RESEARCH ESSENTIALS – Analysing your Data

Zoom

25th August 2021

Presented by

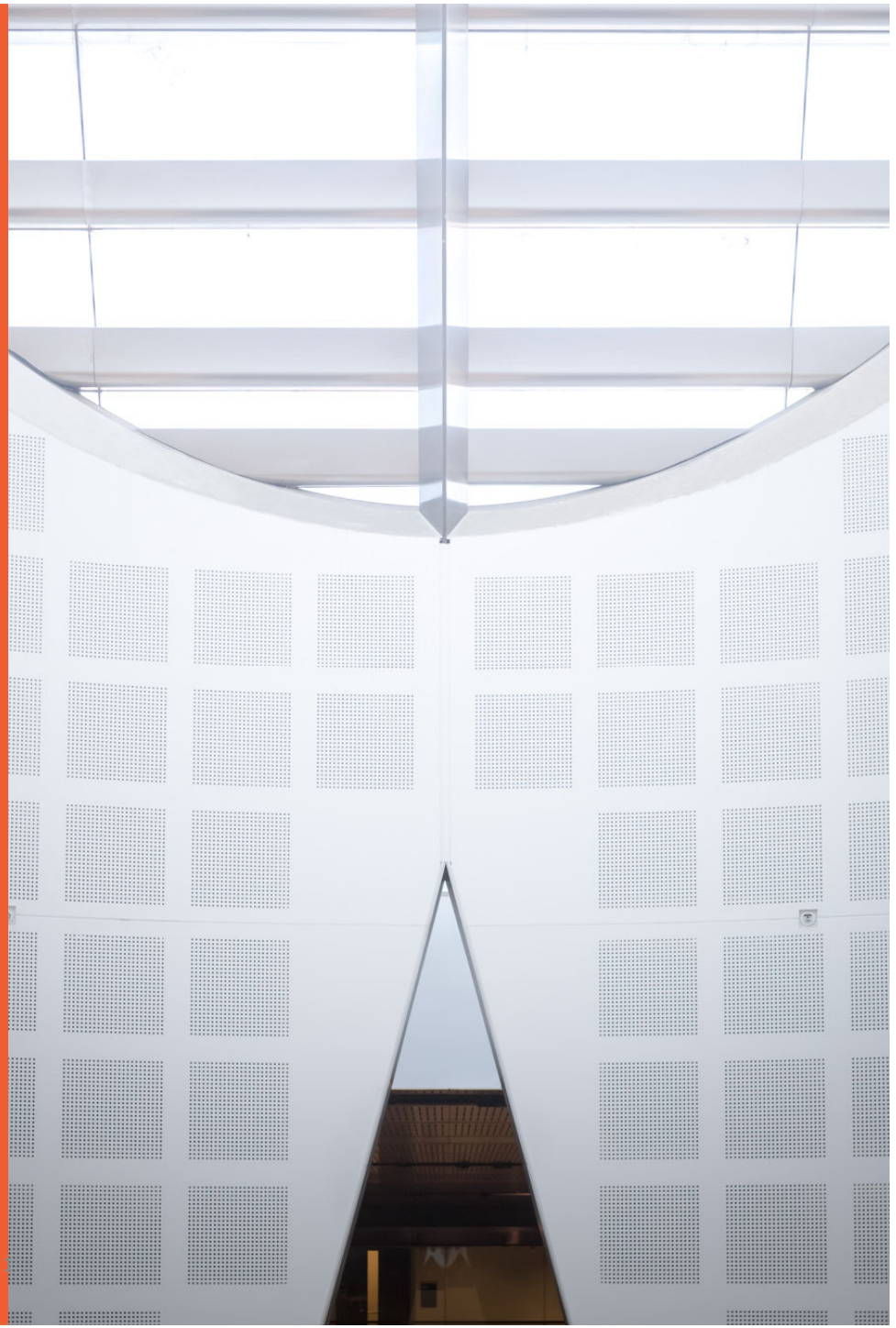
Dr Kathrin Schemann

Sydney Informatics Hub

Core Research Facilities

The University of Sydney

THE UNIVERSITY OF SYDNEY



Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge.** The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

"The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

For further information about acknowledging the Sydney Informatics Hub, please contact us at **sih.info@sydney.edu.au**.

If you want to learn more about The University of Sydney Core Research Facilities:
https://www.youtube.com/watch?v=XWYVXZ2nZIs&feature=youtu.be

How to use this workshop

- These slides have a dual purpose:
  - To guide our interactive workshops
  - As self-contained reference material to be read after the workshop

- Some slides are for your reference, and not all of the material will be discussed in the workshop. Such slides are marked with this blackboard icon

- Ask short questions or clarifications during the workshop. There will be breaks during the workshop for longer questions. You can email us about the material in these workshops at any time, or request a consultation for more in-depth discussion of the material as it relates to your specific project

Research Essentials Workshop overview

I. 8-step research workflow and other resources

Where does this Workshop fit into the research process ?

Where does it fit in with other SIH training and support on offer?

II. Setting up your data for most analyses:

Workflow Step 3: Collect and store data

Workflow Step 4: Cleaning data

III. Workflow examples for common analyses – brief introduction to:

Step 5: Exploratory data analysis

Step 6: Inferential analysis

Workshop Part I:
8-step research workflow and other resources

THE UNIVERSITY OF SYDNEY

Acknowledging SIH

All University of Sydney resources are available to Sydney researchers free of charge. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

“The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.”

For further information about acknowledging the Sydney Informatics Hub, please contact us at sih.info@sydney.edu.au.

How to use this workshop

Workshops developed by the Statistical Consulting Team within the Sydney Informatics Hub form an integrated modular framework. Researchers are encouraged to choose modules to *create custom programmes tailored to their specific needs.* This is achieved through:

- Short 90 minute workshops, acknowledging researchers rarely have time for long multi day workshops.
- Providing statistical workflows appliable in any software, that give practical step by step instructions which researchers return to when analysing and interpreting their data or designing their study e.g. workflows for designing studies for strong causal inference, model diagnostics, interpretation and presentation of results.
- Each one focusing on a specific statistical method while also integrating and referencing the others to give a holistic understanding of how data can be transformed into knowledge from a statistical perspective from hypothesis generation to publication.

For other workshops that fit into this integrated framework refer to our training link page under statistics https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats

Research Workflows

- Why do we need a research workflow?
  - As researchers we are motivated to find answers *quickly*
  - But we need to be *systematic* in order to
    - Find the right method
    - Use it correctly
    - Interpret and report our results accurately
  - The payoff is huge, we can avoid mistakes that would affect the quality of our work *and* get to the answers sooner

- So… what is a workflow?
  - The process of doing a statistical analysis follows the same general "shape".
  - We provide a general research workflow, and a specific workflow for each major step in your research
    (currently **experimental design, power calculation, analysis using linear models/survival/multivariate/survey methods)**
  - You will need to tweak them to your needs

Using this workshop after today

These slides should be used after the workshop as reference material and include these workflows for you to follow

- Todays workshop gives you the statistical workflow, which is software agnostic in that they can be applied in any software.
- There may also be accompanying software workflows that show you how to do it. We won't be going through these in detail. But if you have problems we have a monthly hacky hour where people can help you.

1on1 assistance You can request a consultation for more in-depth discussion of the material as it relates to your specific project. Consults can be requested via our Webpage (link is at the end of this presentation)

During the workshop

Ask short questions or clarifications during the workshop. There will be breaks during the workshop for longer questions.

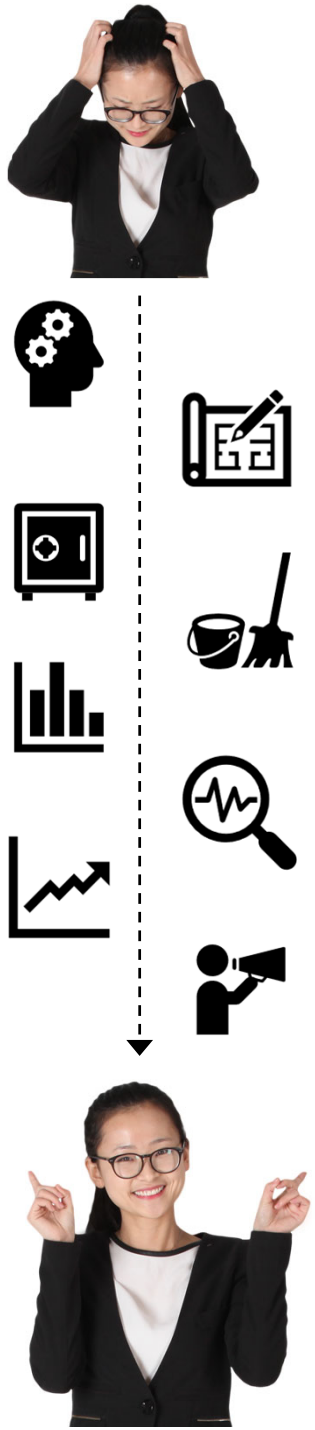
Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

Challenge Question

- A wild boar is coming towards you at 200mph. Do you:?
  - A. Ask it directions
  - B. Wave a red flag
  - C. Wave a white flag
  - D. Begin preparing a trap

General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



6. Statistical Inferential analysis – from sample to population

"Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates."

Upton, G., Cook, I. (2008) *Oxford Dictionary of Statistics*

In English:

Statistical inference:

"The theory, methods, and practice of forming judgments about the parameters of a population, usually on the basis of random sampling."

Collins English Dictionary

➔ *Think p values and confidence intervals to generalise your results from a sample to a population*

7. Predictive modeling: Inferential statistics versus machine learning

Inferential statistics:

- Interested in knowledge about the data, e.g. understand risk factors for disease or demographic factors associated with purchasing decisions.
- Generalises from the sample to the population (makes inference)

Machine learning/predictive analytics:

- Interested in prediction, use algorithm to figure out the pattern on its own directly from the data; workable and reproducible prediction model.
- Implemented based on statistical analysis but can throw off assumptions attached to the statistical methodology.

Other statistical SIH training*:

Workflow Step	Other training
1. Hypothesis generation	
2. Study design	Experimental Design Power and sample size Survey design and analysis 1 + 2 Model Building
3. Collect/store data	**Research Essentials**
4. Data cleaning	**Research Essentials**
5. Exploratory data analysis	Linear models 1-3 + Model Building Survival analysis Meta-analysis Survey design and analysis 1 + 2 Multivariate Analysis
6. Inferential analysis	
7. Predictive modelling	<Currently no WS for inferential statistics> Introduction to machine learning in R/Python – non-inferential - Data Science

* See SIH website for more information on upcoming and new training, to view the training calendar and sign up for the training mailing list: : https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html

Other SIH, Intersect and library trainings:

Workflow Step	SIH training and other support
1. Hypothesis generation	Library research support: Literature and systematic review
2. Study design	
3. Collect/store data	RedCap –various trainings for survey data, from introduction to advanced Collecting web data Database and sql Research data management modules Research data management techniques
4. Data cleaning	
5. Exploratory analysis	
6. Inferential analysis	
7. Predictive modelling	
8. Publication	Library research support: Data publishing, preservation and archiving

SIH's Research Data Management (RDM) support

eNotebooks

DashR & Research Data Management Plans

Research Data Store

REDCap

Dropbox

Office 365

GitHub

cloudstor

- Check out https://informatics.sydney.edu.au/rdm/ if your needs are extensive or complex
- Check out the **Digital Research Café** series for handy tips (short videos): https://web.microsoftstream.com/channel/9278c52a-164d-453f-8c3c-9f933f6d8b2f
- Look for SIH training, RDM drop-in session or contact SIH's Research Data Consulting for further support: digital.research@sydney.edu.au

Software training (Intersect)

Workflow Step	SIH software courses offered
1. H_0 generation	
2. Study design	
3. Collect/store data	
4. Data cleaning	Statistical analysis with SPSS Programming with R Data manipulation and visualisation in R Excel for researchers Data carpentry- Geospatial analysis in R
5. Exploratory analysis	
6. Inferential analysis	Exploring Chi-square + correlations SPSS Statistical comparisons using R Exploring and predicting linear regression
7. Predictive modelling	
8. Publication	

SIH also offers training in Python, Julia and Matlab and high performance research computing/ bioinformatics (Artemis; Galaxy and parallel computing)

The big question: which car will you take?

