Why Do We Need Statistics?

Types of Data Analysis

- Quantitative Methods
  - Testing theories using numbers
  - experimental or observational
- Qualitative Methods
  - Testing theories using language / other representations
  - usually observational

The Research Process

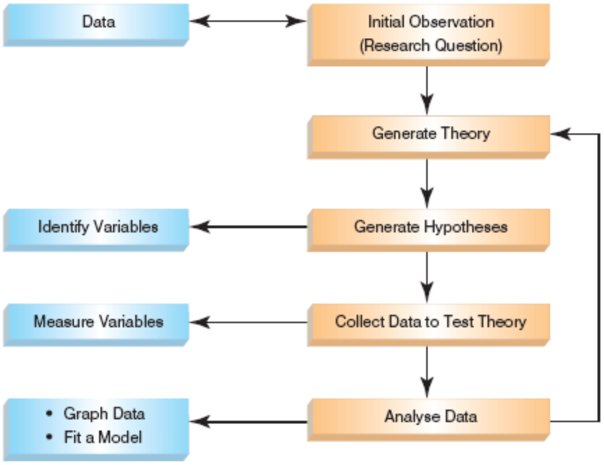


FIGURE 1.2 The research process

Initial Observation

- Find something that needs explaining
  - Observe the real world
  - Read other research
- Test the concept: collect data
  - Collect data to see whether your hunch is correct
  - To do this you need to define variables
    - Anything that can be measured and can differ across entities or time.

The Research Process

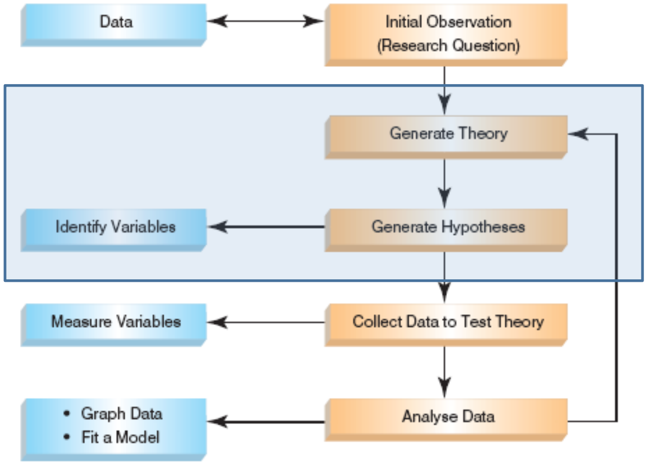


FIGURE 1.2 The research process

Generating and Testing Theories

- Theory
  - A hypothesized general principle or set of principles that explains known findings about a topic and from which new hypotheses can be generated.
- Hypothesis
  - A prediction from a theory.
  - E.g. the number of people turning up for a *Big Brother* audition that have narcissistic personality disorder will be higher than the general level (1%) in the population.
- Falsification
  - The act of disproving a theory or hypothesis.

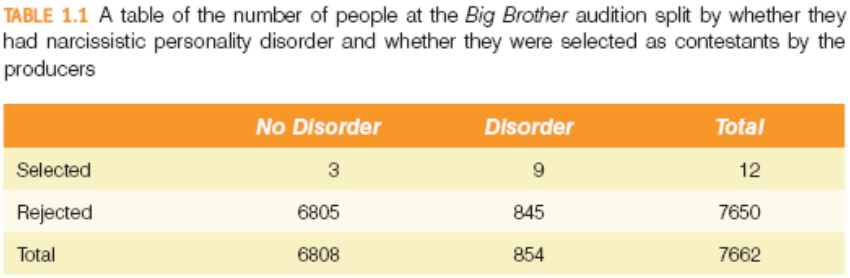


TABLE 1.1 A table of the number of people at the *Big Brother* audition split by whether they had narcissistic personality disorder and whether they were selected as contestants by the producers

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

The Research Process

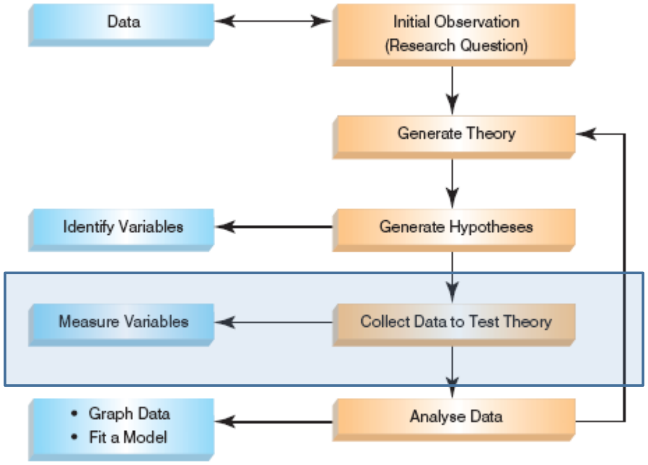


FIGURE 1.2 The research process

Data Collection 1: What to Measure?

