The University Of Sheffield.

Human Centred Systems Design

Experimental Design

Dr Maria-Cruz Villa-Uriol

Evaluation in HCI

✓ Measuring Usability

Week 8 → Assignment hand-out

Tuesday
14th November
(week 8)

✓ Qualitative Research Methods

Week 9

⇨ Experimental Design

Week 10

Quantitative Research Methods

Week 11 → Assignment hand-in via MOLE

Monday
11th December
(week 12)

The University Of Sheffield.

Measuring Usability

- Good design is about ...
  - usability
  - ensuring that systems are *accessible* to all
  - ensuring that designs are *acceptable* for the people and contexts in which they will be used

- Designers need to **evaluate** their designs with people and involve people in the design process

- Access to interactive systems for all people is an important right

- Usability is concerned with balancing the PACT elements in a domain

- Acceptability is concerned with ensuring designs are appropriate to contexts of use

Types of Evaluation

- Automated testing

- Testing involving users in:
  - ‘Controlled settings’
    - usability testing
    - laboratory experiments
    - living labs
  - ‘Natural settings’
    - field studies
      (to see how the product is used in the real world)

- Testing not involving users:
  - ‘Experts’ critique, predict, analyse and model aspects of the interface
    - cognitive walkthrough
    - heuristic evaluation
    - review-based evaluation
    - model-based evaluation

Types of Evaluation

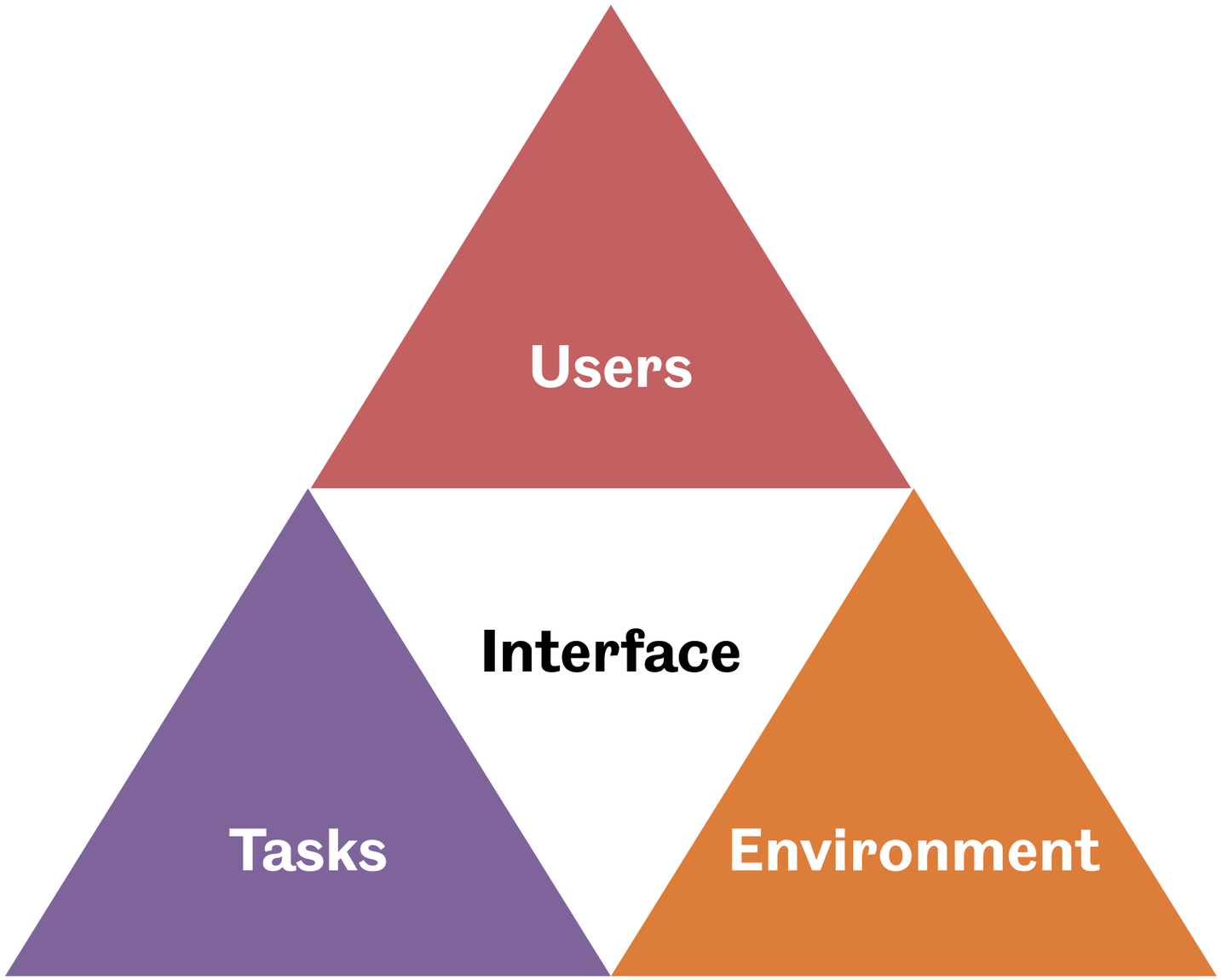
Method	Controlled Settings	Natural Settings	Without Users
Observing	✓	✓	
Asking users	✓	✓	
Asking experts		✓	✓
Usability testing	✓		
Modelling			✓

Choosing and Evaluation Method

❑ What stage in the life-cycle?	Design *vs.* implementation
❑ What style of evaluation?	Laboratory *vs.* field
❑ How objective?	Subjective *vs.* field
❑ What types of measures?	Qualitative *vs.* quantitative
❑ What level of information?	High level *vs.* low level
❑ What level of interference?	Obtrusive *vs.* unobtrusive
❑ What resources are available?	• Time • Subjects • Equipment • Expertise

Usability Testing

- Representative users attempting representative tasks in representative environments



The University Of Sheffield.

