Research Essentials: Analysing your data

Presented by
Dr Kathrin Schemann
Sydney Informatics Hub
Core Research Facilities
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 for use of workshops and workflows:

"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

What is a workflow?

- Every statistical analysis is different, but all follow similar paths. It can be useful to know what these paths are

- We have developed practical, step-by-step instructions that we call '*workflows',* that can you can follow and apply to your research

- We have a general research workflow that you can follow from hypothesis generation to publication

- And statistical workflows that focus on each major step along the way (e.g. experimental design, power calculation, model building, analysis using linear models/survival/multivariate/survey methods)

Statistical Workflows

– Our **statistical workflows** can be found within our workshop slides

– **Statistical workflows** are software agnostic, in that they can be applied using any statistical software

– There may also be accompanying **software workflows** that show you how to perform the statistical workflow using particular software packages (e.g. R or SPSS). We won't be going through these in detail during the workshop. If you are having trouble using them, we suggest you attend our monthly Hacky Hour where SIH staff can help you.

During the workshop

Ask **short questions** or clarifications during the workshop (either by Zoom chat or verbally). 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 questions will be encountered throughout the workshop.

Research Essentials Workshop overview

I.   8-step general 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:
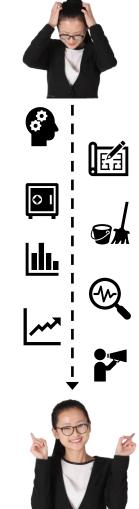Step 5: Exploratory data analysis
Step 6: Inferential analysis

8-step general research workflow

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 modelling:
Inferential predictive statistics versus machine learning predictive analytics

Inferential predictive statistics:

– Interested in knowledge about the data/ predictors, e.g. risk factors for disease or demographic factors associated with purchasing decisions or factors driving temperature increase in climate change.

– Generalises from the sample to the population (makes inference)

Predicting temperature rise in climate change.

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.

Very accurately verify fingerprint to unlock a mobile phone.

Ecosystem of SIH statistical 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	**Research Essentials**
6. Inferential analysis	Linear models 1-3 + Model Building Survival analysis Meta-analysis Survey design and analysis 1 + 2 Multivariate Analysis 1 – Dimension Reduction
7. Predictive modelling	[Predictive analytics: *Introduction to machine learning in R/Python (SIH 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

SIH Training and Outreach

Research Integrity team

Data Science	Research Computing	Research Data Management	Statistics	Outreach
R, Python, Matlab	High Performance Computing	REDCap	Data Visualisation	Hacky Hour
Machine Learning	Cloud computing	eNotebook	Experimental design	Data Bonanzas
Geospatial Analysis	GPUs	Research data storage	Survey design	eResearch Conference
Deep Learning	Containers	Research data management	Linear models, GLMMs	Summer Schools
Natural Language Processing	Workflow languages		Survival analysis	Innovation Week Coding Challenge
Bioinformatics	Simulation		Meta-Analysis	

USyd Core Facilities @Sydney_CRF · Jul 30
A rare #HackyHour at #SydneyInfomaticsHub's home offices. Thanks to Louis Mercorelli (@LouisMercorelli), Nathaniel Butterworth, Kristian Maras, Olya Ryjenko and Tracy Chew.

USyd Core Facilities @Sydney_CRF · Mar 17
#SydInformatics' Dr. Darya Vanichkina running a #MachineLearning in #R workshop this rainy Monday morning 🌧️ Sky water ain't stoppin no coding! 🌧️

sydney.edu.au/informatics-hub/training

Other Sydney uni research integrity 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 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

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Research Data Management

Research data that is managed optimally improves research efficiency and reach, as well as ensuring its integrity and security, and meeting legislative/policy/funding/publishing requirements.