- Hypothesis:
  - *Coca-Cola kills sperm*.
- Independent Variable
  - The proposed cause
  - A predictor variable
  - A manipulated variable (in experiments)
  - Coca-Cola in the hypothesis above
- Dependent Variable
  - The proposed effect
  - An outcome variable
  - Measured not manipulated (in experiments)
  - Sperm in the hypothesis above

Levels of Measurement

- Categorical (entities are divided into distinct categories):
  - Binary variable: There are only two categories
    - e.g. dead or alive.
  - Nominal variable: There are more than two categories
    - e.g. whether someone is an omnivore, vegetarian, vegan, or fruitarian.
  - Ordinal variable: The same as a nominal variable but the categories have a logical order
    - e.g. your place in a race
- Continuous (entities get a distinct score):
  - Interval variable: Equal intervals on the variable represent equal differences in the property being measured
    - e.g. the difference between 6 and 8 is equivalent to the difference between 13 and 15.
  - Ratio variable: The same as an interval variable, but the ratios of scores on the scale must also make sense
    - e.g. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8.

Measurement Error

- Measurement error
  - The discrepancy between the actual value we're trying to measure, and the number we use to represent that value.
- Example:
  - You (in reality) weigh 80 kg.
  - You stand on your bathroom scales and they say 83 kg.
  - The measurement error is 3 kg.

Validity

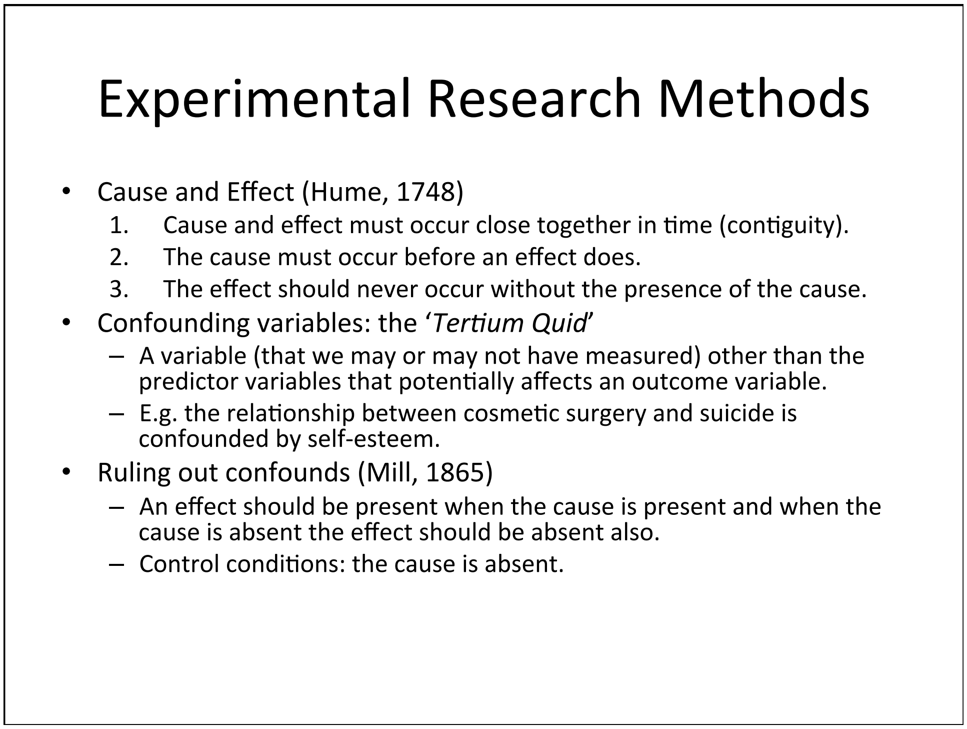
- Whether an instrument measures what it set out to measure.
- Content validity
  - Evidence that the content of a test corresponds to the content of the construct it was designed to cover
- Ecological validity
  - Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions.