Usability Testing

- Representative users attempting representative tasks in representative environments

- This includes testing:
  - Paper prototypes
  - Screen mock-ups
  - Prototypes controlled via Wizard of Oz
  - Partially-functional software
  - Fully-working versions of software
  - ...

Qualitative Research Methods

- Qualitative vs Quantitative Data
- Qualitative Research Methods
- Traditionally, in Human-Computer Interaction:
  1. Surveys
  2. Interviews
  3. Focus groups
  4. Diaries
  5. Ethnographic research

Mostly descriptive!

The University Of Sheffield.

Qualitative research	vs.	Quantitative research
Discover ideas and gain insight and understanding	**Aim**	Test hypotheses and specific research questions
Observe, survey and interpret	**Approach**	Measure and test
Mixed	**Data collection**	Structured
Researcher involved and results subjective	**Researcher independence**	Researcher uninvolved observer, objective results
Small samples, naturalistic setting	**Sample size**	Larger samples for generalisable results

An example

- A researcher observes a class of teenagers and reports the following two observations:
    - 8 out of 10 teenagers who do play a particular computer game can touch type
    - 2 out of 12 teenagers who do not play the game can touch type

Empirical Research Methods

	Type of Research	Focus	General Claims	Typical Methods
1	**Descriptive**	Describe a situation or a set of events	X is happening	
2	**Relational**	Identify relations between multiple variables	X is related to Y	
3	**Experimental**	Identify causes of a situation or a set of events	X is responsible for Y	

The University Of Sheffield.

Empirical Research Methods

	Type of Research	Focus	General Claims	Typical Methods
1	**Descriptive**	Describe a situation or a set of events	X is happening	Observations Field studies Focus groups Interviews
2	**Relational**	Identify relations between multiple variables	X is related to Y	Observations Field studies Surveys
3	**Experimental**	Identify causes of a situation or a set of events	X is responsible for Y	Controlled experiments

The University Of Sheffield.

An example

- A researcher observes a class of teenagers and reports the following two observations:
  - 8 out of 10 teenagers who do play a particular computer game can touch type
  - 2 out of 12 teenagers who do not play the game can touch type

The University Of Sheffield.

Descriptive Research

Type of Research	Focus	General Claims	Typical Methods
Descriptive	Describe a situation or a set of events	X is happening	Observations Field studies Focus groups Interviews

- An example:
  - A researcher observes that 8 out of 10 teenagers in a class who play a particular computer game can touch type
  - But only 2 out of 12 teenagers in the same class who do not play the game can touch type
  - Interesting observations, but…
    - do these observations establish a relationship between game and typing?
    - do they offer any explanation?

Relational Research

Type of Research	Focus	General Claims	Typical Methods
Relational	Identify relations between multiple variables	X is related to Y	Observations Field studies Surveys

- Revisiting the previous example:
  - Record the number of hours the teenagers play the computer game each week AND measure their typing speed
  - Run a correlation analysis between number of hours and typing speed…
    - will the correlation analysis show a relationship between typing speed and time spent playing the game?
    - but will it show the cause of the relationship?

Some reflections…

- Possible reasons for the correlation
  a) playing game might improve typing speed
  b) teenagers who type well may like the game more
  c) teenagers who read well also type faster, and teenagers who read well like the game more and spend more time on it.
  d) …

Oops!

So game may have no impact on teenagers' typing speed

The University Of Sheffield.

Experimental Research

Type of Research	Focus	General Claims	Typical Methods
Experimental	Identify causes of a situation or a set of events	X is responsible for Y	Controlled experiments

- Revisiting the example once more:
  - We could design an experiment to demonstrate a causal effect between 2 factors
    - Recruit teenagers, and randomly assign to 2 groups
      - Group 1 required to play the game for X hours every week
      - Group 2 does not play the game
    - After 3 months, measure teenagers' typing speed

IF Group 1's typing speed is **significantly** faster than Group 2's… can we conclude that there is good reason to believe that playing the game improves typing speed?

What might be a major complication when designing experiments in HCI?

What might be a major complication when designing experiments in HCI?

US

(i.e. humans)