The Research Data Consulting team assists researchers to enhance their research productivity and improve data management practices. They provide:

– Short consultations to integrate digital tools and data management into your research

– Training and functional support for university supported tools/platforms

Research Data Consulting

Research Integrity & Ethics Administration
digital.research@sydney.edu.au

Book a consultation

Supported platforms

- eNotebook
- REDCap
- Research Data Store (RDS)
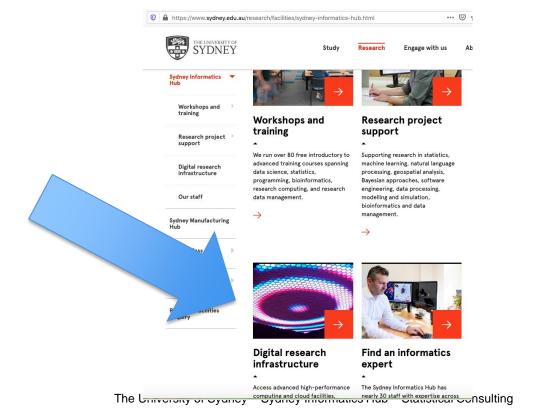- OneDrive (Office365)
- Github
- CloudStor

DO NOT USE

- Google Drive
- Survey Monkey
- Portable Drives e.g. USB

eNotebooks

DashR & Research Data Management Plans

Research Data Store

REDCap

Office 365

GitHub

cloudstor
FileSender

Further information: How do I manage my research data?

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Research computing –
when your analysis outgrows your computer
sydney.edu.au/informatics-hub/

Software training (OLE's + LinkedIn Learning)

HDR degree includes compulsory coursework – some units include R coding

Workflow Step	SIH software courses offered
1. H_0 generation	
2. Study design	
3. Collect/store data	
4. Data cleaning	**SPSS Statistics Essential Training**
5. Exploratory analysis	**R Essential Training** **Learning the R Tidyverse**
6. Inferential analysis	**R Essential Training**
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 University of Sydney – Sydney Informatics Hub – Statistical Consulting

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

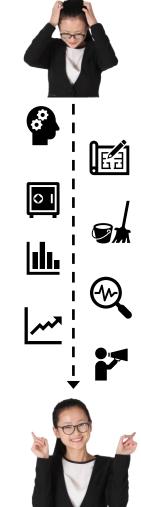| Workflow Step	Other training
1. Hypothesis generation	Library research support consultancy
2. Study design	Hacky hour; One-on-one statistical consultancy
3. Collect/store data	Hacky hour; RDM drop in session or one-on-one RDM consultancy
4. Data cleaning	Hacky hour; Drop in or one-on-one statistical 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
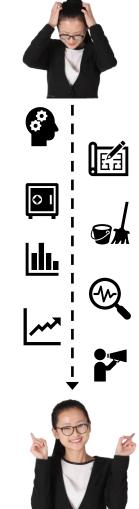
General Research Workflow

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**

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**

The first question: which car will you take?

Getting from step 3 to step 8 will involve using software. Will it be:

Graphical User Interface? (GUI)
- Interactive, point and click
- Easier to get started

A Good GUI

CLI

Command line interface (CLI)
- Writing code
- Easier to handle complex and/or large data sets

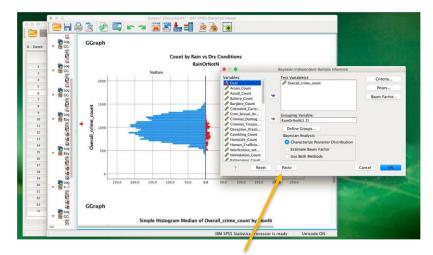The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Software choice: programming (CLI) versus menu-driven (GUI)

– Which are you more familiar or comfortable with?

– How do you record your analysis for reproducible research?