Reliability

- Reliability
  - – The ability of the measure to produce the same results under the same conditions.
- Test–Retest Reliability
  - – The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

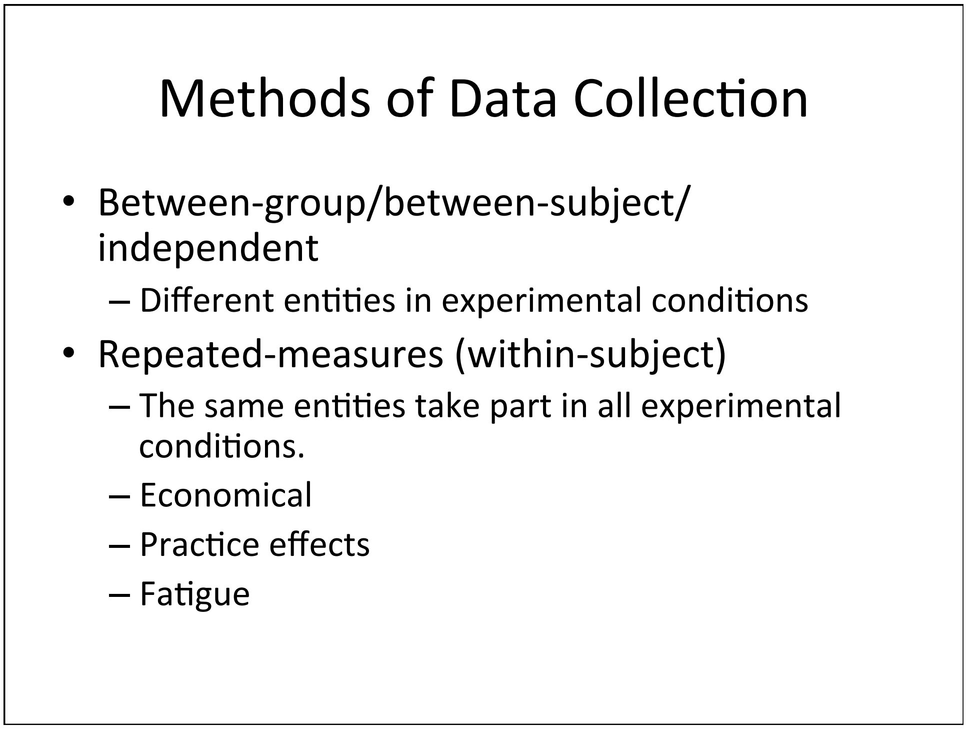
Data Collection 2: How to Measure

- Correlational/observational research:
  - – Observing what naturally goes on in the world without directly interfering with it.
- Experimental research:
  - – One or more variables is systematically manipulated to see their effect (alone or in combination) on an outcome variable.
  - – Statements can be made about cause and effect.



Experimental Research Methods

- Cause and Effect (Hume, 1748)
  1. Cause and effect must occur close together in time (contiguity).
  2. The cause must occur before an effect does.
  3. The effect should never occur without the presence of the cause.
- Confounding variables: the '*Tertium Quid*'
  - A variable (that we may or may not have measured) other than the predictor variables that potentially affects an outcome variable.
  - E.g. the relationship between cosmetic surgery and suicide is confounded by self-esteem.
- Ruling out confounds (Mill, 1865)
  - An effect should be present when the cause is present and when the cause is absent the effect should be absent also.
  - Control conditions: the cause is absent.



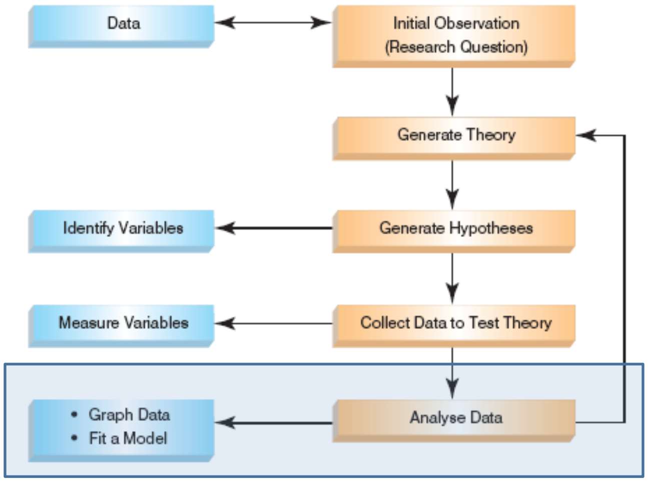
Methods of Data Collection

- Between-group/between-subject/ independent
  - Different entities in experimental conditions
- Repeated-measures (within-subject)
  - The same entities take part in all experimental conditions.
  - Economical
  - Practice effects
  - Fatigue

Types of Variation

- Systematic Variation
  - Differences in performance created by a specific experimental manipulation.
- Unsystematic Variation
  - Differences in performance created by unknown factors.
    - Age, gender, IQ, time of day, measurement error, etc.
- Randomization
  - Minimizes unsystematic variation.