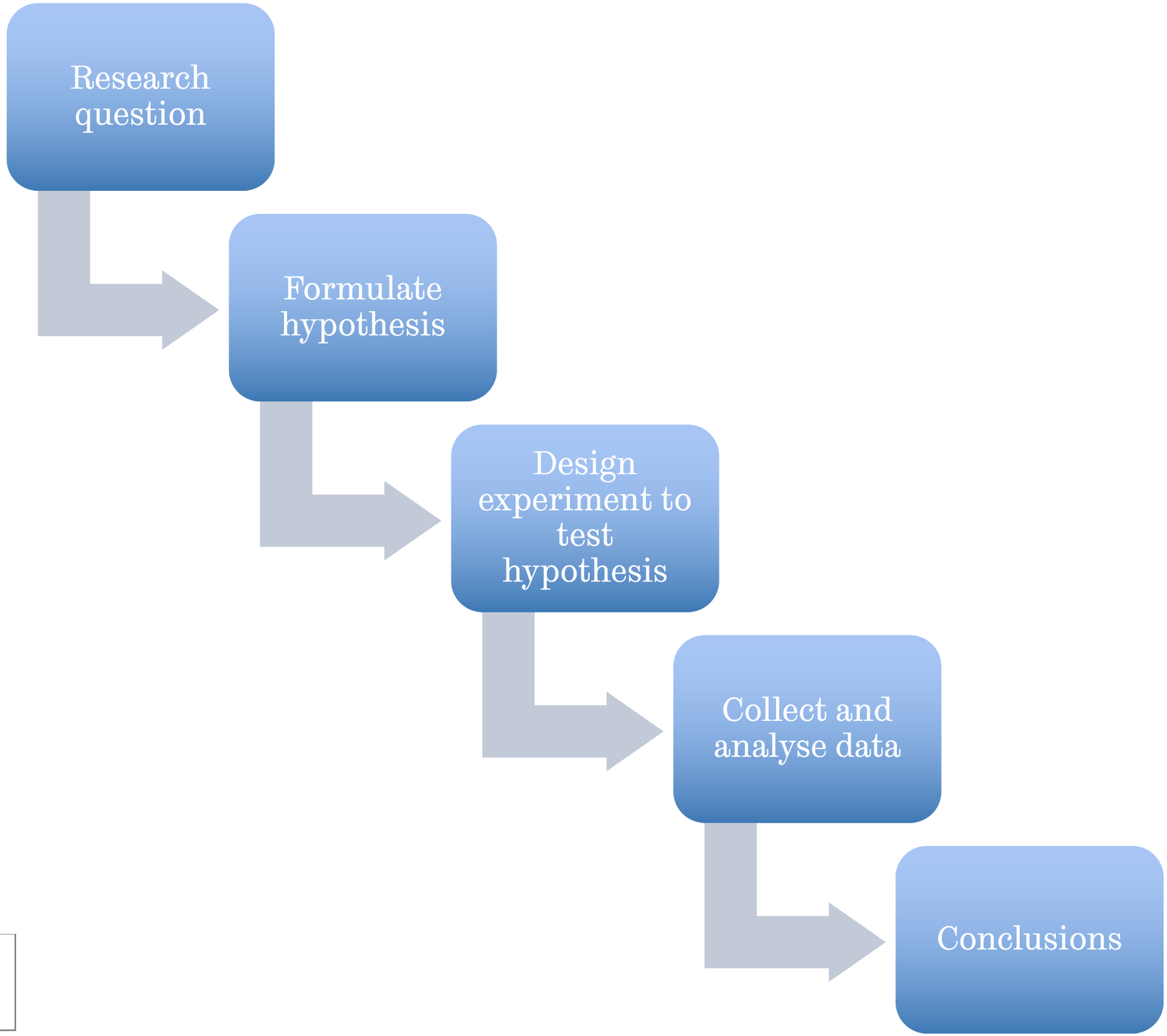
Experimental design in HCI

- Dealing with people complicates things…..
- People can be influenced by many different things
- Their answer could be affected by
  - Trying to please you
  - The colours you are wearing
  - The impression they want to create (moody and interesting?)
  - The predominant colours in the room
  - How they interpret your question

Research with humans requires careful planning and design:

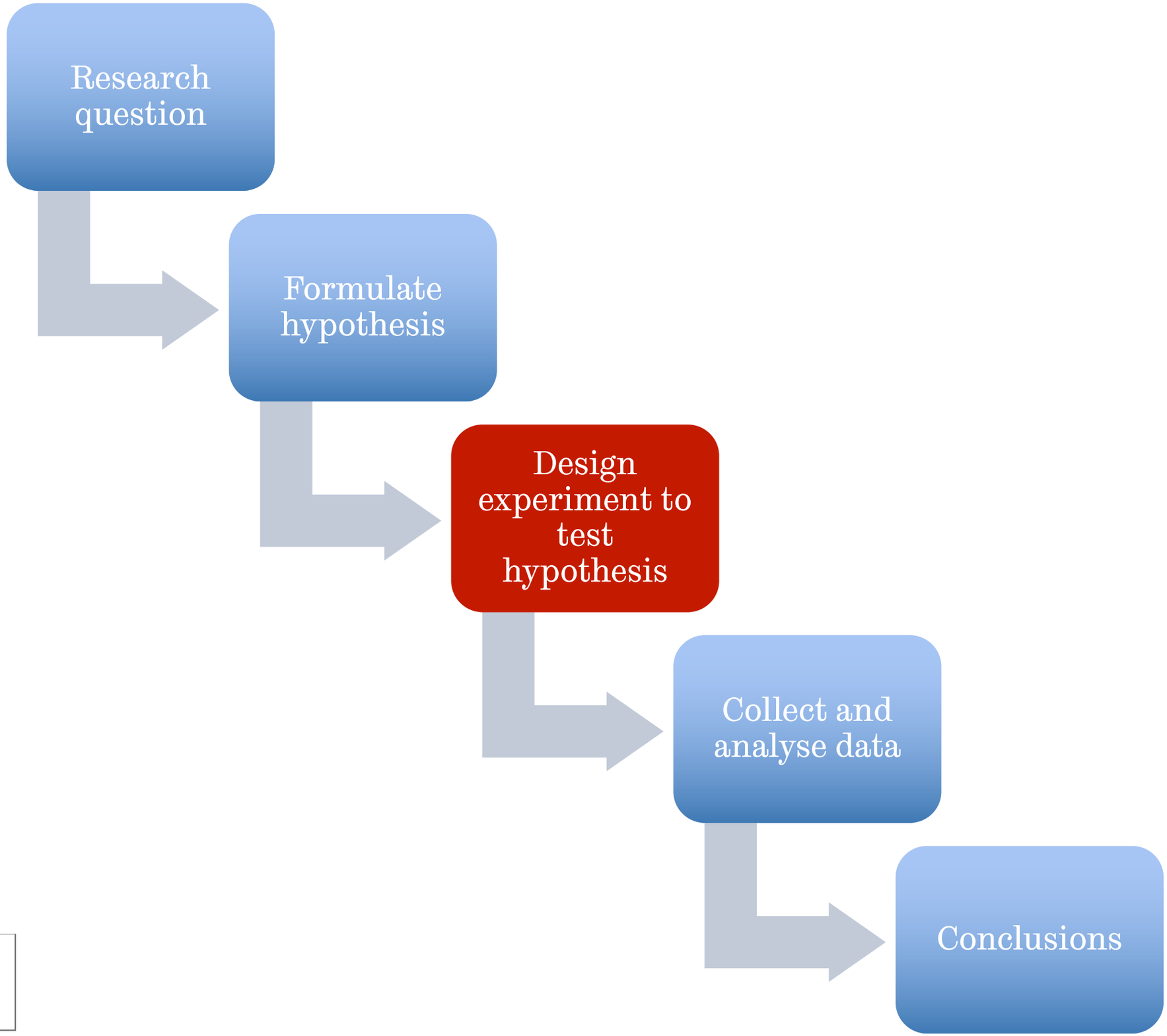
1. Need to identify a research question
2. Use the research question to formulate a testable hypothesis

