Week 1: Introduction to Research Methods

Lecture notes: cautionary tale of Simpson's paradox

Introduction to Simpson's paradox

- **Simpson's paradox**: phenomenon in probability and statistics whereby a trend occurs in several different groups of data but then disappears or reverses when these groups are combined.
- **Berkeley postgraduate admissions** (1973):
    - In the past, people investigated the gender balance of Berkeley's admissions.
    - It was discovered that there were a lot more men than women, with too much data supporting this for it to be pure chance.

	Applicants	Admitted
Men	8442	**44%**
Women	4321	**35%**

-
    - The University of California, Berkeley, set out to further investigate the culprits for supposedly gender discrimination after the data raised a lot of eyebrows.
    - They did this by breaking open the data into different departments to see which ones were responsible for gender bias.
    - Interestingly, they found that out of six departments, four of them accepted more women than men – there was a gender bias, but it was in favour for women.

Department	**# of men**	**# of women**	**Men accepted**	**Women accepted**
A	825	108	62%	**82%**
B	560	25	63%	**68%**
C	325	593	**37%**	34%
D	417	375	33%	**35%**
E	191	393	**28%**	24%
F	393	341	6%	**7%**
Total	8442	4321		

-
    - Therefore, the aggregated data told a different story from the ungrouped data – classic case of *Simpson's paradox*: when grouped-up data demonstrates the opposite trend of the ungrouped data.
    - The truth was that women were not being discriminated against. Rather, a large proportion of them were applying to a low-acceptance rate department

while a large proportion of men were applying to a high-acceptance rate department, resulting in skewed overall results.

Lessons for potential researchers

- **Data** can be sneaky:
  - *Aggregated data*: refers to the overall data by combining multiple groups of data. Can show a bias in one direction.
  - *Disaggregated data*: refers to the separate data of different data groups, which can show no bias or a bias in an opposite direction from the aggregated data.
- **Statistics** and good data analysis:
  - Helps keep researchers on the right track.
  - Reduces the chances of researchers drawing the wrong conclusions from data.
- **Psychological research**:
  - Important to understand research methods.
  - Important to understand data analysis.

Video notes: introduction to R

Operators

- **R** is a statistical programming language used to:
  - Perform basic calculations.
  - Run statistical analyses.
  - Draw graphs.
  - Write programs.
  - Etc.
- **Pros of R**:
  - Open source and costs nothing.
  - Very powerful for statistics.
  - Rapidly becoming the most popular data analysis tool.
  - An introduction to programming, which is a valuable skill.
- **Operators**:
  - Used to carry out a particular kind of operation.
  - *Numerical operators*: used to carry out simple calculations.
  - *Logical operators*: used to provide a TRUE or FALSE response or for more complex comparisons.

Operator	**Type**	**Description**	**Example**
+	Numerical	Addition	5+2
-	Numerical	Subtraction	5-2
*	Numerical	Multiplication	2*2

/	Numerical	Division	8/2
^	Numerical	Power	3^3
==	Logical	Equality	1+1==2
!=	Logical	Inequality	1+1!=3
>	Logical	Greater than	5>3
<	Logical	Less than	5<8
>=	Logical	Greater than or equal to	5>=5
<=	Logical	Less than or equal to	3<=3
&	Logical	And	
|	Logical	Or	
!	Logical	Not	

Functions

- **Functions**:
  - Involve most of the other things that are not operators as there are not enough symbols on the keyboard to perform everything one might need to do.
  - Set of statements organised together to perform a specific task.

Function	Description	Example
sqrt()	Square root	sqrt(4)
round()	Round to nearest whole number	round(5.8)
log()	Logarithm	log(4)
exp()	Exponentiation	exp(4)
abs()	Absolute value	abs(-4)

- **Argument**:
  - Every function has this.
  - *Functions* can be thought of as recipes and *arguments* like ingredients, such that the recipe combines the right ingredients in a specific way.
  - *Arguments*: go within the brackets right after a function.

  - *Default values*: many arguments have these, which are used when the user does not tell R what value to use (e.g., the default number of digits to round to is zero).
- **Functions**:
  - Many can take more than one argument, which are separated with commas.
  - *Arguments*: most also have names and can be used when typing commands in any order (e.g., round(3.1415, 2) is the same as round(digits=2, x=3.1415).
  - *Equal signs*: only one (=) is used inside functions, while two (==) is used to compare two things.
- **Silent fail**: occurs when the input does not make sense, leading to the default value being used without warning.
- **Nesting functions**:
  - Just as a recipe can use the output of other recipes as ingredients, so too can functions use the output of other functions as arguments.
  - Hence, functions can take other functions as arguments (e.g., sqrt(round(4.45)).
  - Important to note that the parentheses are balanced.
- **Navigation tips**:
  - *Tab autocomplete*: for example, if the user types 'ro' and then hits tab, a window will be brought up showing possible commands the user might want to use (such as 'round').
  - *Help function*: if the user wants to know more about a function, they can use this function as help().

Variables

- **Variables**:
  - Likened to a box as it could store things.
  - Stores values (e.g., variable <- 'word').
  - Note that variable names are not in quotes.

Variable type	Stores	Example
Numeric	Numbers	NumericVar <- 4.78
Character	Text (via speech marks)	CharacterVar <- 'hi'
Logical	True/false values	LogicalVar <- TRUE

- Creating and using **variables**:
  - Used to store and label information.
  - Refer to the contents of a block of computer memory.
  - Use the 'assignment operator' (<-) to create one.
  - Variables in R behave the same way as their values do.
  - By assigning a new value, the old one will disappear since variables only contain one thing.