The Research Process



Data

Initial Observation (Research Question)

Generate Theory

Identify Variables

Generate Hypotheses

Measure Variables

Collect Data to Test Theory

- Graph Data
- Fit a Model

Analyse Data

FIGURE 1.2 The research process

Analysing Data: Histograms

- Frequency Distributions (aka Histograms)
  - A graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set.
- The 'Normal' Distribution
  - Bell-shaped
  - Symmetrical around the centre

The Normal Distribution

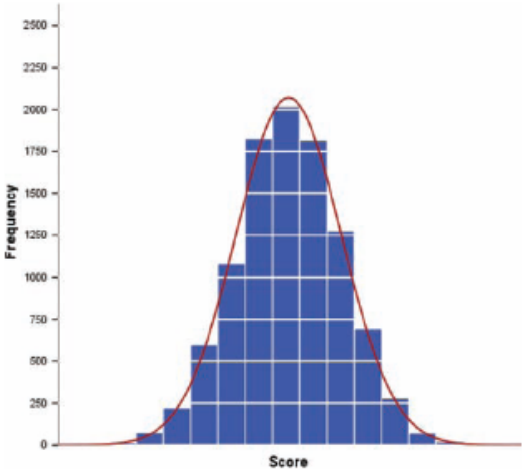


FIGURE 1.3 A 'normal' distribution (the curve shows the idealized shape)

Properties of Frequency Distributions

- Skew
  - The symmetry of the distribution.
  - Positive skew (scores bunched at low values with the tail pointing to high values).
  - Negative skew (scores bunched at high values with the tail pointing to low values).
- Kurtosis
  - The 'heaviness' of the tails.
  - Leptokurtic = heavy tails.
  - Platykurtic = light tails.

Skew

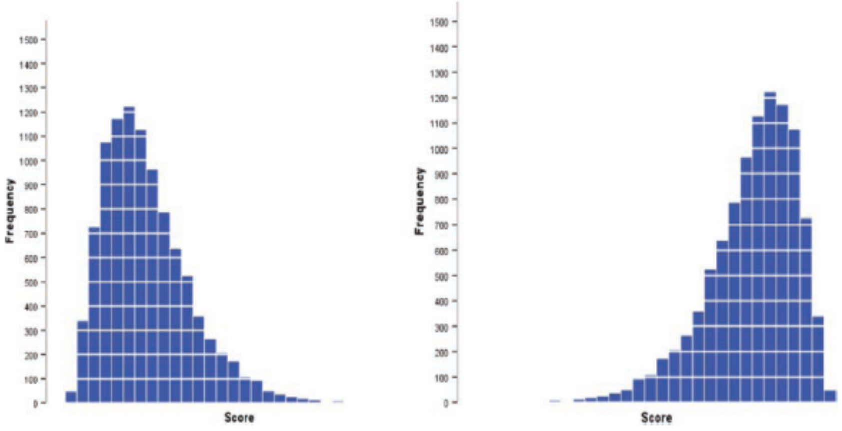


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

Kurtosis

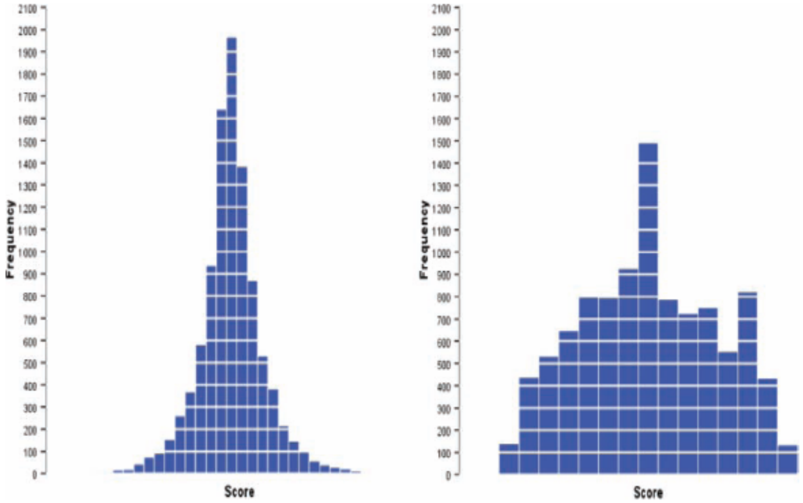


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

Central tendency: The Mode

- Mode
  - The most frequent score
- Bimodal
  - Having two modes
- Multimodal
  - Having several modes