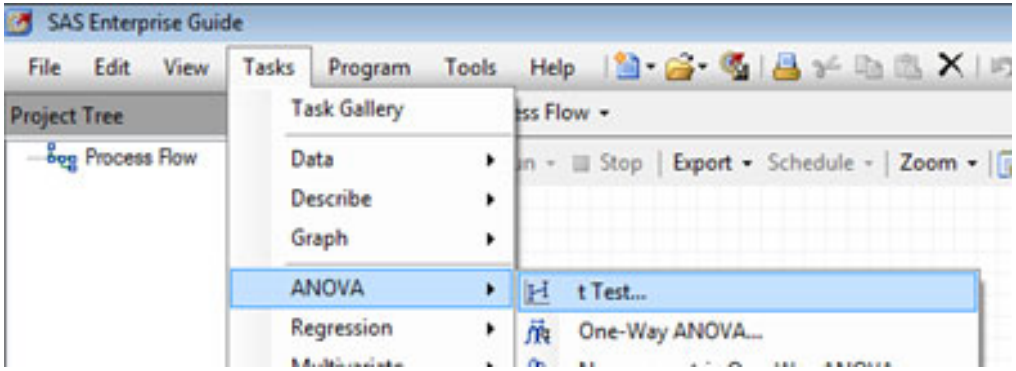
CLI

Software choice: programming versus menu-driven

- How do you record your analysis for reproducible research?

R code versus **Interactive SAS Enterprise Guide Menu**

```
> t.test(x,y)
```



- If using interactive processing, you should keep a track of the commands you ran

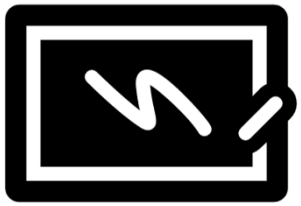
- By documenting you should always be able to rerun your analysis from start to finish (and get the same result!)

Other non-training, face-to-face support*:

Workflow Step	Other training
1. Hypothesis generation	Library research support consultancy
2. Study design	Hacky hour; One-on-one statistical consultancy
3. Collect/store data	Hacky hour; RDM drop in session or one-on-one RDM consultancy
4. Data cleaning	Hacky hour; Drop in or one-on-one statistical consultancy
5. Exploratory data analysis	Hacky hour; Drop in or one-on-one statistical consultancy
6. Inferential analysis	Hacky hour; Drop in or one-on-one statistical consultancy
7. Predictive modelling	Hacky hour; Drop in or one-on-one statistical consultancy/Data science consultancy
8. Publication	Library research support consultancy

* See SIH website for hacky hour/drop in session times or to request assistance: https://www.sydney.edu.au/research/facilities/sydney-informatics-hub.html

Questions so far?



Research Workflow

1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. **Collect/Store Data**
4. Data cleaning
5. Exploratory Data Analysis (EDA)
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication

3. Collect/store your data

a. Research data management
b. Organise your data for input into statistical software

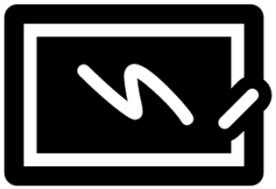
THE UNIVERSITY OF SYDNEY

a) Research data management

- Data storage
  - Back up EVERYTHING including original data collection forms or raw data (images, electrical signals, DNA sequences, whatever)

- Data entry - will you be using manual data entry?
  - Ideally double-data entry followed by comparison
  - Be wary of spreadsheets – especially entering and editing in the same sheet
  - Statistical software preferred

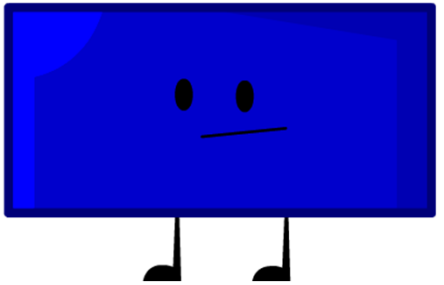
a) Research data management



- Have you got a Research data management plan according to University policy?
  - Research Data Management Guide?
  - What are the university supported tools for data collection and storage?
    - What is an eNotebook?
    - Where can I store my data?

- Consider appropriate folder/directory structure, file naming and version control for your project, *or at least your part of it*
  - "Good enough practices for scientific computing"

b) Organising a dataset for analysis



- Consider which variables are required for analysis

- Most programs read in data in a rectangular format:
  - A header including column names in the first row
  - Each row thereafter being the data itself (often corresponding to a single unit of interest – e.g. person, udder quarter, animal, plant, plot, farm, etc)
  - Each column represents one variable
    - ID variable – identifies the subject
    - Demographic variable – characteristics of the subject including their treatment
    - Measurement variable – some observation on the subject
  - A delimiter between each column (comma .csv and tab .tsv/.tab)

- Check your data once it is imported into the statistical software

Pitfalls when coming from Excel:

- Blank rows
- Watch out for:
  - Merged cells
  - Cell comments
  - Colour coding
- Data in multiple sheets
- Coding of missing data/blanks/non-applicable
- Deal with the above in Excel before exporting to text. If you need to restructure your data, usually easier to do this in stats software
- A good summary of these pitfalls is provided in [this paper]

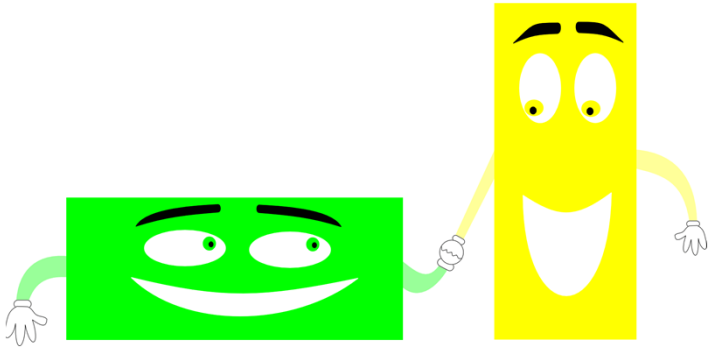
b.) Data formats - transformations

Animal ID	Time 1	Time 2	Time 3
1	50	55	60
2	47	49	50
…	…	…	…

Wide/unstacked format

Animal ID	Time	Body weight
1	1	50
1	2	55
1	3	60
2	1	47
2	2	49
2	3	50
…	…	…

Long/stacked format



Data transformation

- Required format depends on type of analysis

	A	B
1	Patient	Cholesterol
2	1	181
3	2	182
4	3	193
5	4	193
6	5	203
7	6	190
8	7	201
9	8	191
10	9	208
11	10	203

Sheet 1: Treatment

	A	B
1	Patient	Cholesterol
2	1	209
3	2	192
4	3	203
5	4	217
6	5	204
7	6	209
8	7	220
9	8	214
10	9	212
11	10	202

Sheet 2: Control

Unstacked/wide format in multiple sheets

	A	B	C
1	Patient	Treatment	Cholesterol
2	1	treatment	181
3	2	treatment	182
4	3	treatment	193
5	4	treatment	193
6	5	treatment	203
7	6	treatment	190
8	7	treatment	201
9	8	treatment	191
10	9	treatment	208
11	10	treatment	203
12	1	control	209
13	2	control	192
14	3	control	203
15	4	control	217
16	5	control	204
17	6	control	209
18	7	control	220
19	8	control	214
20	9	control	212
21	10	control	202

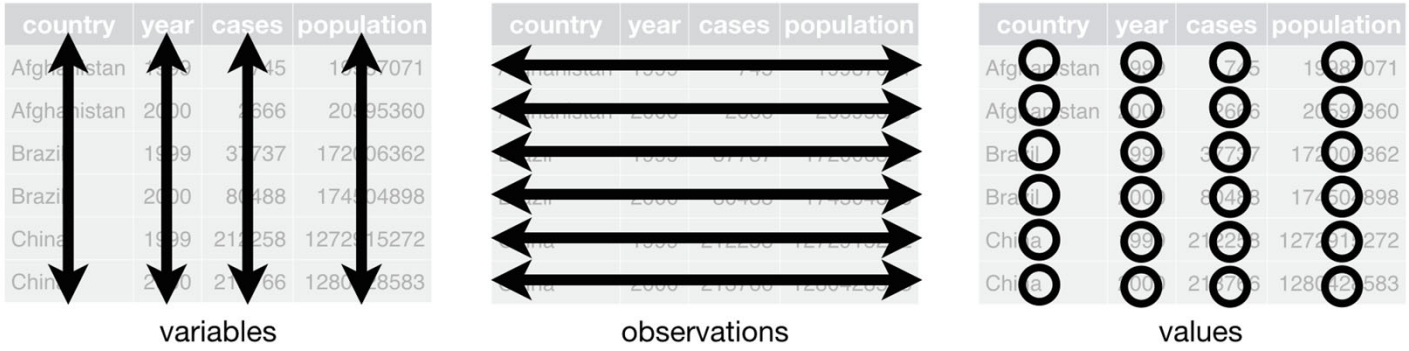
Create Sheet 3: Stacked/long format In one file

Data formats – tidy data

- Depending on the design of your experiment/survey you may have a mix of demographic data on each individual, and measurements
  - You may need multiple tables and a unique ID for each individual to link them

- Wide and long can become relative terms if you have multiple measurement and ID variables in one table

- Tidy data is an absolute term, which describes data transformed to:



https://r4ds.had.co.nz/tidy-data.html

b.) Organising a dataset for statistical analysis: Data coding

- Specify type of variable: ensure your analysis software knows whether a variable is continuous (numeric), nominal or ordinal (text)

- Label variables (e.g. Age = Age at interview in years) and values/'levels' within categorical variables, e.g. 1 = "Male", 2="Female", 3="Non-binary"

- Correctly code missing values according to software program: ensure your analysis software knows that the data is missing and not '0' or some other value

Version control - keeping track of files

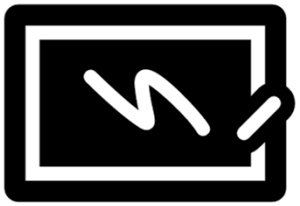
- Use a separate directory for each discrete analysis
- When processing data and intermediate files save with a new name
- Create a log file in the same directory and use version control (e.g. name sequentially, date/time stamp, for example:
  - *"20200401_stats101_workshop.ppt"*
  - *"2020_stats101_workshop_v2.0.ppt"*

Example of a version log file:

File name	Date created	Description	# Obs	#Vars
Mydata01.csv	30/3/2020	Original data entry by KS, 1 record per person	250	34
Mydata02.csv	1/4/2020	Eligible records only based on study inclusion criteria with new variables created for analysis	204	37

Further information: https://library.sydney.edu.au/research/manage-data.html

Research Workflow



1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. **Data cleaning (Descriptive data analysis 1)**
5. Exploratory Data Analysis (EDA; Descriptive analysis 2)
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication

4. Data Cleaning

A. Descriptive analysis

i. Identify variable types

ii. Describe the distribution of individual variables

THE UNIVERSITY OF SYDNEY

Why worry about variable types?

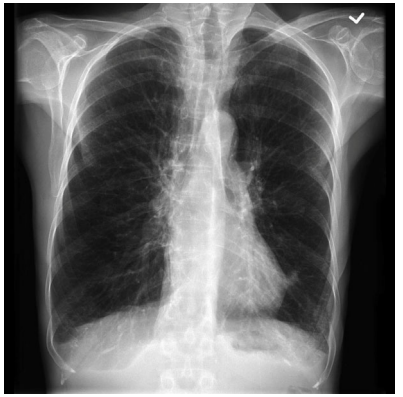
- **Types of variables determine the appropriate statistical methods for data analysis**

- You need to know what data type your variable is AND how it is recorded in your data

- You may need to convert a continuous variable to a categorical variable depending on its distribution

Data type versus functional classification

- Functional classification:



Smoking

Predictor
Explanatory variable
~~Independent variable~~

Lung disease

Response
Outcome
Dependent variable

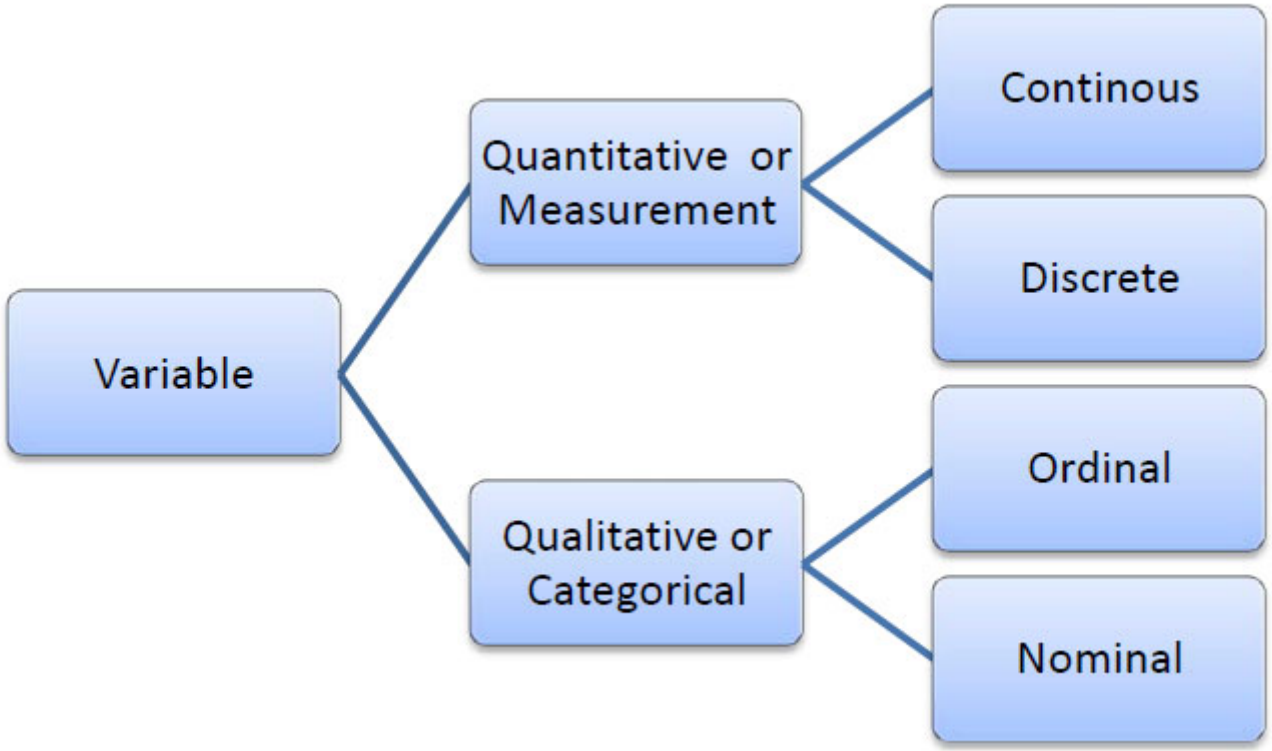
Other functional classifications for variable types

- Covariate/exposure variable: a variable measured on the sampling units of which we have no control over

- Experimental design variables:
  - Design variables: Based on the physical design of the experiment. They are often included in the analysis even if not 'significant' in order to correctly partition the variance e.g. Block, id, etc.
  - Treatment: Variables of interest, e.g. diet, drug treatment, intervention etc.

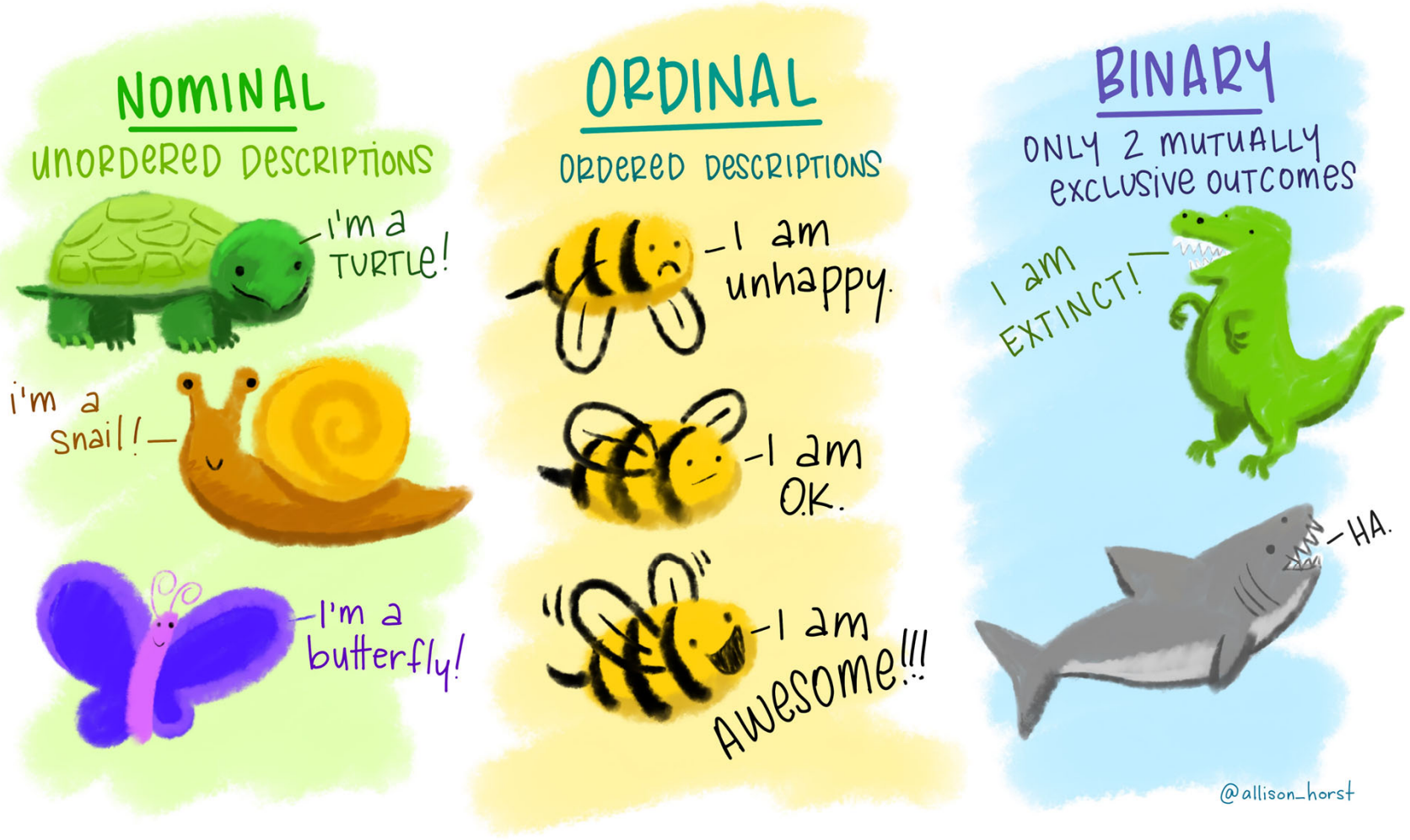
➔ Consider our SIH "Experimental design" Workshop!

i.) Identify variable types:



Variable types

Variable types



ii.) Describe individual variables
Data processing: the outcome variable(s)

- Review study aim and objectives
  - E.g. vaccine RCT - daily mortality outcome data could be analysed as:
    - mean daily rate
    - cumulative mortality
    - peak mortality
    - outbreak presence/absence
    - time to infection/disease outbreak

ii.) Describe individual variables
Data processing:

- Assess all variables for missing observations – if many missing consider analysing with and without that predictor

- Check the distribution of all variables individually
  - Continuous predictors: handle as continuous or categorical?
  - Categorical: may have to combine categories if there are low frequency counts (if it makes sense to do so)

- Multi-level (clustered) data
  - Each observation/row uniquely identified? E.g. herd, animal, ID
  - Evaluate hierarchical structure of your data: Average/range of observations at one level in each higher level?
    - E.g. mean, min, max of students/class; mean, min, max of classes/school

Data wrangling and data dictionary - example

- Use short but informative variable names
- Names should keep track of transformations/recoding, e.g.
  - age = original data in years
  - age_ct = age after centering by subtracting the mean age
  - age_ctsq = quadratic term (age_ct squared)
  - Age_c2 = age categorised into two categories (young vs old)
  - Use a single letter prefix to help keep groups of variables together, e.g. b_ecoli, b_staphau, etc.

	A	C	D
1			
2	Questions	Categories	Code used
3	Q1_Age(years)	20-30	1
4		31-40	2
5		41-50	3
6		51-60	4
7		>60	5
8	Q2_Gender	male	1
9		female	2

Keep track of analyses

- Remember you should be able to repeat analysis from the start, to demonstrate/enable reproducibility

- For statistical programming languages:

  - Name the program file logically

  - Use structure, work in blocks or ‘chunks’ of code for different sections, e.g. ‘descriptive analyses’ – do it for all predictors in one go

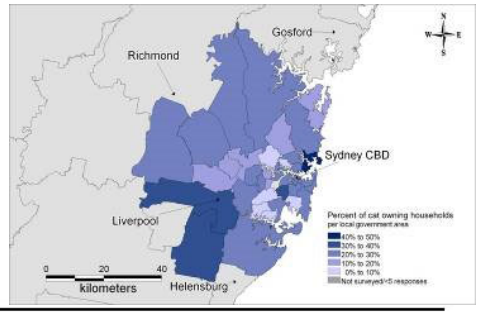
  - Log file – same name as program file, different extension – VERY important as record for interactive mode!

  - Use functions to avoid repetition.

  - Use appropriate level of comments, e.g. key steps and results

  - Consider using Rmarkdown notebook if using R

Variable types – example Sydney cat study



Journal of Feline Medicine and Surgery (2009) **11**, 449–461
doi:10.1016/j.jfms.2008.06.010

ELSEVIER

jfms

Demographics and husbandry of pet cats living in Sydney, Australia: results of cross-sectional survey of pet ownership

Jenny-Ann LM Toribio BVSc, PhD[1,a], **Jacqueline M Norris** BVSc, MVS, PhD, MASM, GradCertHigherEd[1,a], **Joanna D White** BVSc, MACVSc[1], **Nanveet K Dhand** BVSc&AH, MVSc, PhD, MACVSc[1], **Samuel A Hamilton** BSc(Vet), BVSc, MACVSc[1], **Richard Malik** DVSc, DipVetAn, MVetClinStud, PhD, FACVSc, FASM[1,2*,a]

Sydney cat study data

Cat ID	Age (yrs)	Breed	Sex	Vaccinated?	Years since last vet visit	Never gone to vet
1	5	DSH	M	1	0	FALSE
2	8	Russian Blue	F	0		TRUE
4	14	DSH	M	1	3	FALSE
5	6	Barman	F	1	1	FALSE
6	6	DSH	F	1	0	FALSE
7	2	DSH	M	1	0	FALSE
8	3	Persian/Ragdoll	F	1	0	FALSE
9	12	DLH	F	1	0	FALSE
10	10	DSH	F	1	1	FALSE
11	9	DSH	M	1		FALSE

Step 2.1: Descriptive analysis for individual variables

Outline:

- Categorical variables
  - Frequency tables
  - Bar charts

- Continuous variables
  - Graphical summaries
    - Histogram
    - Box-and-whisker plot
  - Numerical summaries
    - Mean
    - Median
    - Mode
    - Quartiles
    - Percentiles

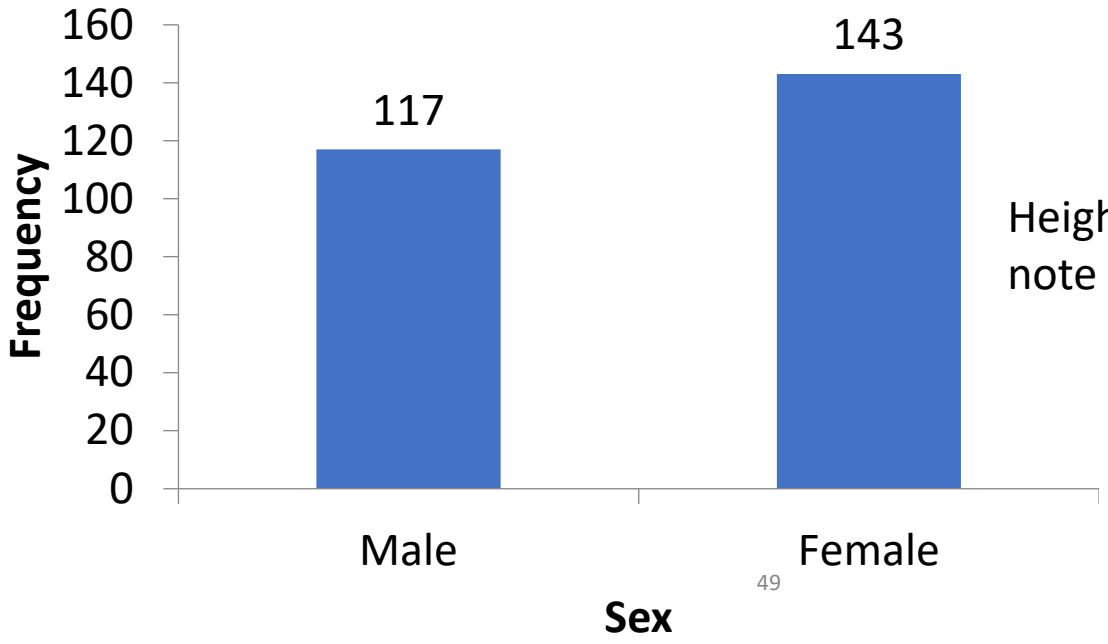
How to summarise categorical variables?