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

R code   versus   SPSS

`>t.test(x,y)`

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

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

3. Collect/store your data

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

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, editing analysing in the same sheet
  - Statistical software generally doesn't allow easy editing once you have entered your data

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"

Version control - keeping track of files
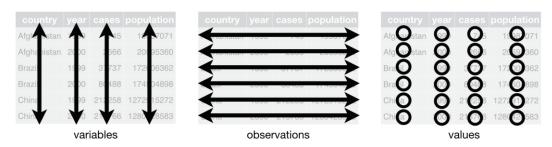
- Use a separate directory for each discrete analysis
- When processing data and intermediate files save with a new name
- save frequently so if you lose a version you do not have to redo too much work.
- For collaborative research consider using eNotebook or other version control systems, e.g. Git (free)
- Create a log file in the same directory and use version control (e.g. name sequentially, date/time stamp, for example:
  - *"20230208 stats101 workshop v2.0.xlsx" (orders files chronologically)*

Example of a version log file in Excel:

File name	Description	# Obs	#Vars
Mydata_v1_30102022.csv	Original data entry by KS, 1 record per person	250	34
Mydata_v2_01112022.csv	Eligible records only based on study inclusion criteria with new variables created for analysis: BMI calculated from recorded height and weight; babies age processed to be consistently in months instead of days and weeks as well; number of pets categorised (none, 1-2, 3+)	204	37
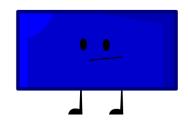
Data formats – tidy data

- **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, or just have the demographic data repeated when transforming to long format

- **Wide and long can become relative terms especially if you have clusters of subjects**

- **Tidy data is an absolute term, which describes data transformed to:**
  - One variable in each column
  - One observation per row
  - One value per cell

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

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

b) Organising a dataset for analysis

- **Most programs read in data in a rectangular format:**
  - E.g. A text file – you can read it in Notepad or any text editor or Excel, csv etc.
  - 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, animal, plant, plot, farm, machine, business, school, hospital 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/.txt)

Pitfalls when coming from Excel:

- Watch out for:
    - Merged cells
    - Cell comments
    - Colour coding
    - Blank rows
    - Data in multiple sheets
    - Particular coding of missing data/blanks/non-applicable

- Deal with the above in Excel before exporting to text. Sometimes these have been added to annotate the data, or make it easier to read. Other times, they are *part* of the data and must be represented some way in a text file

- A good summary of these pitfalls is provided in <u>this paper</u>

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

b.) Data formats - transformations
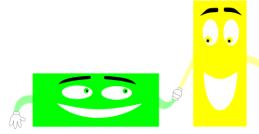
Animal ID	Time 1	Time 2	Time 3
1	50	55	60
2	47	49	50
…	…	…	…

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

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

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

- Specify type of variable: ensure your analysis software knows whether a variable is continuous (numeric), categorical/factor/string (text)

- Label variables, either within the software or by keeping your own record (e.g. Age = Age at interview in years)

- Label variable 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

4. Data cleaning

THE UNIVERSITY OF
SYDNEY

Data cleaning: data wrangling and data dictionary

- Data cleaning involves examining[*] the variables in the dataset and creating new variables for analysis by recoding/processing variables as required
- Use short but informative variable names; it's a good idea to have a data dictionary
- Names should keep track of transformations/recoding, e.g.
  - age = original data in years
  - Age_c2 = age categorised into two categories (young vs old)

A simple example of a data dictionary:

	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

[*] **More on how to examine, i.e. describe the distribution of, variables later in this workshop.**

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

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

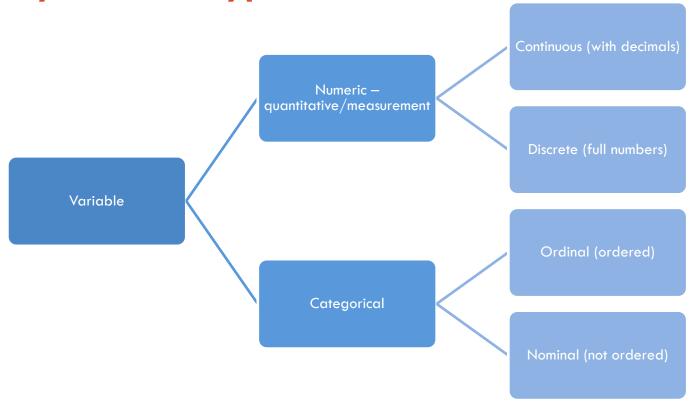   - <u>Also covered in "Good enough practices for scientific computing"</u>

NEWS flash: Variable types are important!