A Bimodal Distribution

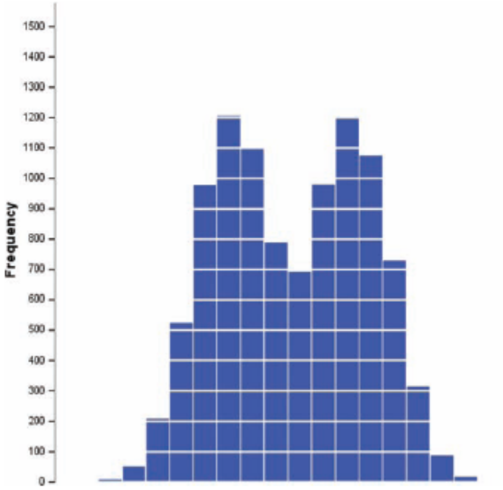
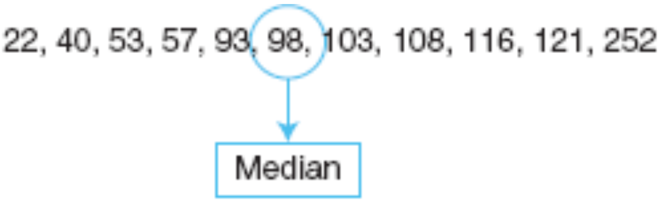


FIGURE 1.6 A bimodal distribution

Central Tendency: The Median

- Median
  - The middle score when scores are ordered.
- Example
  - Number of friends of 11 Facebook users.



22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252

Median

Central Tendency: The Mean

- Mean
  - – The sum of scores divided by the number of scores.
  - – Number of friends of 11 Facebook users.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

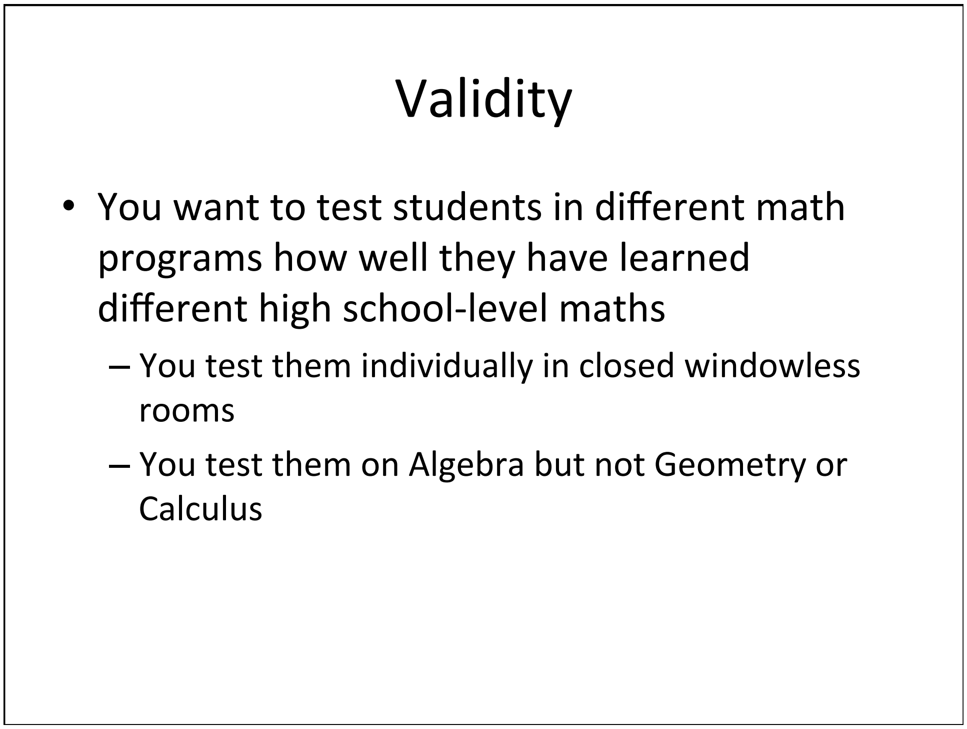
$$\sum_{i=1}^{n} x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 252$$

$$= 1063$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1063}{11} = 96.64$$