Experimental research methods



The University Of Sheffield.

Experimental research methods



The University Of Sheffield.

Class experiment

Nine	Swap	Cell	Ring	Lust
Plugs	Lamp	Apple	Table	Sway
Army	Bank	Fire	Hold	Worm
Clock	Horse	Colour	Baby	Sword
Desk	Hold	Find	Bird	Rock

Write down all the words you can remember

Horse	Cat	Dog	Fish	Bird
Orange	Yellow	Blue	Green	Black
Table	Chair	Desk	Bookcase	Bed
Teacher	School	Student	Homework	Class
Apple	Banana	Kiwi	Grape	Mango

The University Of Sheffield.

Write down all the words you can remember

Class experiment (explanation)

- Short term memory:
  In 1956 George Miller reported that people can remember 7 + or – 2 items

- But interestingly, when items can be grouped meaningfully, (chunking), memory capacity increases.

George A. Miller (1956) “The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information”.

The University Of Sheffield.

Working Memory Capacity

The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information (1956).

Ready?

M U T G I K T L R S Y P

You should be able to recall 7±2 letters.

George A. Miller
1920-2012

Experimental Design: Key Aspects

- Formulate a testable hypothesis
  - identify the dependent and independent variables
- Reduce any bias
- Counteract any variables that can not be controlled

Formulation of hypothesis

- **Independent variables** in an experimental study
  - the factors that the researchers are interested in studying
  - the possible ‘cause’ of the change in the dependent variable
    - the treatments/conditions the researchers can control
- **Dependent variables**
  - the outcome or effect that the researchers measure

The University Of Sheffield.

Formulating a hypothesis: Example

Someone running for the bus, carrying bags

- A plausible <u>hypothesis</u>:
  - ‘the more bags you carry the slower you run’
- <u>Independent variable</u>:
  - the number of bags (1,2,3)
- <u>Dependent variable</u>:
  - the time taken
    to run a set distance

The University Of Sheffield.

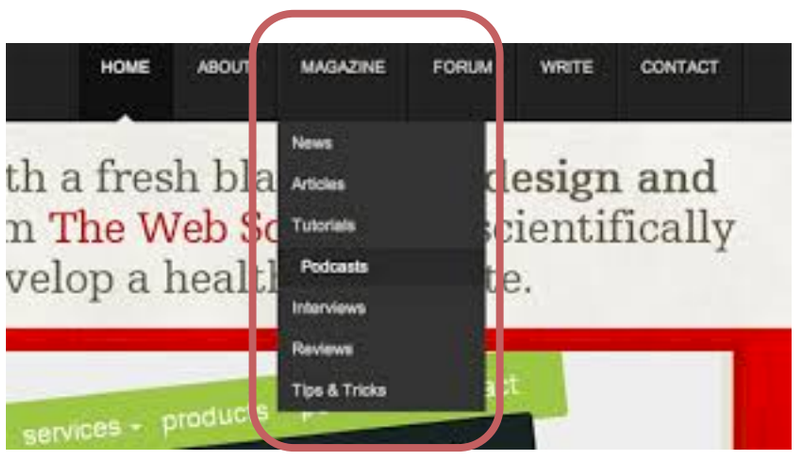
- Do manipulations of the independent variables cause changes in the dependent variables?

The University Of Sheffield.

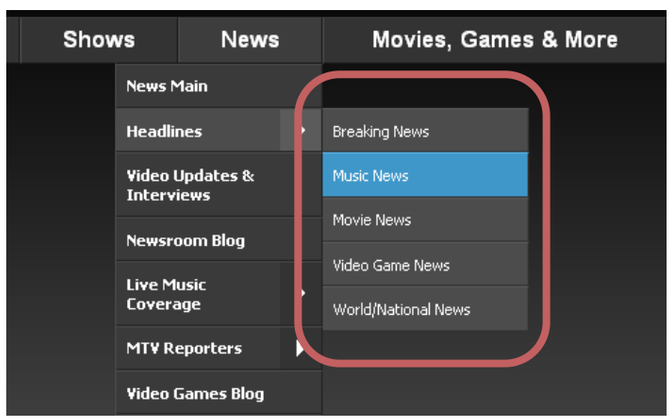
Formulating a hypothesis: Example

The developers of a website want to know if a pull-down menu or a pop-up menu is better

- Research question:
  - Which menu type is best for navigating the site?
- Independent variable:
  - the type of menu (pull-down or pop-up)
- Dependent variable:
  - time spent locating web pages



vs.