Why worry about variable types?

- **Variable types determine the appropriate statistical methods for 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 numeric variable to a categorical variable depending on its distribution

i.) Identify variable types:

```
                                          ┌──────────────────────────────┐
                                          │  Continuous (with decimals)  │
                    ┌──────────────────┐ ─┤                              │
                    │    Numeric –     │  └──────────────────────────────┘
              ┌─────┤ quantitative/    │
              │     │  measurement     │  ┌──────────────────────────────┐
              │     └──────────────────┘ ─┤   Discrete (full numbers)    │
┌──────────┐  │                           └──────────────────────────────┘
│ Variable │ ─┤
└──────────┘  │                           ┌──────────────────────────────┐
              │     ┌──────────────────┐ ─┤     Ordinal (ordered)        │
              └─────┤   Categorical    │  └──────────────────────────────┘
                    └──────────────────┘
                                          ┌──────────────────────────────┐
                                         ─┤    Nominal (not ordered)     │
                                          └──────────────────────────────┘
```

Variable types

CONTINUOUS
measured data, can have ∞ values within possible range.

I AM 3.1" TALL
I WEIGH 34.16 grams

DISCRETE
observations can only exist at limited values, often counts.

I HAVE 8 LEGS
and
4 SPOTS!

@allison_horst
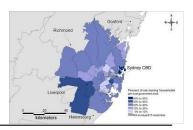
Variable types

Variable types – example Sydney cat study

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

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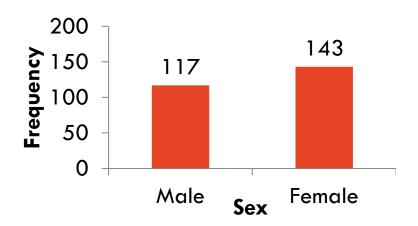   - Median
    - Mode
    - Quartiles
    - Percentiles

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

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
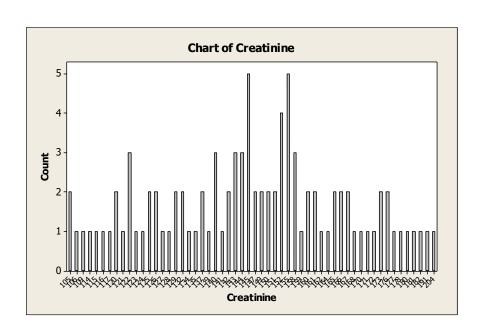   - Quartiles
    - Percentiles

How to summarise the variable: Creatinine

– **Creatinine levels (μmol/L) of 96 cats**

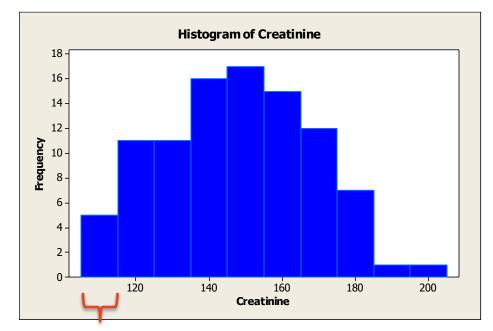|												
170	164	173	106	160	139	105	178	140	140	172	155	122
152	125	114	144	155	180	137	150	105	132	120	145	162
166	176	137	152	155	122	145	123	165	145	161	124	128
182	171	155	149	158	161	177	158	151	147	142	143	126
144	159	166	117	167	127	142	149	120	151	125	121	155
181	191	134	158	143	147	109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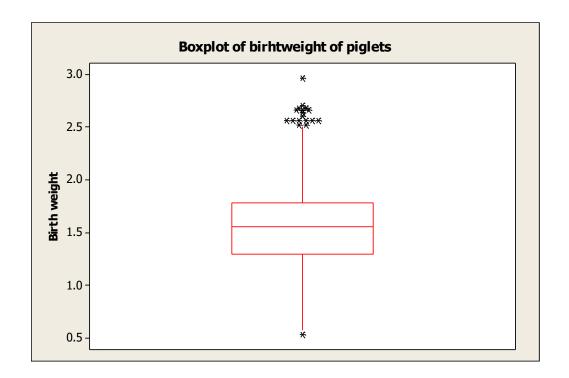