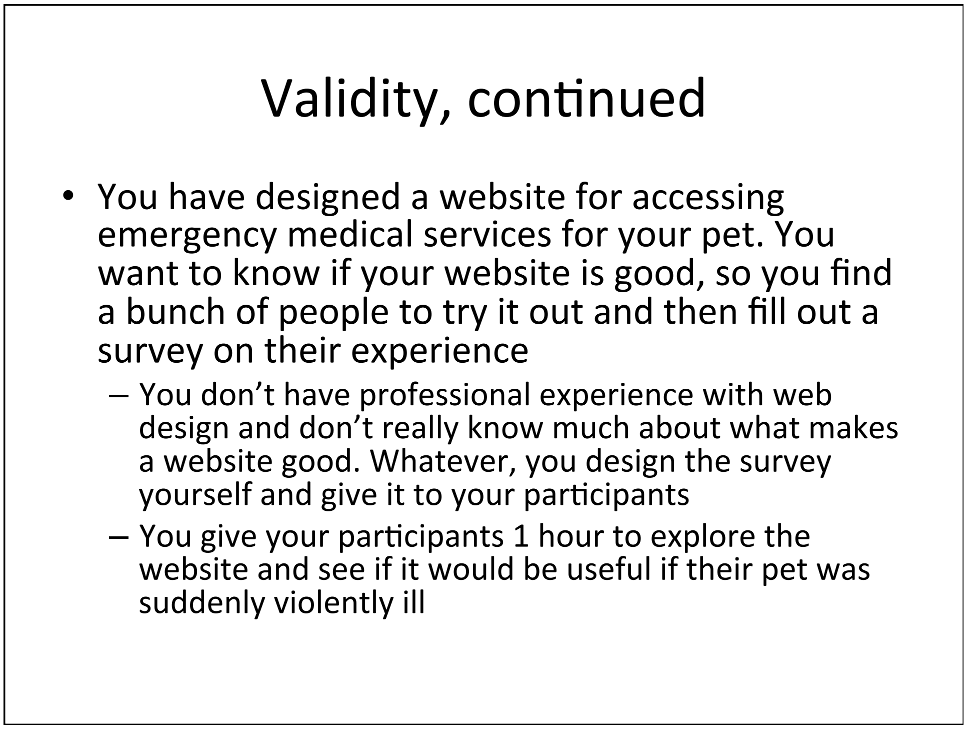
February 9

- Review some terminology from last week and add examples
  - Validity
  - Reliability
  - Tertium Quid
  - Variability
  - Randomization
- Introduction to RStudio and R
- Remember: first practical session tomorrow!
  - – activity & guide will be uploaded before tomorrow
  - – for this week: hw is what you don’t finish
  - – how many people can bring their own devices?



Validity

- You want to test students in different math programs how well they have learned different high school-level maths
  - You test them individually in closed windowless rooms
  - You test them on Algebra but not Geometry or Calculus



Validity, continued

- You have designed a website for accessing emergency medical services for your pet. You want to know if your website is good, so you find a bunch of people to try it out and then fill out a survey on their experience
  - You don't have professional experience with web design and don't really know much about what makes a website good. Whatever, you design the survey yourself and give it to your participants
  - You give your participants 1 hour to explore the website and see if it would be useful if their pet was suddenly violently ill

Reliability

- How is validity different from reliability?
- Examples of unreliability
  - an IQ test gives different results for the same person at two different time points
  - an fMRI study shows activation in completely different areas of the brain for different participants doing the same task

Tertium Quid

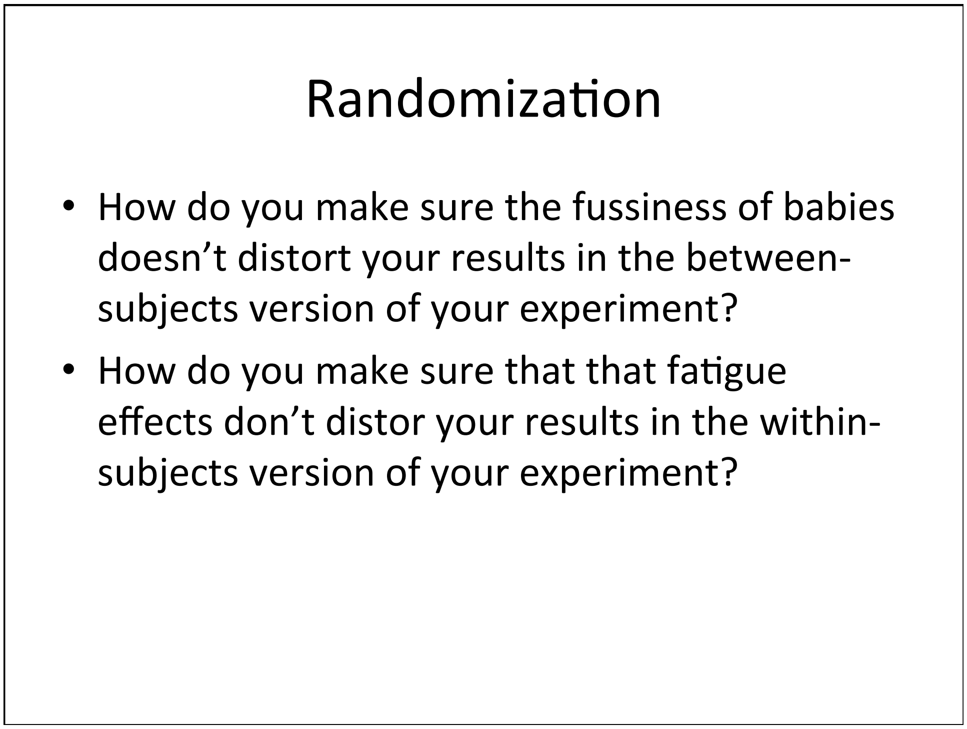
- Murder rates correlated with ice cream consumption
- Does eating ice cream make you commit murder?
- Do murderers crave ice cream?
- Confound variable: the heat