Cat ID	Age (yrs)	Breed	Sex	Vaccinated?	Years since last vet visit	Never gone to vet
1	5	DSH	M	1	0	FALSE
2	8	Russian Blue	F	0		TRUE
4	14	DSH	M	1	3	FALSE
5	6	Barman	F	1	1	FALSE
6	6	DSH	F	1	0	FALSE
7	2	DSH	M	1	0	FALSE
8	3	Persian/Ragdoll	F	1	0	FALSE
9	12	DLH	F	1	0	FALSE
10	10	DSH	F	1	1	FALSE
11	9	DSH	M	1		FALSE

Frequency - count the number of Male and Female cats

Summarising sex in the Sydney cat study

Sex	Frequency /count	Relative Frequency (%)
Female	143	55
Male	117	45
Total	**260**	



Height proportional to frequency; note space between bars

Step 1: Descriptive analysis for individual variables

Outline:

- Categorical variables
  - Frequency tables
  - Bar charts
- Continuous variables
  - Graphical summaries
    - Histogram
    - Box-and-whisker plot
  - Numerical summaries
    - Mean
    - Median
    - Mode
    - Quartiles
    - Percentiles

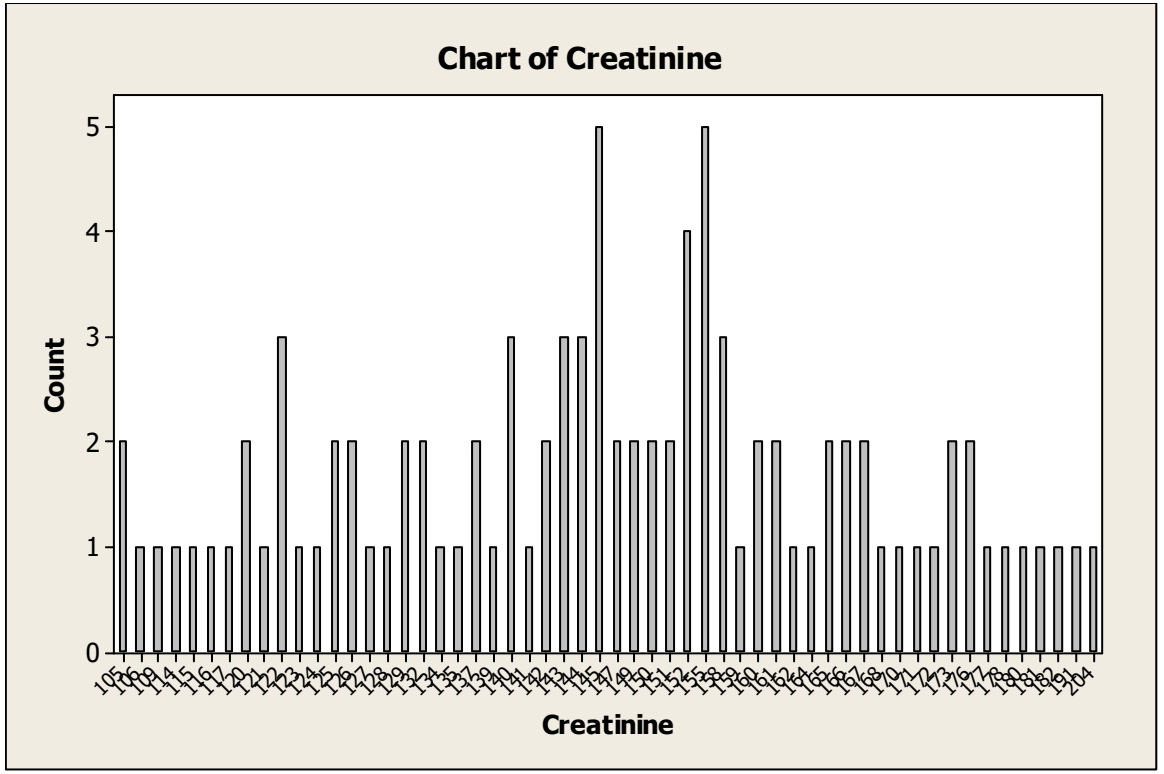
How to summarise the variable: Creatinine

- Creatinine levels (µmol/L) of 96 cats

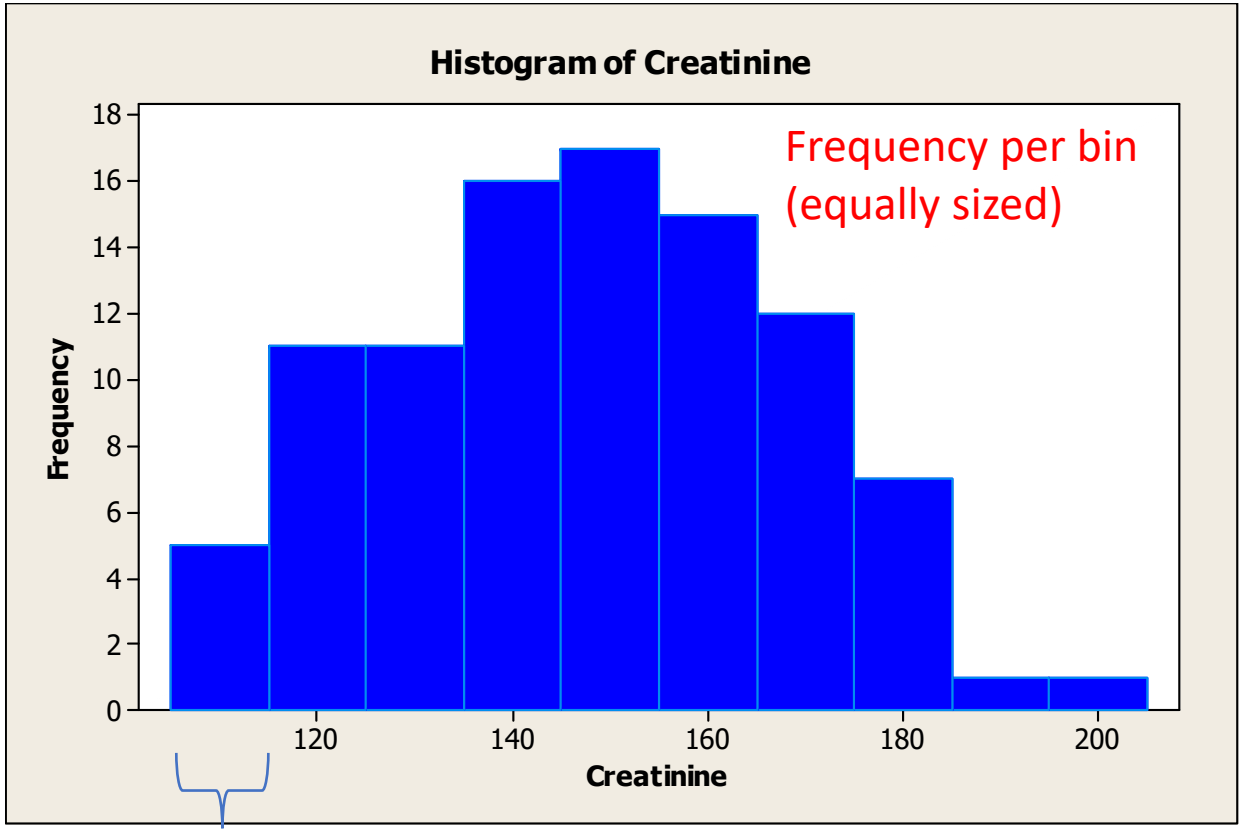
170	164	173	106	160	139	105	178	140	140	172	155	122
152	125	114	144	155	180	137	150	105	132	120	145	162
166	176	137	152	155	122	145	123	165	145	161	124	128
182	171	155	149	158	161	177	158	151	147	142	143	126
144	159	166	117	167	127	142	149	120	151	125	121	155
181	191	134	158	143	147	109	167	141	152	122	144	145
116	160	173	145	204	135	143	129	150	152	129	126	132
176	115	168	165	140								

Frequency table would be long and messy! Not a great summary.