Frequency per (equally sized) bin

Class/bin:
104-110

Shapes of the distribution

Symmetric Distribution

Histogram of bwt

Asymmetric Distributions

Asymmetric Distributions

Histogram of Gene+

Boxplot

Boxplot of birhtweight of piglets
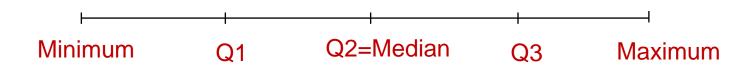
Summarising the picture

- **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)

| Minimum | Q1 | Q2=Median | Q3 | Maximum |

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

– **Categorical variables**

  – Frequency table

  – Bar chart

– **Numeric 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

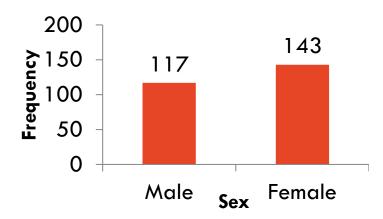Don't forget to check for missing data/NA's!

Summary – descriptive data analysis for individual variables
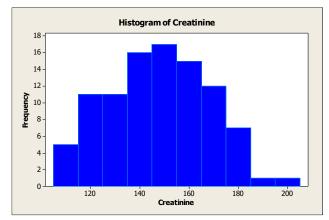
- A categorical variable
    - Frequency table
    - Bar chart

- A numeric variable
    - Histogram
    - Box-and-whisker plot

    - Mean ± std deviation
    - Median and quartiles

Histogram of Creatinine

OK, so variable types are important for describing the distribution of each variable and checking/ cleaning each individual variable, but…

What about my Research question and analysis??

THE UNIVERSITY OF
SYDNEY

What is the Research Question?

For any analysis we need to be clear what the functional classification of the variables in the data set is, e.g. we want to investigate the effect of smoking on lung disease:

Smoking (yes/no)

Predictor
Explanatory variable
Risk factor
Independent variable*

Lung disease (yes/no)

Response
Outcome
Dependent variable

Depending on who taught you statistics, your discipline, stats software or text book you may have come across any of these terms or more!

*this term is frequently used, but we don't promote it as predictors may be correlated and hence are not independent

Other functional classifications for variable types

- Covariate: a measured predictor (numeric predictor variable)
- Factor (a categorical predictor variable)

- **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 (batch of reagent, source of lab mice), subject ID, etc.
  - Treatment: Variables of interest, e.g. diet, drug treatment, intervention etc. NB: The 'levels' of a 'treatment variable' might include 'control (placebo)', 'treatment 1 (drug 1)', 'treatment 2 (drug 2)'

  ➔ More information on design variables in our "Experimental design" Workshop!

What is the outcome variable?

- **Review study aim and objectives**
  - E.g. vaccine RCT - daily morbidity outcome data could be analysed as:
    - mean daily rate (average – numeric)
    - cumulative morbidity (sum – numeric)
    - peak morbidity (maximum – numeric)
    - outbreak presence/absence (binary group – categorical)
    - time to infection/disease outbreak (time to binary event data – survival analysis)

More 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 (previous step)**
  - 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

A quick primer to Step 5: EDA

Step 5: Exploratory Data Analysis (EDA)

→ depends on the analysis/variables involved

→ basic EDA: plot the relationship of each predictor
   with the outcome; reconsider data processing,
   e.g. need to collapse more categories?