The University Of Sheffield.

Designing the experiment: Example

- Two alternatives are possible:

1. Recruit <u>two</u> groups of people:
   - Group 1 will see the version of the website with pop up menus,
   - Group 2 will see a version of the website with pull down menus

2. Recruit <u>only one</u> group of people:
   - All participants will experience both websites

Designing the experiment: Example

- Two alternatives are possible:

Between subjects design

1. Recruit <u>two</u> groups of people:
   - Group 1 will see the version of the website with pop up menus,
   - Group 2 will see a version of the website with pull down menus

Within subjects design

2. Recruit <u>only one</u> group of people:
   - All participants will experience both websites

Type of Experiment Design

	Between-group	Within-group
Advantages	Cleaner Avoids learning effect Better control of confounding factors, e.g. fatigue	Smaller sample size Effective isolation of individual differences More powerful tests available
Disadvantages	Larger sample size required Large impact of individual differences Harder to get statistically significant results	Hard to control learning effect Large impact of fatigue

Experiment Design

- **Between subjects design**
  - Different groups of people are allocated to different conditions
  - Need to ensure that the participants in different conditions are *similar*

- **Within subjects design**
  - Same group of people experience all the conditions
  - Need to ensure that order of conditions is *randomised*

- Important to try to eliminate systematic errors and possible sources of bias
  - uncontrolled variables that could affect the results

Class exercise 1

- **H1**:
  There is no difference in typing speed when using a QWERTY keyboard, a DVORAK keyboard, or an alphabetically ordered keyboard
- **H2**:
  There is no difference in the time required to locate an item in an online store between novice and experienced users
- **H3**:
  There is no difference in the perceived trust toward an online agent among customers who are from the United States, Russia, China, and Nigeria.

Discuss in pairs –
What kind of experiment type would you recommend to test each of these 3 hypotheses?

The University Of Sheffield.

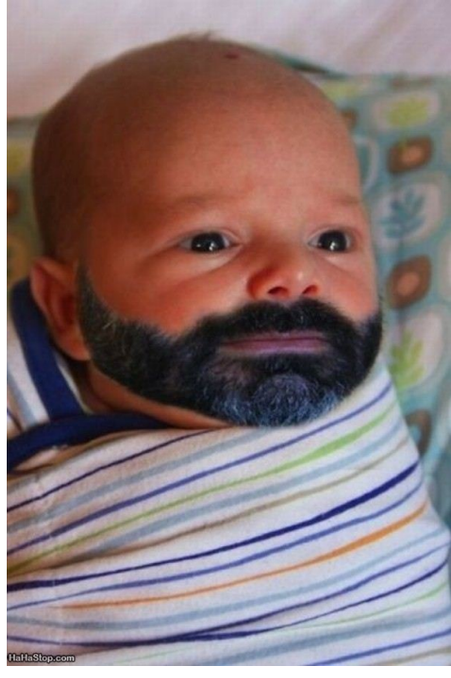
Class exercise 2

- Research question:
  - Does a diet rich in tomatoes increases life expectancy?
- Experiment design:
  - Two groups of 1000 people each :
    - Group 1: from an island in Greece, where the diet is traditionally rich in tomatoes
    - Group 2: from a religious sect based in the remote highlands of Chile which forbids the eating of tomatoes
  - After 10 years they followed up both groups :
    - Group 1: 132 people had died
    - Group 2: 68 people had died
- Conclusion:
  - Researchers concluded that eating tomatoes is associated with a lower life expectancy.

Discuss in pairs –
Is this conclusion likely to be robust and reliable?
What factors might bias the results?

The University Of Sheffield.

Class exercise 3



- A rumour spreads in the media that babies who are made to wear a false beard for 30 minutes each day become cleverer as a result.
- Research question:
  - Does wearing a false beard for 30 minutes each day (when you are a baby) makes that baby cleverer?
- Experiment design:
  - Two groups of 500 babies each :
    - Group 1: parents have started making their babies wear stick on beards
    - Group 2: parents do not make their babies wear stick on beards
  - Children are followed up and when they turn 4 years old :
    - Group 1 performed a 12% better on tests of preschool skills with respect Group 2
- Conclusion:
  - Researchers concluded that wearing false beards did have a significant impact on the children's development.

Discuss in pairs –
(a) are the conclusions likely to be reliable
(b) what could have biased the results
(c) how could the study be improved?

The University Of Sheffield.

Reliability of experiments

- Random errors
  - Also called ‘chance errors’ or ‘noises’
  - Cause variations in both directions
  - Occur by chance
  - Can be controlled by a large sample size
- Systematic errors
  - Also called ‘biases’
  - Always push actual value in the same direction
  - Can never be offset no matter how large the sample is

Random errors : Example

After observing a participant typing several text documents during 5 sessions, we obtained an actual typing speed of 50 words per minute

- Session 1: 46 words/min
- Session 2: 52 words/min
- Session 3: 47 words/min
- Session 4: 51 words/min
- Session 5: 53 words/min

The University Of Sheffield.