Bar chart of creatinine ☹



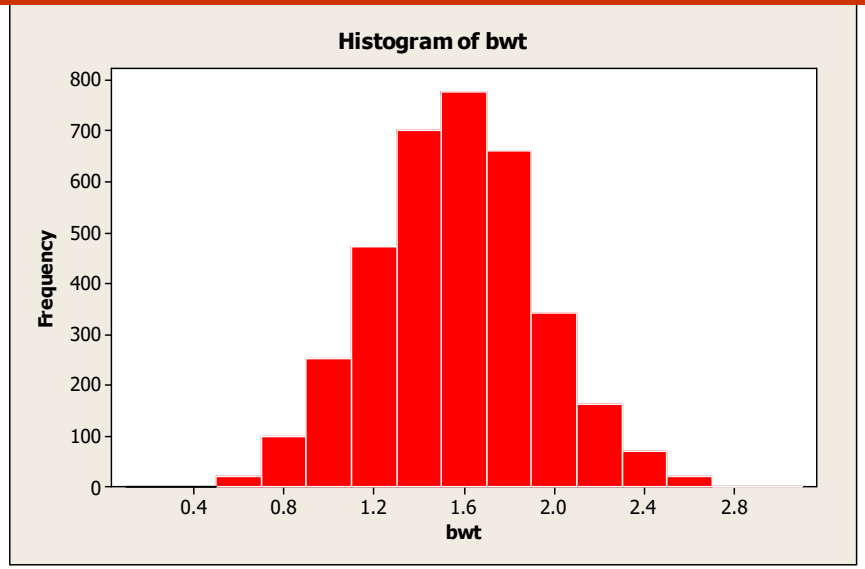
Histogram of creatinine



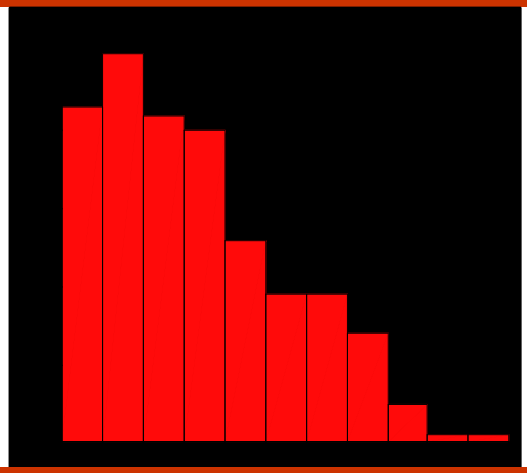
Class/bin: 104-110

Shapes of the distribution

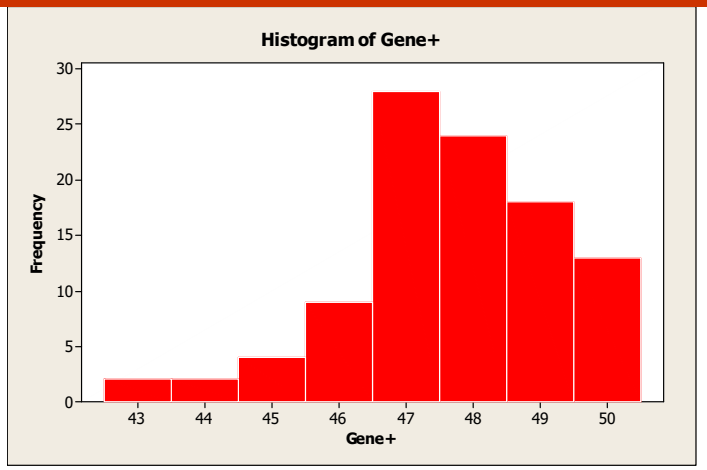
Symmetric Distribution



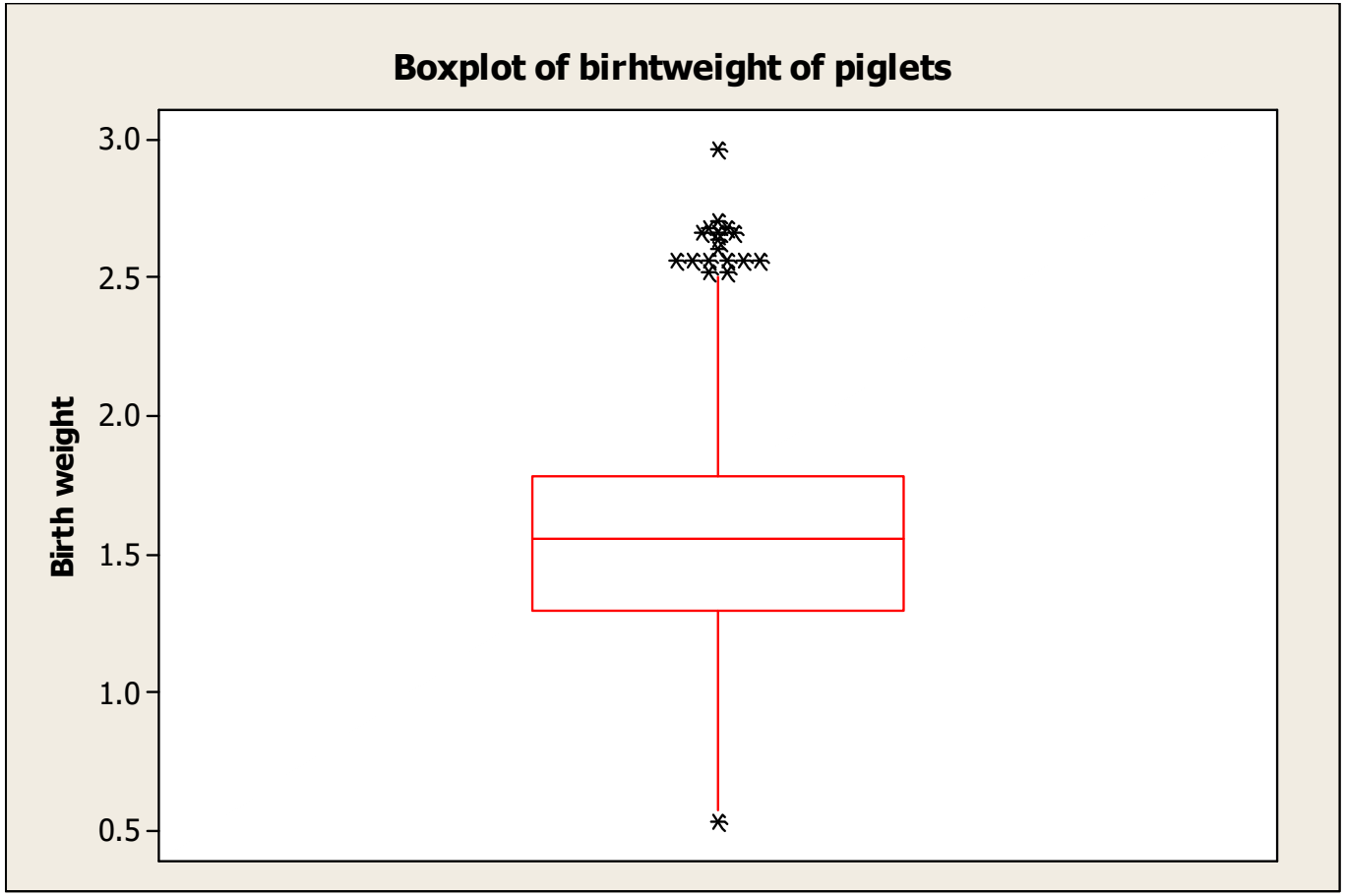
Asymmetric Distributions



Asymmetric Distributions

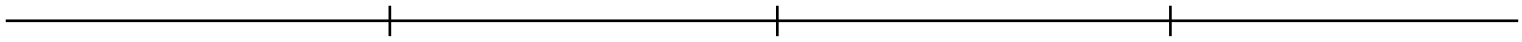


Boxplot


Boxplot of birhtweight of piglets
Birth weight
3.0
2.5
2.0
1.5
1.0
0.5

Summarising the picture

- A numerical distribution can be summarised by giving descriptions/measures of:
  - its shape
    (symmetric, right skewed, left skewed)
  - its centre
    (measures of central value or central location)
  - its spread
    (measures of spread/dispersion)



So…what is a systematic approach to conduct descriptive analyses for individual variables?

- Categorical variables
  - Frequency table
  - Bar chart

- Continuous variables
  - Histogram
  - Box-and whisker plot

- Symmetric??
  - Yes
    - Mean
    - Standard deviation
    - Min and Max
  - No
    - Median
    - Quartiles
    - Min and Max

Median and quartiles can be used for symmetric data

but it is not a good idea to use mean and standard deviation for asymmetric data

Don’t forget to check for missing data/non-applicable’s!

Data checks and manipulation in R

- Highly recommend using the tidyverse packages within R

- Learning the R Tidyverse Welcome (linkedin.com)

Questions?

Research Workflow

1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. Data cleaning
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. Predictive modelling
8. Publication

Steps for statistical analysis versus Workflow

Step: 4. Data cleaning **4. Data cleaning** **6. Inferential Analysis**

5. Exploratory data analysis

Descriptive data analysis:

A.) Description of individual variables (Workflow Step 4)

B.) Evaluation of association between two variables (Workflow Step 5)

Summary – descriptive data analysis

- A categorical variable
  - Frequency table
  - Bar chart
- A continuous variable
  - Histogram
  - Box-and-whisker plot
  - Mean ± std deviation
  - Median and quartiles
- Two categorical variables
  - Contingency table
- A categorical and a continuous variable
  - Tabulate summary statistics by groups
  - Box-and-whisker plot by groups
- Two continuous variables
  - Scatter plot

Data Analysis Workflow: 4 Examples

A – Linear Models examples:

Simple regression, ANOVA, ANCOVA, Repeated measures.

B – Extended Linear Models example:

Piecewise, linear mixed model.

C – Extended Linear Models example:

Generalised Linear Model – Poisson regression

D – Multivariate Analysis

Confirmatory Factor Analysis

Example A: Linear models examples

Scenario: We are interested in studying a continuous outcome variable, e.g. weight gain (kg) or blood cell count (cells/µL)

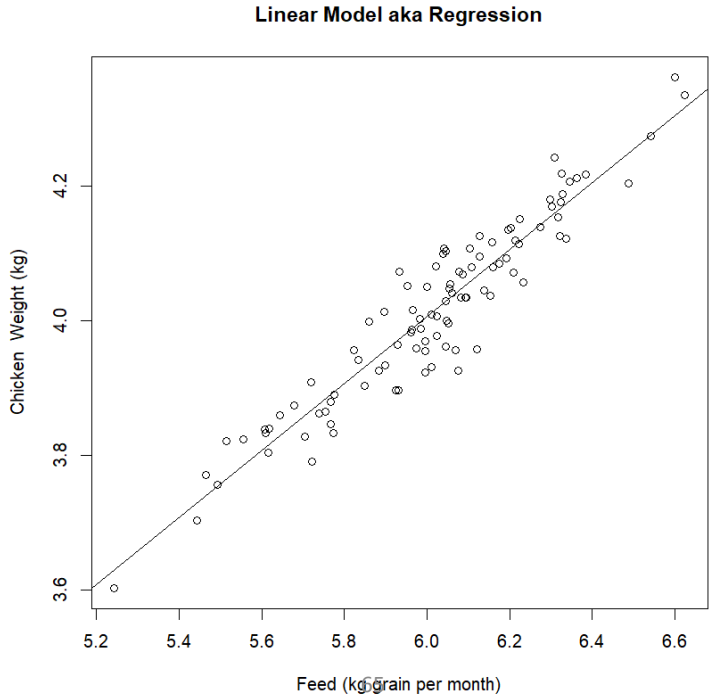
a) Simple Linear Regression – one continuous predictor variable
b) ANOVA (Control vs Treatment) – for 2 groups = 2 sample t-test = simple linear regression – one binary predictor variable
c) ANCOVA – ANOVA with a covariate
d) Repeated Measures (basic mixed model)

➔ For more detail on how to do these analyses and for R code, attend our SIH "Linear models 1" workshop!

Example A: Linear models – Simple Linear regression

Step 5: EDA – Plot the data in a scatter plot

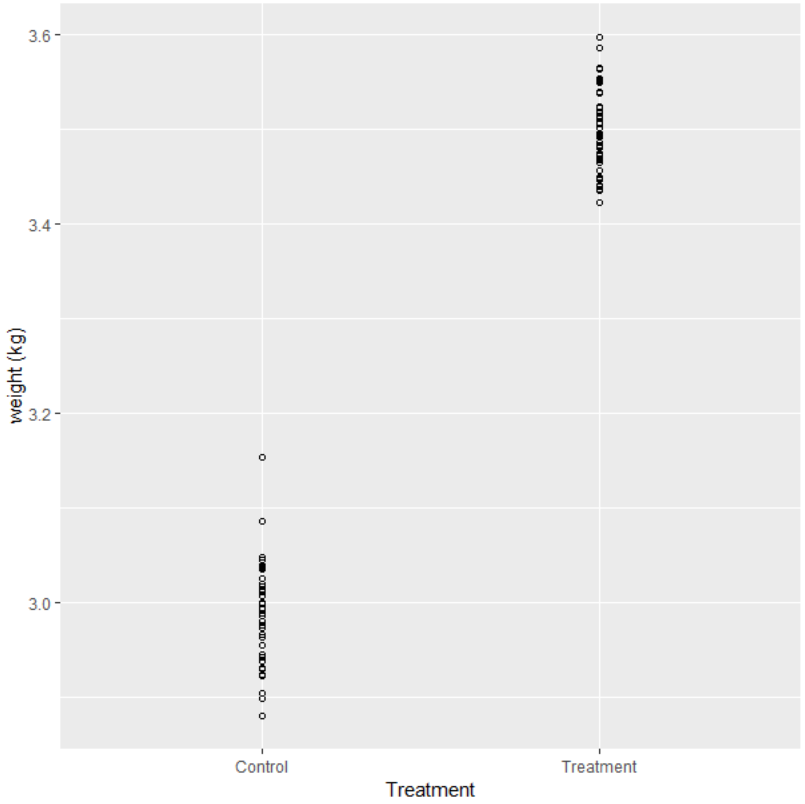
Step 6: Inferential analysis – fit a linear regression line and test if the slope is different from 0; p <0.001; report slope/regression estimate and 95% CI.

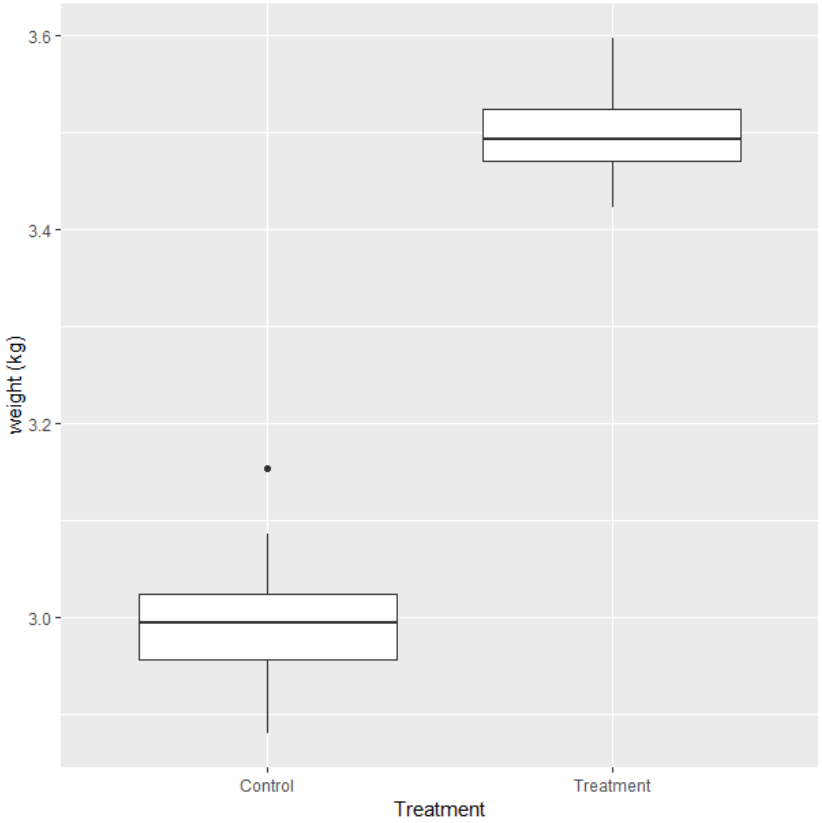


Example A: Linear models – Control versus Treatment experiment

Step 5: EDA – plot the data; side-by-side box plots

Step 6: Inferential analysis – ANOVA/ 2 sample t-test; $p<0.001$. Report predicted means and 95% CI's.