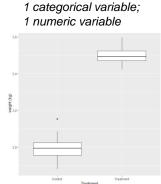## Two categorical variables
- Contingency table
- Side-by-side bar chart
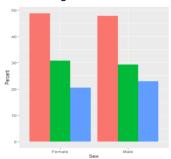
A categorical and a numeric variable
- Tabulate summary statistics by groups
- Box-and-whisker plot by groups

Two numeric variables
- Scatter plot and correlation coefficient r
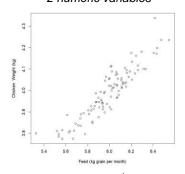
2 categorical variables

Side-by-side bar charts

1 categorical variable;
1 numeric variable
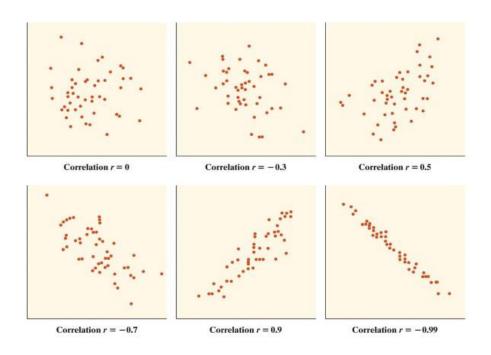
2 numeric variables

xy scatter plot

XY scatter plot and Pearson correlation coefficient r

A quick review of correlation coefficient r to describe the relationship of two numeric variables in a scatter plot; r = 0 means no relationship – data points in a horizontal line

Correlation r = 0   Correlation r = −0.3   Correlation r = 0.5

Correlation r = −0.7   Correlation r = 0.9   Correlation r = −0.99

Examples for reporting of descriptive analyses

-Plotting gives a quick visual summary during EDA + highlights issues

-Tables are more publication friendly as they save space

Numeric variables:
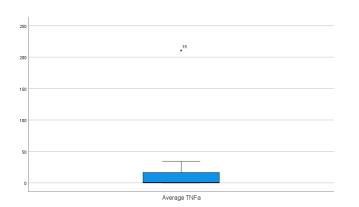
Table 1: Summary statistics for numeric predictor variables in a study of animal reproductive performance

	Minimum	Q1	Median	Q3	Maximum	Mean	Standard deviation
TNF-alpha	0.0	0.0	0.4	17.0	210.6	18.4	47.8
Age	4.0	6.0	8.0	17.0	22.0	10.5	5.9

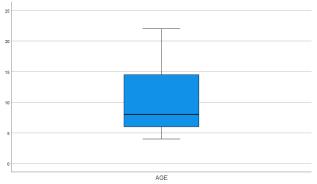The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Examples for reporting of descriptive analyses

-Plotting gives a quick visual summary during EDA + highlights issues

-Tables are more publication friendly as they save space

Categorical variables:
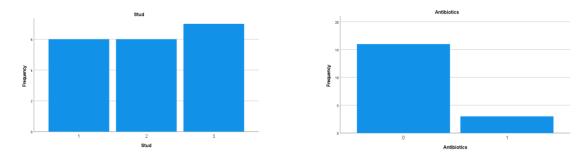
Table 2: Summary statistics for categorical predictor variables in a study of animal reproductive performance

Predictor	Categories	N (%)
Stud	1	6 (32)
	2	6 (32)
	3	7 (36)
Antibiotic	No	16 (84)
	Yes	3 (16)

Some analysis examples

5. Exploratory Data Analysis (EDA)
6. Inferential analysis

Data Analysis Workflow: 4 Examples

A – Linear Models examples:

Simple regression, ANOVA, ANCOVA, Repeated measures.