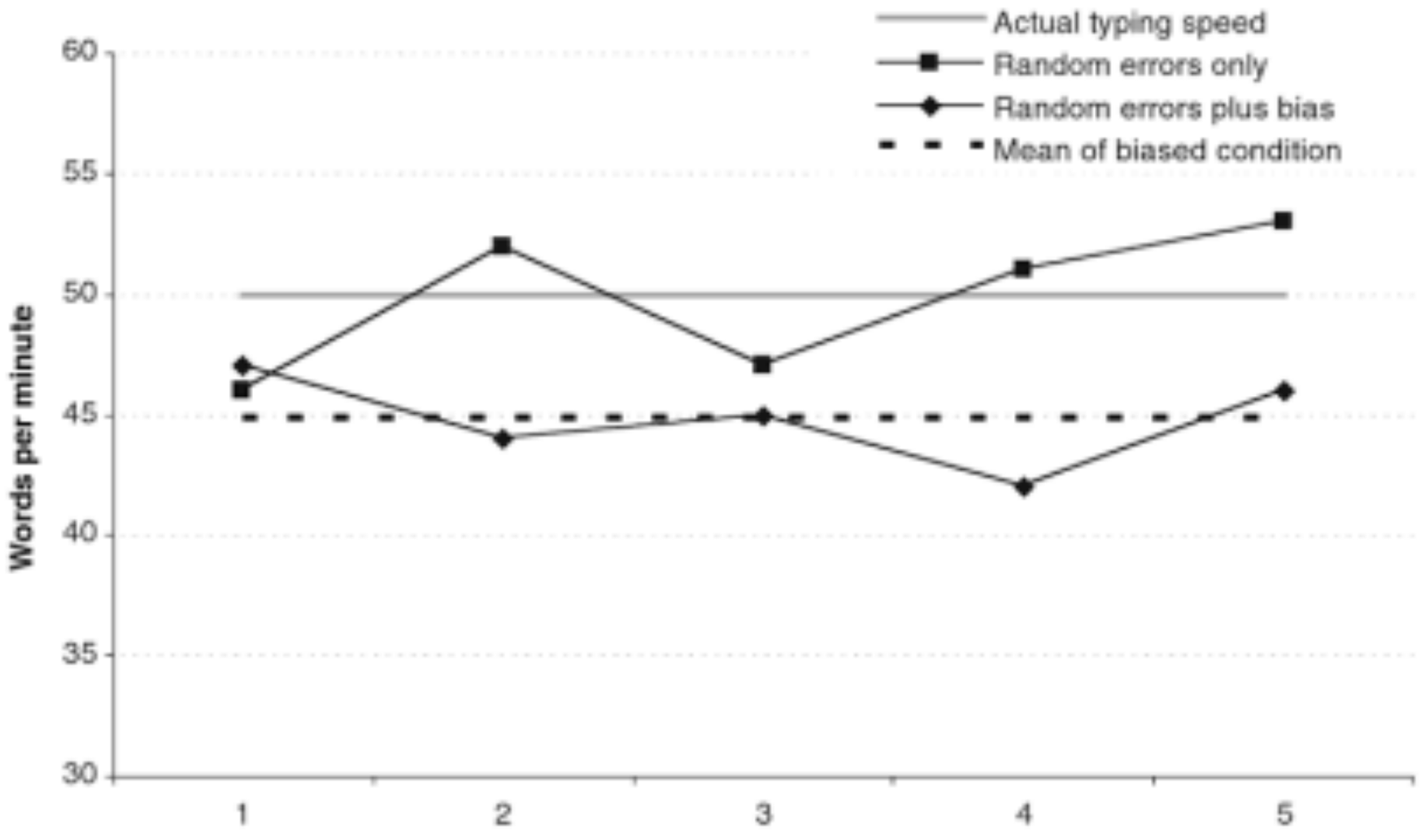
Systematic errors : Example

After observing a participant typing several text documents during 5 sessions, we obtained an actual typing speed of 44.8 words per minute
This participant was tired and nervous

- Session 1: 47 words/min
- Session 2: 44 words/min
- Session 3: 45 words/min
- Session 4: 42 words/min
- Session 5: 46 words/min

The University Of Sheffield.

Random *vs.* systematic errors



The University Of Sheffield.

Systematic errors

- Most common sources
  - Bias caused by measurement instruments
  - Bias caused by experimental procedures
  - Bias caused by participants
  - Bias caused by experimental behaviour
  - Bias caused by the experimental environment

Examples of systematic errors

- Bias caused by measurement instruments
  - e.g. using an inaccurate stop watch to record times
- Bias caused by experimental procedures
  - e.g. not randomising the order of task conditions
  - not using the same instructions for all participants

Examples of systematic errors

- Bias caused by participants
  - e.g. recruiting people who are more knowledgeable than the general public about the task

Examples of systematic errors

- Bias due to experimenter behaviour
  - e.g. An experimenter is introducing an interface to a participant. The experimenter says “Now you get to the pull-down menus. I think you will really like them …I designed them myself!”
  - Experimenter should be as neutral as possible

Examples of systematic errors

- Bias due to environmental factors
  - social environment
    - e.g. someone performing a task with a person watching them may perform differently to someone working on their own
  - physical environment
    - e.g. a speech recognition study when there is a lot of background noise