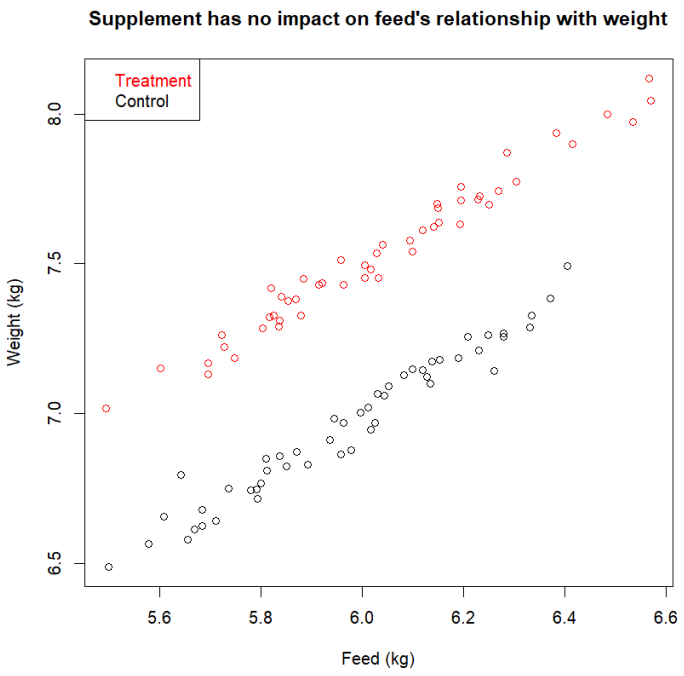




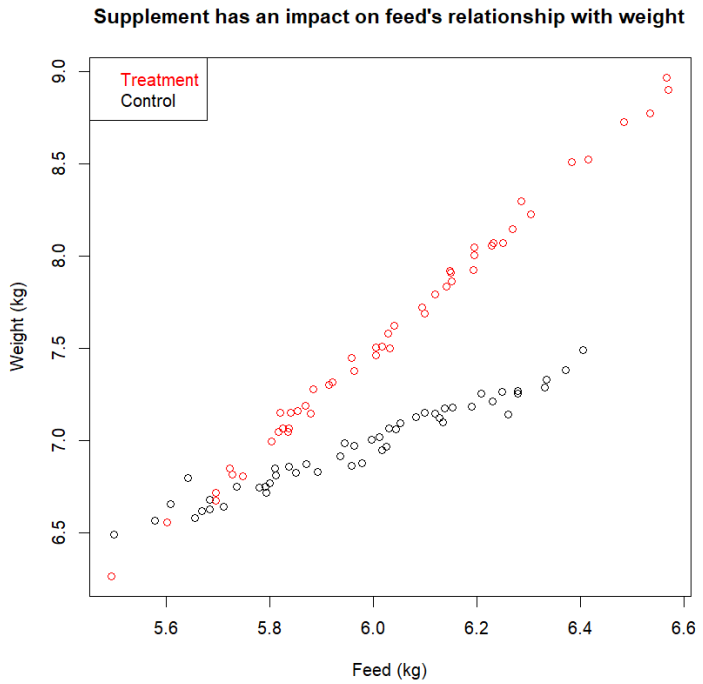
Example A: Linear models – ANCOVA - ANOVA with a continuous covariate

Step 5: EDA – plot the data; differentiate categories of the treatment variable

Step 6: Inferential analysis – ANCOVA/ multivariable regression



without interaction



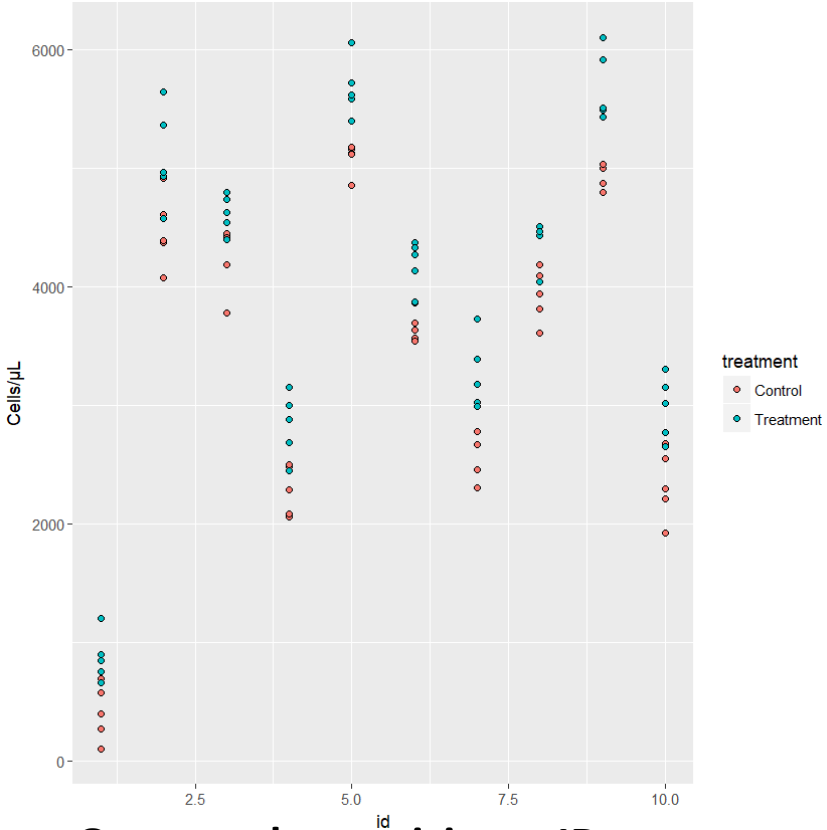
with interaction

Example A: Linear models – Repeated measures

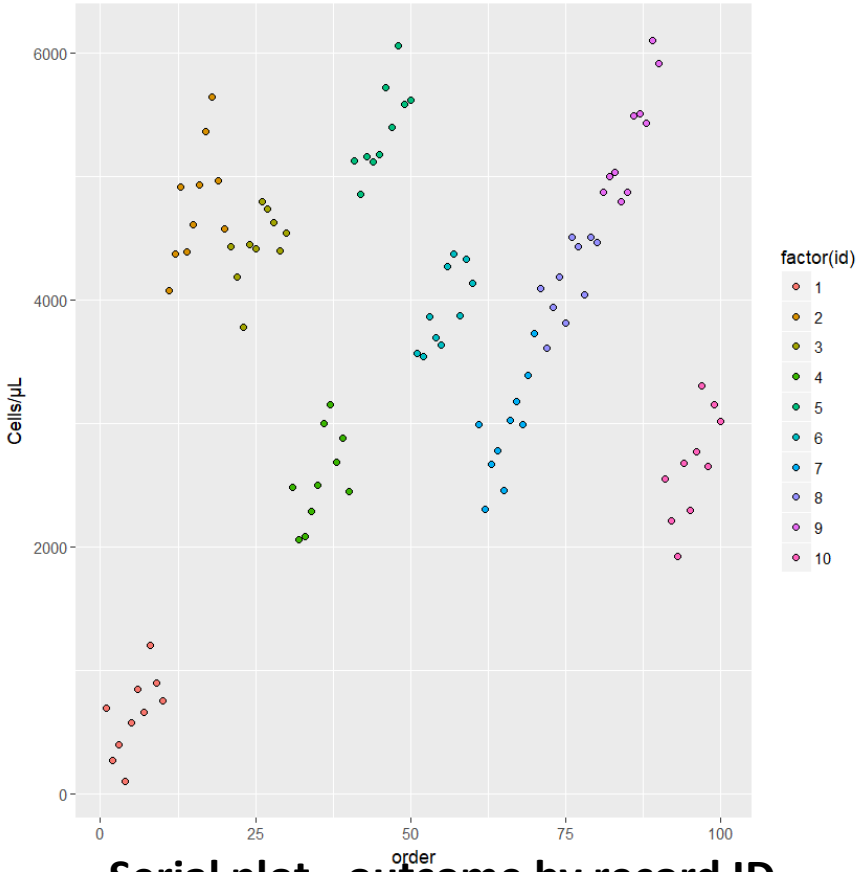
Scenario: n=10 with 5 before and 5 after treatment measurements

Step 5: EDA – plot the data by participant ID and record ID

Step 6: Inferential analysis – repeated measures ANOVA; linear mixed model



Outcome by participant ID



Serial plot - outcome by record ID

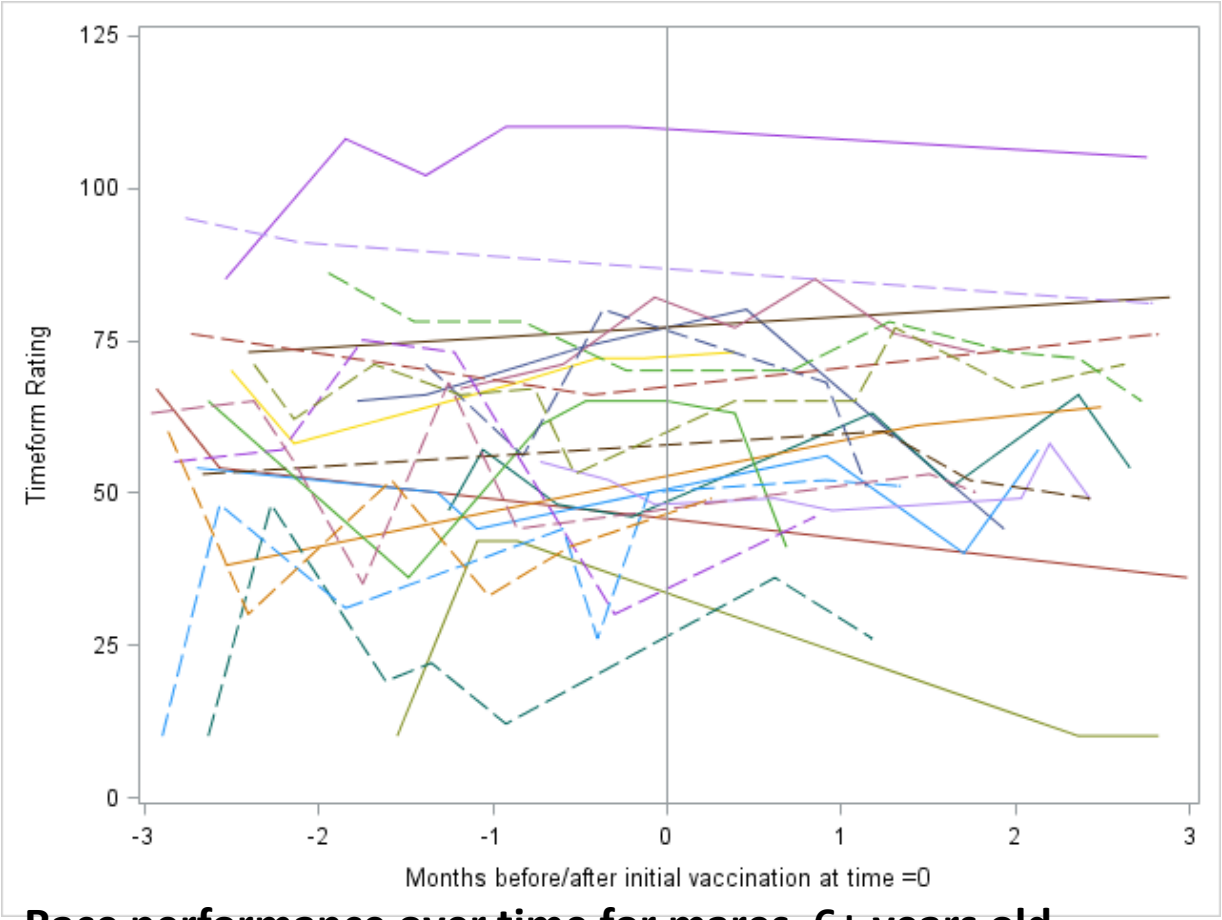
Example B: Extended Linear Model – Before-after piecewise linear mixed model

Scenario: To assess the impact of Hendra virus vaccination on racehorse performance

Reference: Schemann K, Annand E, Reid P, Lenz M, Thompson P and N Dhand (2018) Investigation of the effect of Equivac® HeV Hendra virus vaccination on thoroughbred racing performance, Australian Veterinary Journal, 96(4): 132-141.

