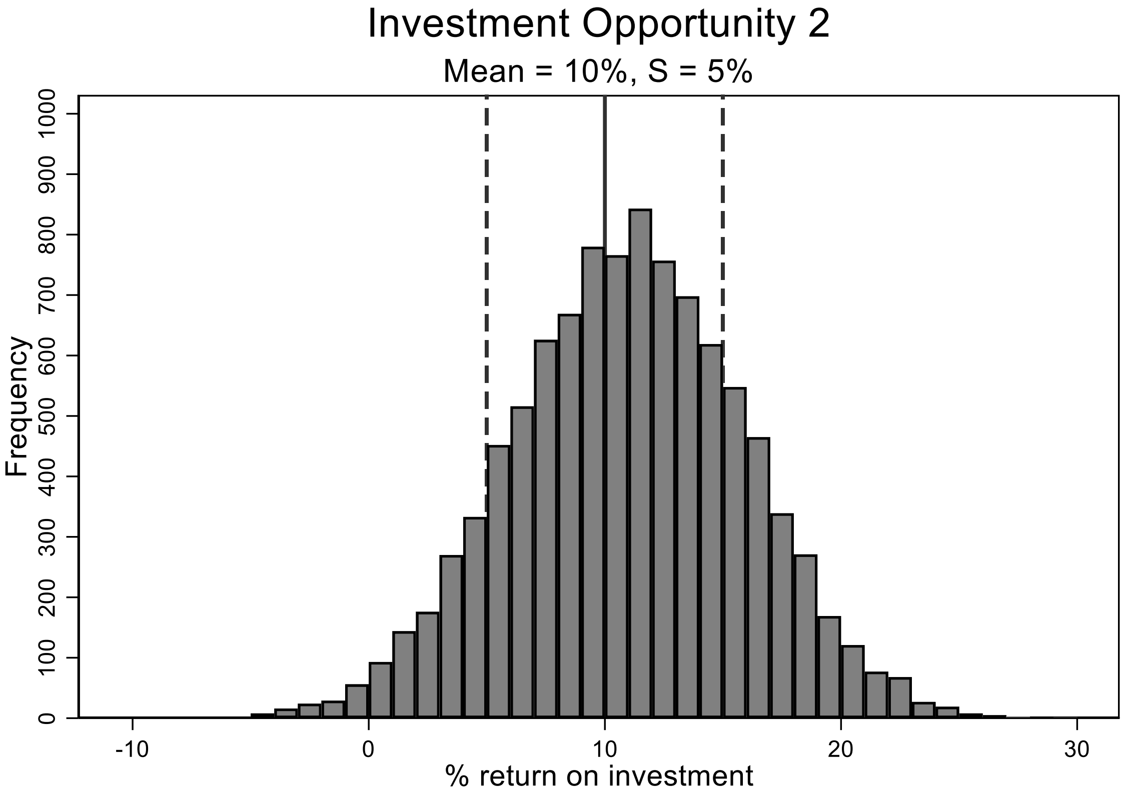
= 88 – 16 = 72

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





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**

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)

Multivariate

Multivariate analysis

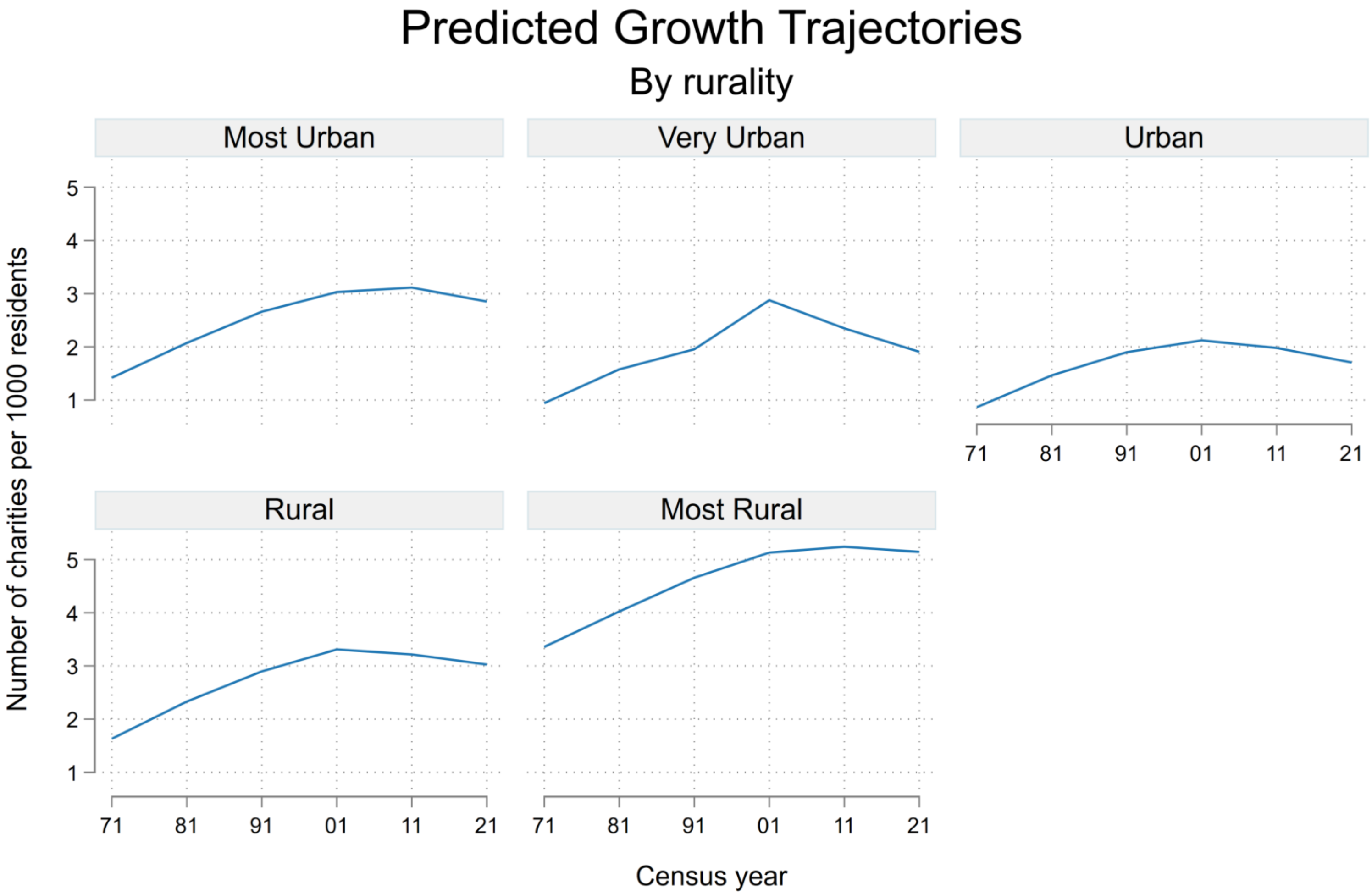
Multivariate analysis is concerned with testing whether the bivariate analyses vary across values of a third / fourth / fifth etc variable.

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?

Multivariate analysis


Predicted Growth Trajectories
By rurality
Most Urban
Very Urban
Urban
Rural
Most Rural
Number of charities per 1000 residents
Census year