“Between” vs. “Within” designs

- You want to test a new magical baby food to see if it makes babies fall asleep. First, you give the babies the baby food and see what happens. Later, you give the babies a placebo and see what happens.
- Alternatively, you get two groups of babies. You give one group the baby food and the other group the placebo

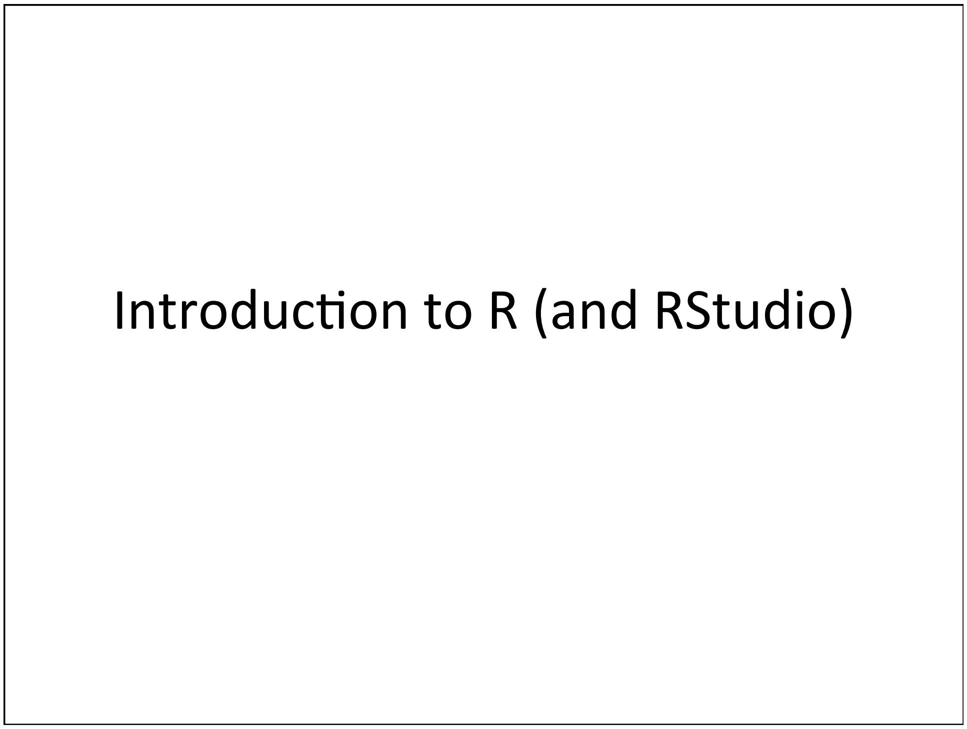
Variability

- systematic versus unsystematic variability
- In the between-subjects version of your baby food experiment, you find that some of the babies given the food didn’t fall asleep. Their mothers describe them as “fussy”
- On average, the babies given the food sleep more readily than those who were not given the food
- In the within-subjects version of your baby food experiment, the babies get no food in the first condition, and then the magical food in the second condition. The results turn out great! But then when you start marketing the food, consumers report worse results than you found in your experiment. What happened?



Randomization

- How do you make sure the fussiness of babies doesn’t distort your results in the between-subjects version of your experiment?
- How do you make sure that that fatigue effects don’t distor your results in the within-subjects version of your experiment?