Example B: Extended Linear Model – Before-after piecewise linear mixed model

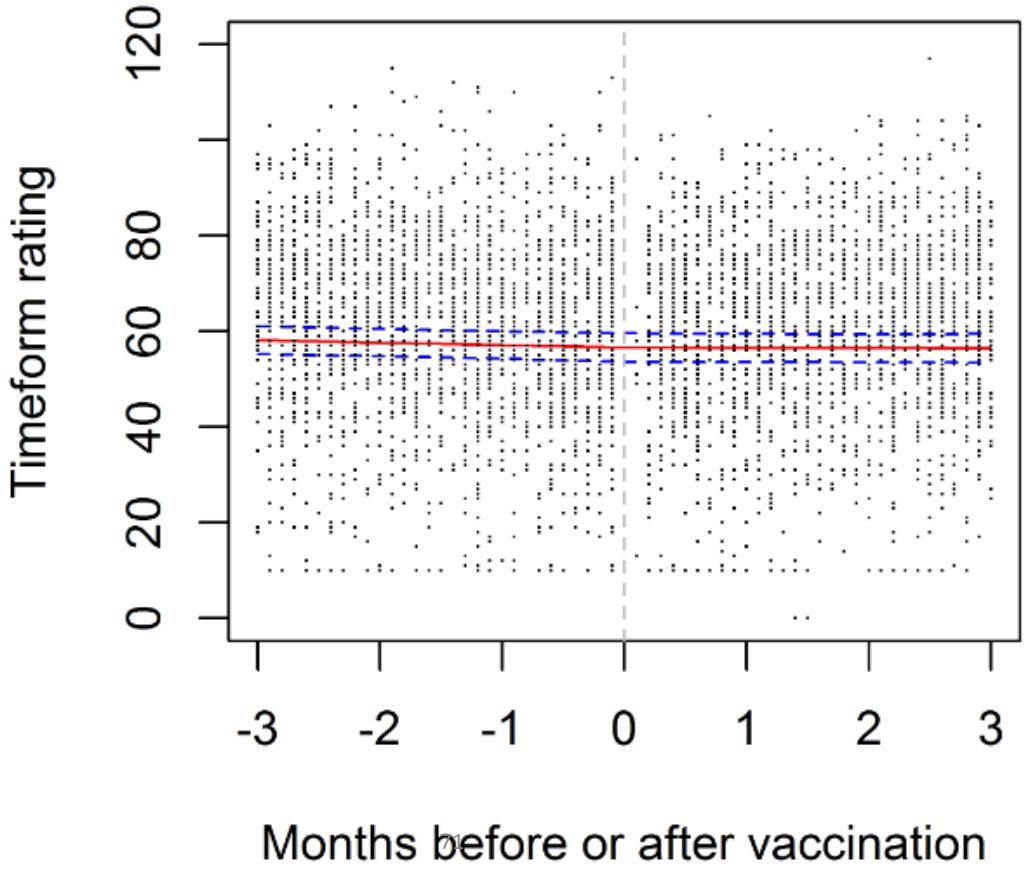
Step 5: EDA – ‘Spaghetti plot’, consider stratifying by age, sex, etc.



Race performance over time for mares, 6+ years old

Example B: Extended Linear Model – Before-after piecewise linear mixed model

Step 6: Inferential analysis – multivariable, piecewise, linear mixed model to fit and compare two separate regression lines (before versus after), adjusted for multiple covariates and for repeated measures



Example C: Extended Linear Model – Generalised Linear Model (GLM) / Poisson regression

Scenario: Quasi-before-after-control-impact experiment to assess how dingoes respond to a decline in anthropogenic foods.

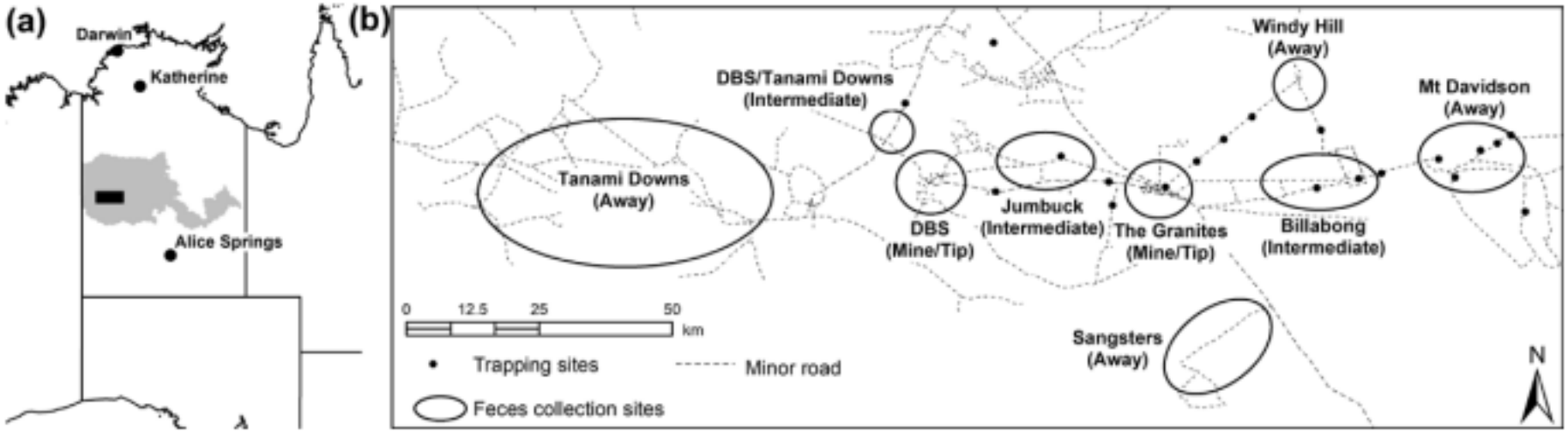
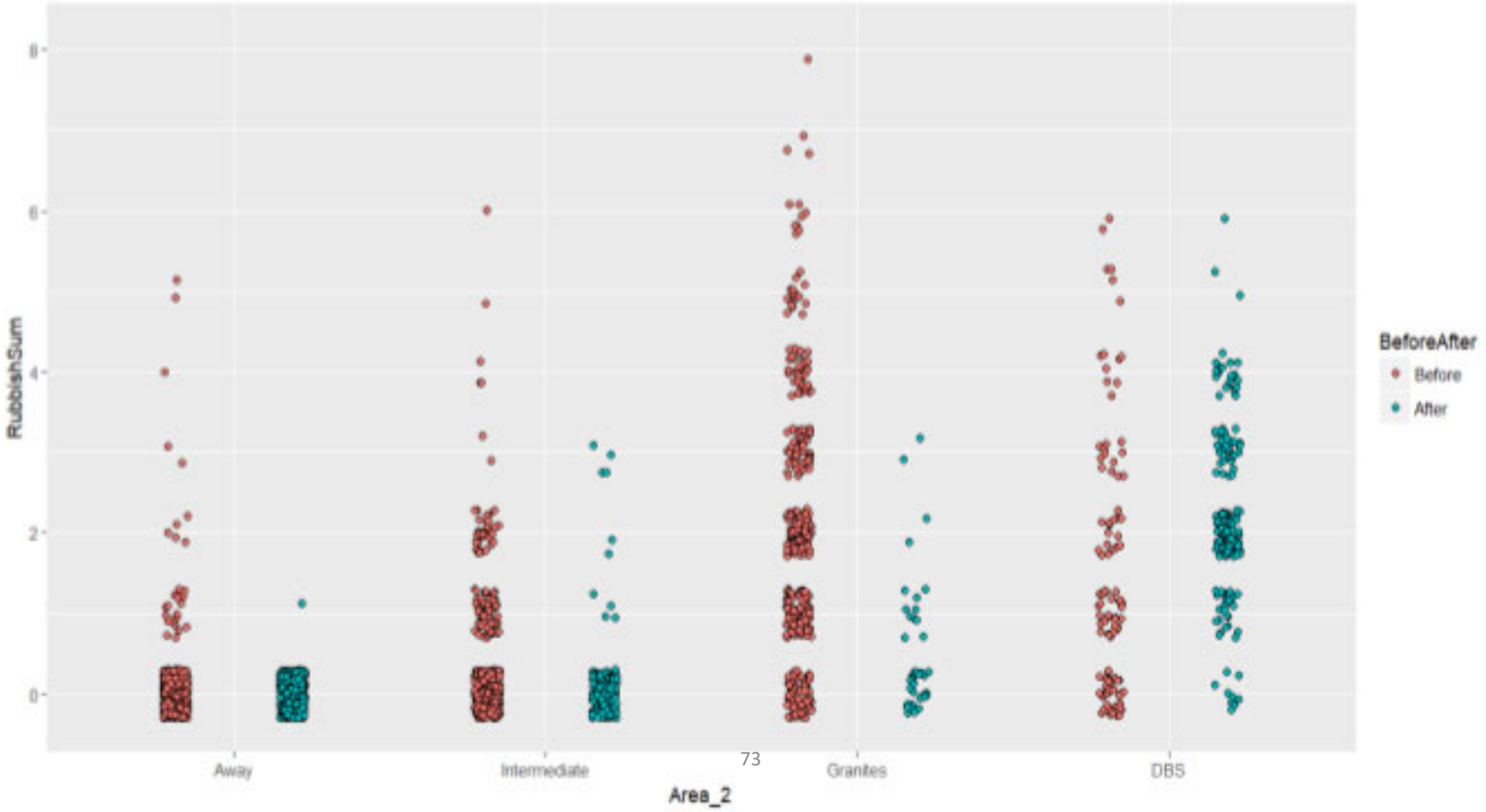


Fig. 1.—(a) Study site (box) in the Tanami Desert (gray shade), central Australia (130°18′E, 20°30′S), and (b) locations where dingo (*Canis dingo*) feces were collected for this study (2008/2009 and 2016) in relation to trapping sites that provided prey availability estimates for the region. Area categories used for the analysis are indicated in parentheses.

Reference: Newsome TM, Howden C and AJ Wirsing (2019) Restriction of anthropogenic foods alters a top predator's diet and intraspecific interactions, *Journal of Mammalogy*, 100(5), pp. 1522–1532.

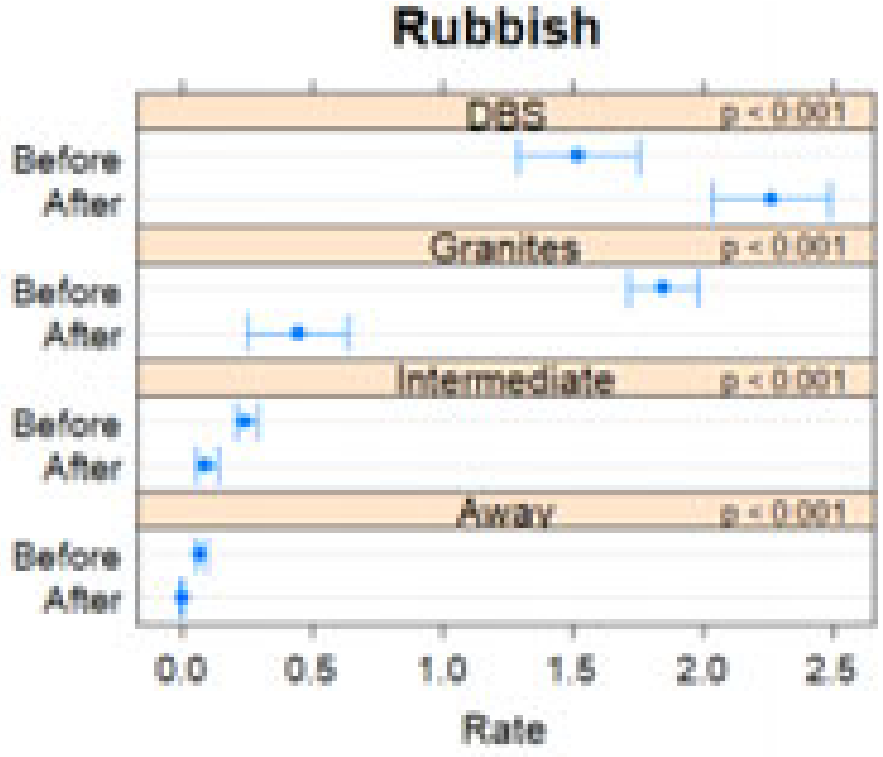
Example C: Extended Linear Model – Generalised Linear Model (GLM) / Poisson regression

Step 5: EDA – plot count data by area and treatment



Example C: Extended Linear Model – Generalised Linear Model (GLM) / Poisson regression

Step 6: Inferential analysis – Poisson regression to compare the rate of rubbish at different sites before and after the intervention



➔ To learn more about logistic and Poisson regression attend our SIH "Linear Models 2" Workshop!