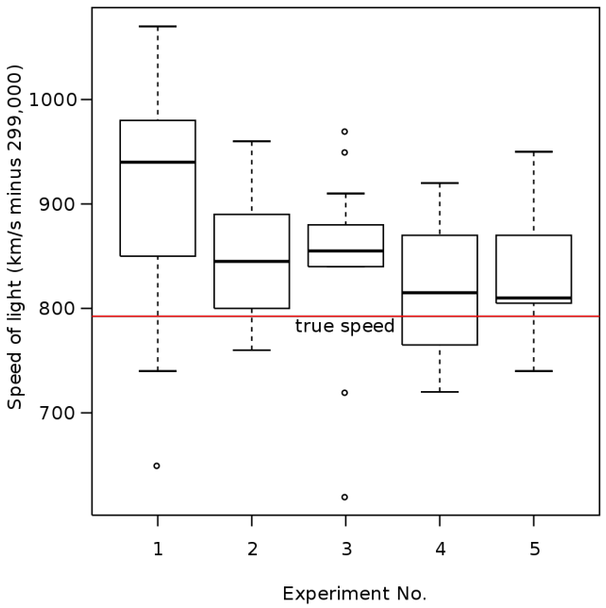
Quantitative methods

- Descriptive statistics
- Inferential statistics

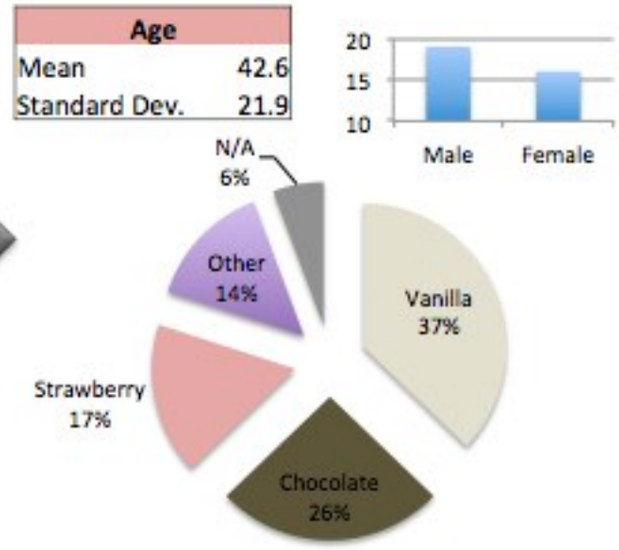
Descriptive statistics

- Describes a set of data with summary charts and tables.
  - No attempt to draw conclusions about the population from which the sample was taken.
  - A bit like telling someone the key points of a book (executive summary) instead of handing them the book itself (raw data)

Descriptive statistics



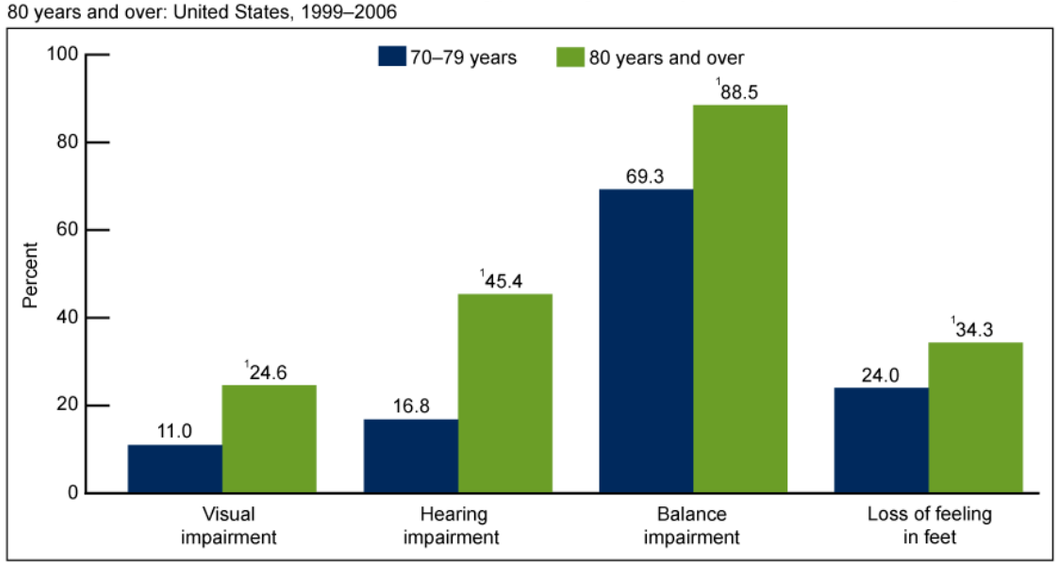
	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15	14	55	m	Vanilla
16	15	42	f	Chocolate
17	16	18	f	Vanilla