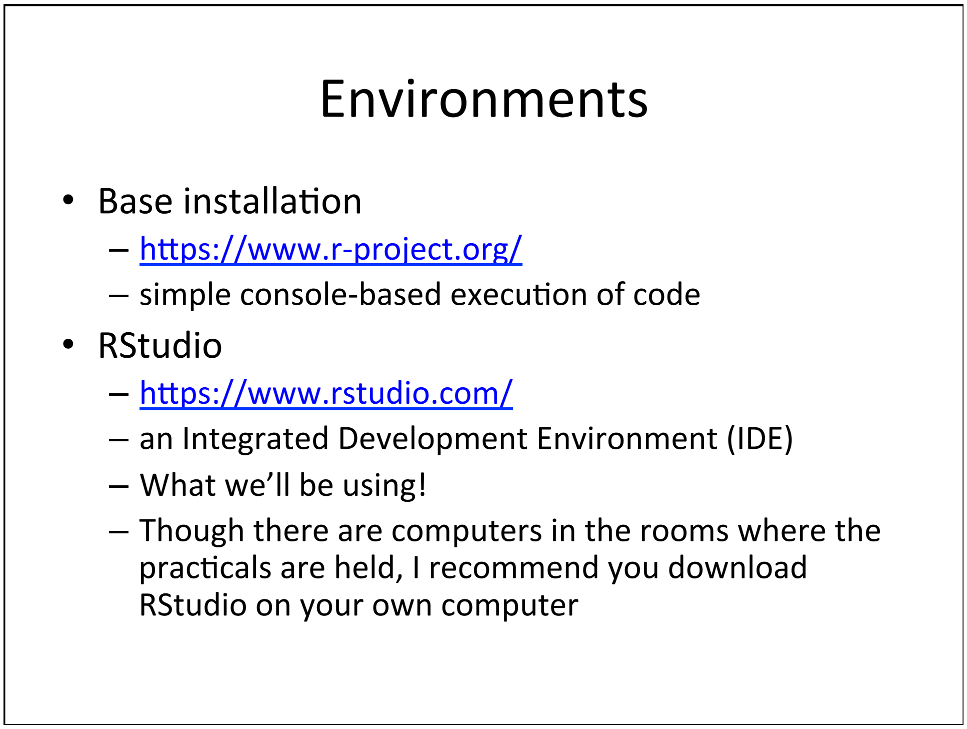
Introduction to R (and RStudio)

What is R?

- Programming language
  - interactive, unlike Java
  - widely used in statistical computing
    - open source
    - extensive statistical packages available
    - vectorization
  - though still general-purpose
    - control flow, function declaration, objects
  - similar syntax to other popular languages

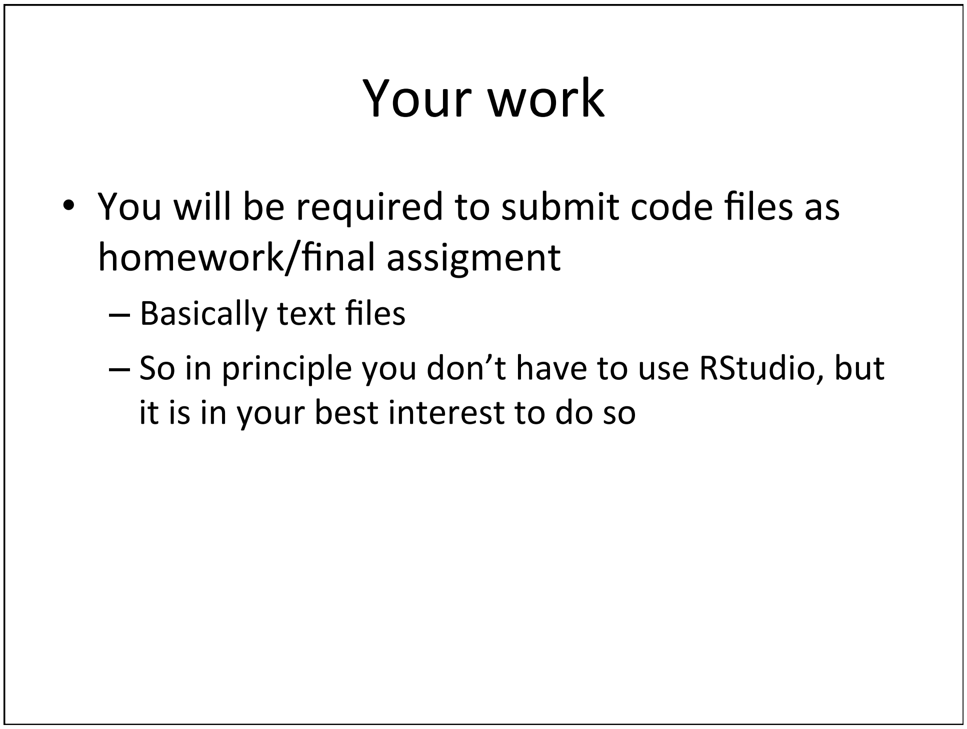
Why use a scripting language?

- Why not SPSS?
  - scripts offer MUCH more flexibility
  - this is AI
    - the more experience w/ programming & programming languages, the better



Environments

- Base installation
  - https://www.r-project.org/
  - simple console-based execution of code
- RStudio
  - https://www.rstudio.com/
  - an Integrated Development Environment (IDE)
  - What we'll be using!
  - Though there are computers in the rooms where the practicals are held, I recommend you download RStudio on your own computer



Your work

- You will be required to submit code files as homework/final assigment
  - Basically text files
  - So in principle you don't have to use RStudio, but it is in your best interest to do so

HOW WOULD YOU RATE YOUR COURSE PROGRAMME?

Take the 2017 National Student Survey!

www.nse.nl

Radboud University

Picnic with your sweetheart? Take part for a chance to win a picnic basket for two from Huize Heyendael!