Example D: Multivariate analysis – Confirmatory Factor Analysis

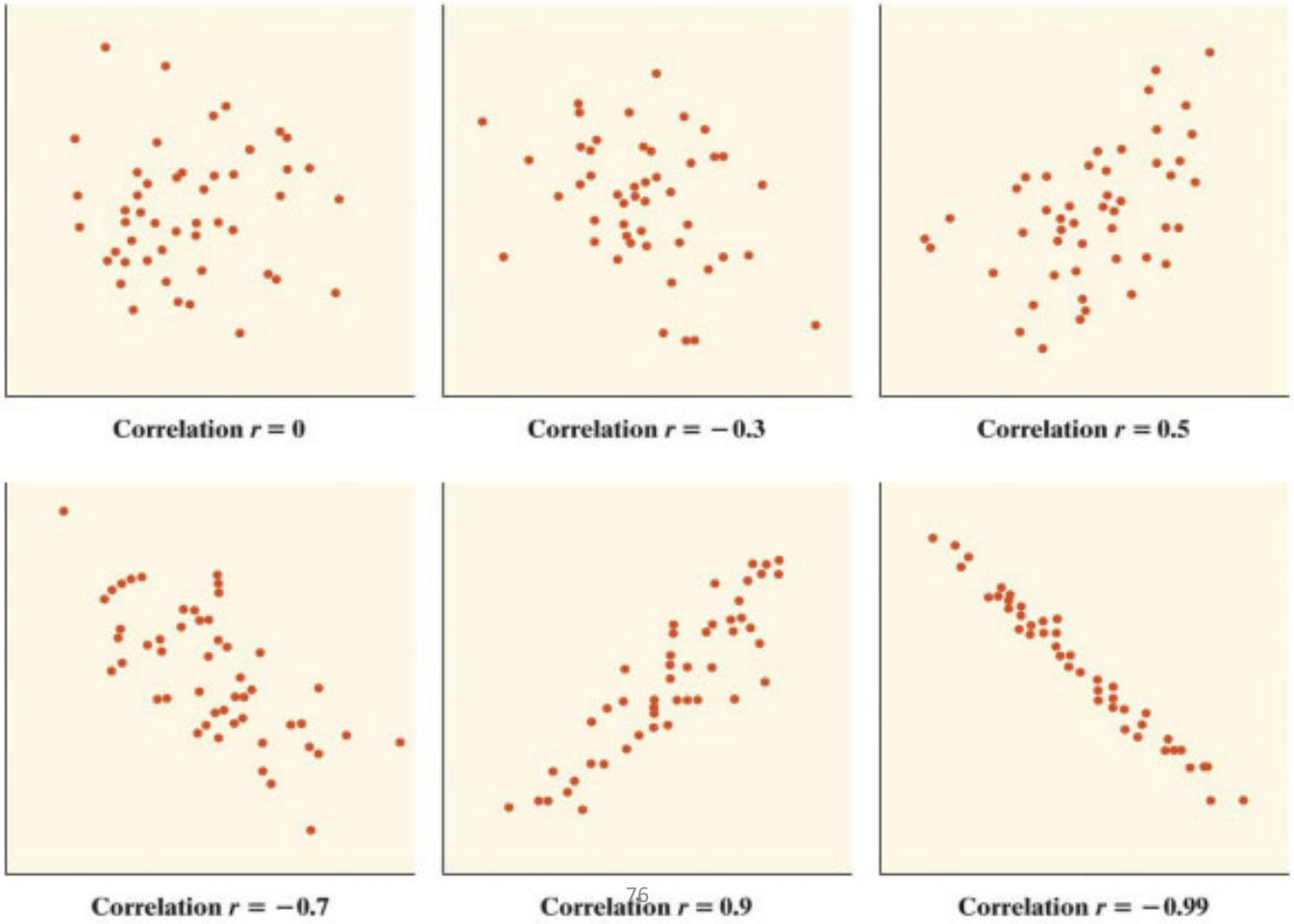
Scenario: To test if 'SPSS statistical software Anxiety' explains the common variance among 7 questionnaire items:

1. I dream that Pearson is attacking me with correlation coefficients.
2. I have little experience with computers.
3. All computers hate me.
4. I have never been good at mathematics.
5. My friends are better at statistics than me.
6. Computers are useful only for playing games.
7. I did badly at mathematics at school.

Example adapted from: "A practical introduction to Factor Analysis: Confirmatory Factor Analysis". UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis-confirmatory-factor-analysis/

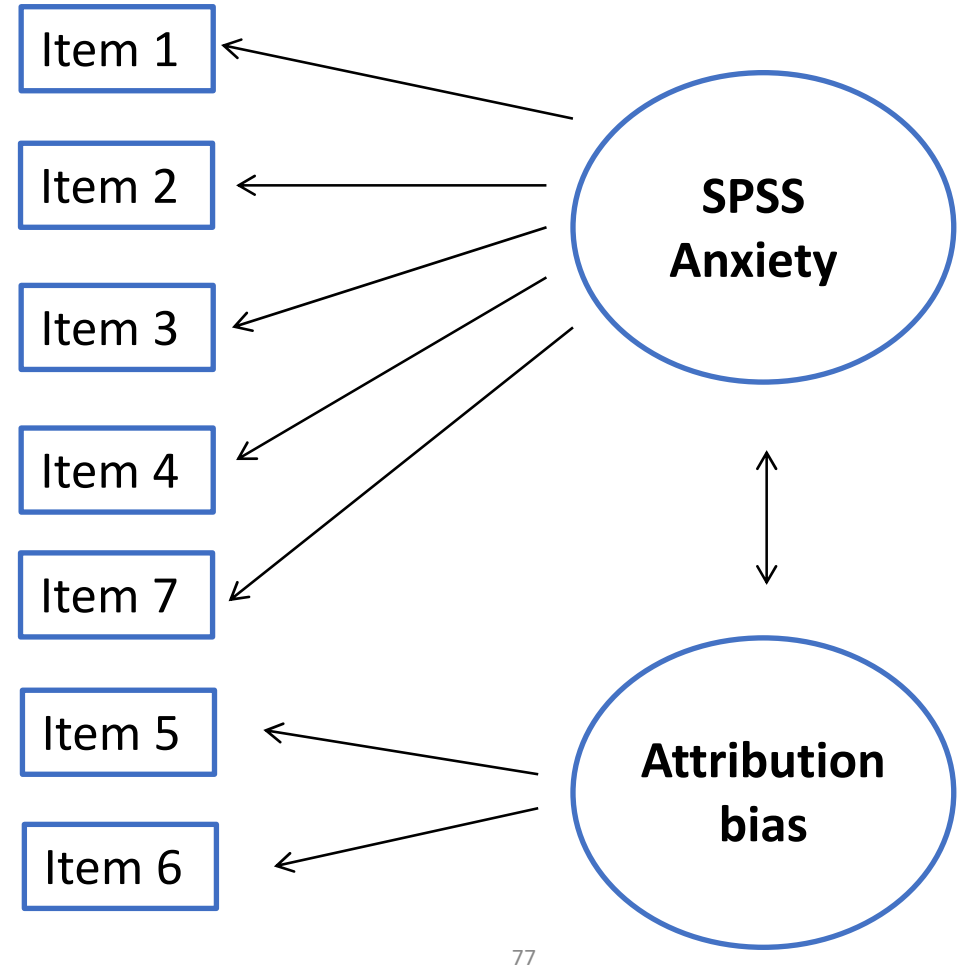
Example D: Multivariate analysis – Confirmatory Factor Analysis

Step 5: EDA – scatter plots + Pearson's correlation coefficient r; correlation matrix

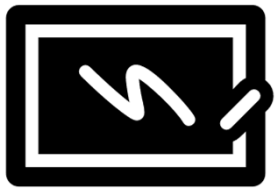


Example D: Multivariate analysis – Confirmatory Factor Analysis

Step 6: Inferential analysis



Final notes on Step 6: Inferential Analysis



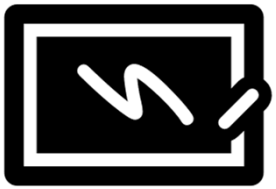
- We only showed some more common examples - there are many different types of analyses, e.g. consider
  - Other Linear Models extensions such as logistic regression and more complex mixed models - see our SIH 'Linear Models' training!
  - Survival Analysis for 'time-to-event' outcome data – see our SIH training!
  - Survey Data analysis – see our SIH training!
  - Other Multivariate Analyses – for example PCA, Factor Analysis - see our SIH training!

- Start simple and increase complexity step by step
- Always consider/check the test/model assumptions
- Report 95% CI's for estimates, e.g. predicted means/ probabilities/rates
- For basic analyses consider more powerful analyses first and use less powerful tests if assumptions are violated, e.g.:

  → 2 sample t-test with equal or unequal variance for means before Mann-Whitney Test

  → Chi-squared test to compare proportions before Fisher's exact test

Inferential analysis



- ➔ Use knowledge of variable types to guide you through the systematic tree roadmap
- Don’t forget to check test/model assumptions!

Outcome (i.e. dependent) variable	Exposure (i.e. independent) variable	Statistical test	Key assumptions[1]
Unpaired data			
Dichotomous/binary, nominal or ordinal data	Two or more groups (i.e. dichotomous/binary, nominal or ordinal data)	Chi-square test[2]	Expected numbers are <5 in <20% of cells
As above	As above	Fisher's exact test[2]	
Ordinal data	Two groups (i.e. dichotomous/binary data)	Mann-Whitney U test (Wilcoxon rank-sum test) [2]	
Ordinal data	Three or more groups (i.e. nominal or ordinal data)	Kruskal-Wallis test[2]	
Continuous data	Two groups (i.e. dichotomous/binary data)	2-sample t-test[3]	Variance same in both groups Residuals have normal distribution
Continuous data	Two groups (i.e. dichotomous/binary data)	2-sample t-test for unequal variances[3]	Residuals have normal distribution
Continuous data	Two or more groups (i.e. dichotomous/binary, nominal or ordinal data)	One-way ANOVA[3]	Variance same in all groups Residuals have normal distribution

Statistical inferential analysis roadmap

How many outcome/ dependent variables?	What type of outcome?	How many predictor/ independent variables?	What type of predictor?	If a categorical predictor, how many categories?	If a categorical predictor, are the same or different entities in each category?	Assumptions of linear model met, yes use GLM	Assumptions of linear model not met, use non parametric or bootstrap

Adapted from "Discovering Statistics using IBM SPSS Statistics" by Andy Field

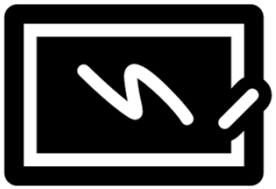
Data Analysis – some terminology:

- Univariate – one outcome per analysis
- Multivariate – multiple outcomes in the same analysis
- Multivariable – multiple explanatory variables

- Linear models (LM – continuous outcome)
- Generalised linear models (GLM – categorical outcomes, e.g. binary, ordinal, multinomial (for nominal outcome data) or Poisson regression (for count/rate outcome data)

- Mixed models (i.e. LM or GLM with random effect = LMM or GLMM)
  - Data clustered in space or time, e.g. repeated measures/ longitudinal)

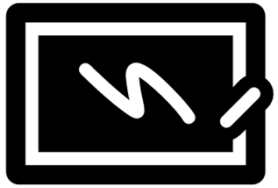
R resources



- There is a large online community of R users contributing free ‘packages’ with data analysis functions, which leads to many ways of doing an analysis in R. This can be confusing. We recommend using tidyverse packages.

Starting points for conducting descriptive data analyses and basic inferential tests are:

- Learning the R Tidyverse Welcome (linkedin.com)
- Learning R markdown https://www.linkedin.com/learning/creating-reports-and-presentations-with-r-markdown-and-rstudio/report-your-data-with-r-markdown?u=2196204
- Tidyverse style guide and lintr The tidyverse style guide



Upcoming SIH statistical workshops:

Check out the SIH training calendar or mailing list for details on all upcoming trainings: https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/training-calendar.html

- 25th August 3.15pm - 4pm - Statistical drop-in session
- 15th September – Hacky Hour
- 29th September – Meta-Analysis, Linear Models 3, Model Building
- 13th October - Surveys 1 and 2, Multivariate Analysis
- 20th October – Hacky Hour
- 27th October – Linear Models 1 + 2, Model Building

End of workshop

Survey

Asking for more help

THE UNIVERSITY OF SYDNEY

Further Assistance at Sydney University

SIH

- Workshops
  - **Create your own custom programmes tailored to your research** needs by attending more of our **Statistical Consulting** workshops. Look for the statistics workshops on our training page https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats. Some we recommend are:
    - Research Essentials
    - Experimental Design
    - Power Analysis
  - **Other SIH workshops** https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html
  - **Training** **Sign up to our mailing list** to be notified of upcoming training: mailman.sydney.edu.au/mailman/listinfo/computing_training
- **Hacky Hour** www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html OR Google "Sydney Hacky Hour"
- **1on1 Consults** can be requested on our website www.sydney.edu.au/research/facilities/sydney-informatics-hub.html OR Google "Sydney Informatics Hub"

OTHER

- **Open Learning Environment (OLE) courses**
- **Linkedin Learning**: https://linkedin.com/learning/
  - **SPSS** https://www.linkedin.com/learning/machine-learning-ai-foundations-linear-regression/welcome?u=2196204

Workshop Survey

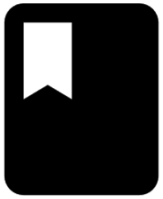
- Please use this link to provide feedback on this workshop:

https://redcap.sydney.edu.au/surveys/?s=FJ33MYNCRR&training=37

Questions?

Contact: Kathrin Schemann – Kathrin.Schemann@sydney.edu.au

Other references used:



- Library resource for conducting a systematic review

https://library.sydney.edu.au/research/systematic-review/?section=analyse-and-interpret

- Good enough practices in scientific computing.

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK.

PLoS Comput Biol. 2017 Jun 22;13(6):e1005510. doi: 10.1371/journal.pcbi.1005510. eCollection 2017 Jun.