Raw Data

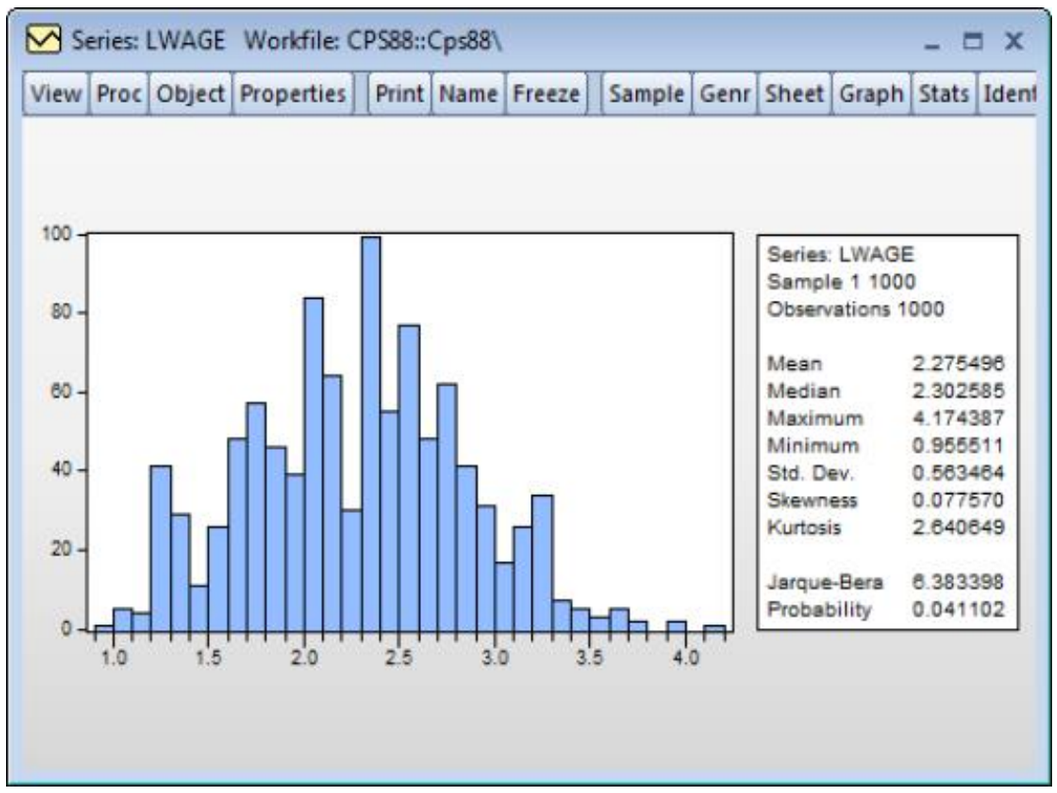
Descriptive Statistics

Figure 3. The prevalence of sensory impairments among persons aged 70–79 years compared with persons aged 80 years and over: United States, 1999–2006



[1]Significantly different from the 70–79 age group.

SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey.



The University Of Sheffield.

Inferential statistics

- Tests a hypothesis and draws conclusions about a population based on the sample
- Uses tests of significance
- Tries to draw conclusions that are robust and reliable

- Major difficulty: Producing robust findings!

The University Of Sheffield.

Null hypothesis testing

- 1920s, Ronald Fisher a statistician formulated the idea of null hypothesis testing

The University Of Sheffield.

Example: Tossing a coin

- Toss a coin 10 times and record the order of heads and tails

- e.g.
  - HTHHTHHHTH

The University Of Sheffield.

Example: Tossing a coin

- Which sequence is more likely to come up?
  - HHHHHHHHHH?
  - TTTTTTTTTT?
  - HTTHHTTHHT?

The University Of Sheffield.

Some reflections

- How can we prove that a coin is biased, given the outcome above?

- We can't
  - But we can estimate the probability that if the coin is not biased that we would see the given outcome

The University Of Sheffield.

Example: A box of crayons

- Same idea
  - Box with 2 crayons in it, one blue and one red.
  - You keep taking a crayon out and replacing it.
  - The crayons you take out are always red ones.
  - After doing this 20,000 times can you be certain that there is no blue crayon in the box?

The University Of Sheffield.

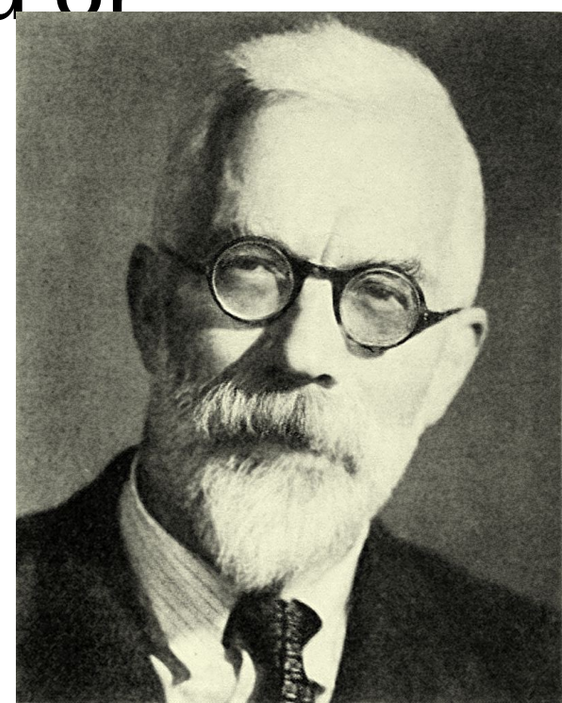
Example: Holtodol drug

- Same idea
  - New drug for gout, Holtodol
  - Will it cause side effects?
  - Give it to one user – no side effects
  - Give it to 10,000 people – no side effects
  - Can we conclude there will never be any side effects?

Null hypothesis

- Framework for statistical analysis
- Approach is the called the Null Hypothesis
- Ronald Fisher, a statistician in the 1920s – developed significance testing approach called the Null Hypothesis and popularised the idea of p-values

- The Null Hypothesis (that there is no difference between two sets of data) can be rejected or disproved.



The University Of Sheffield.

Null hypothesis

- H0 – The hypothesis that nothing has changed
  - E.g. That our actions have not had a corresponding effect

- We can never prove the null hypothesis
  - We may find evidence to reject it

Null hypothesis

- Going for the simplest option, that there is no relationship between two measures, giving people a new drug won't affect them

Null hypothesis: Example

- Research question: do dogs eat bananas?
- Experimental hypothesis: dogs eat bananas
- Null hypothesis: dogs do not eat bananas

- Null hypothesis is tested –

It can be rejected if a dog is seen eating a banana

The University Of Sheffield.

Null hypothesis: Example

- We believe that people will be able to remember more words from a list when
  (a) they are organised into meaningful groups than when (b) they are presented randomly.

- <u>Null Hypothesis</u>:
  - There is no difference in the number of words people can remember under condition (a) and condition (b)
- <u>Experimental Hypothesis</u>:
  - People will remember more words in condition (a) than in condition (b)

The University Of Sheffield.

Summary

- Empirical research methods
  - Descriptive *vs* relational *vs* experimental
- Experiment design in HCI
  - Research question → Formulation of hypothesis → Design experiment → Collect and analyse data → Conclusions
- Types of experiments
  - Between subjects *vs* within subjects
- Reliability of experiments:
  - random *vs* systematic errors

Summary

- Quantitative methods
  - Descriptive statistics
  - Inferential statistics
- Null hypothesis