B – Extended Linear Models example:

Survival Analysis

C – Extended Linear Models example:

Generalised Linear Model – Poisson regression

D – Multivariate Analysis

Confirmatory Factor Analysis

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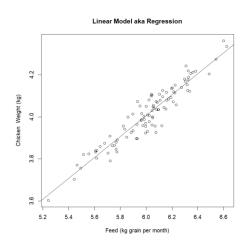
Scenario: We are interested in studying a continuous outcome variable, e.g. weight gain (kg) or blood cell count (cells/μL)

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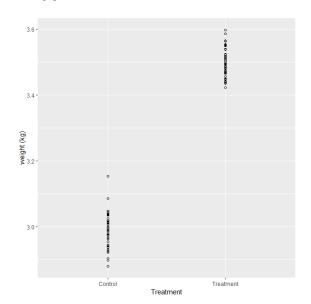
Step 6: Inferential analysis – fit a linear regression line and test if the slope is different from 0; p <0.001; report slope/regression estimate and 95% CI.

Linear Model aka Regression

The University of Sydney – Sydney Informatics Hub – Statistical Consulting
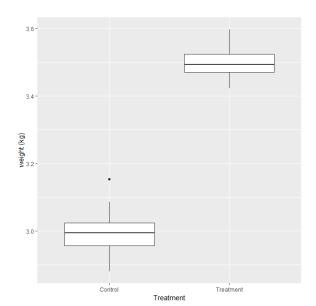
Example A: Linear models – Control versus Treatment experiment

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.

The University of Sydney – Sydney Informatics Hub – Statistical Consulting
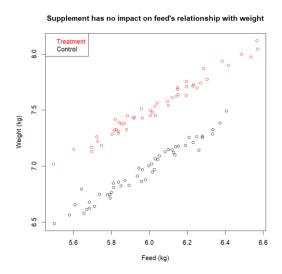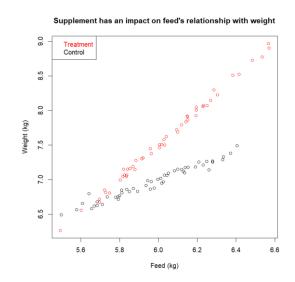
Example A: Linear models – ANCOVA - ANOVA with a continuous covariate

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

Step 6: Inferential analysis – ANCOVA/ multivariable regression

without interaction
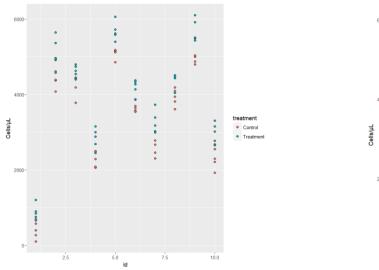
with interaction

Example A: Linear models – Repeated measures
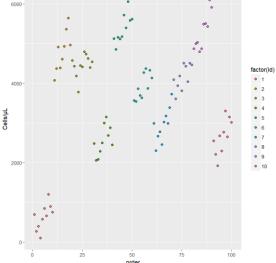
Scenario: n=10 with 5 before and 5 after treatment measurements

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

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Example B: Survival Analysis

Scenario: Worcester Heart Attack Study (WHAS)

Aim: To examine time trends in the incidence rate of acute heart attacks

Objective: Investigate if different demographic and clinical factors are associated with the time to a heart attack.

Data: longitudinal, observational data
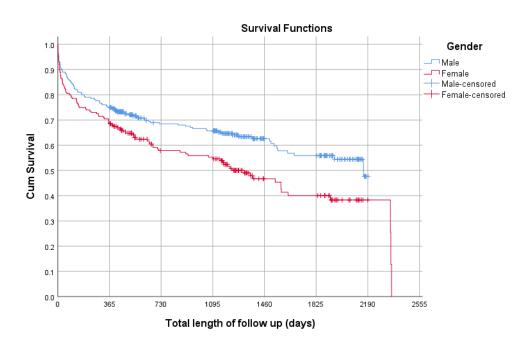
Outcome: heart attack – yes/no

Predictors: demographic and clinical data

Key feature: Data is censored – see our Introduction to Survival Analysis WS

Example B: Survival Analysis

Step 5: EDA – Kaplan Meier curve is the EDA plot for Survival Analysis

Example B: Survival Analysis

Step 6: Inferential analysis:
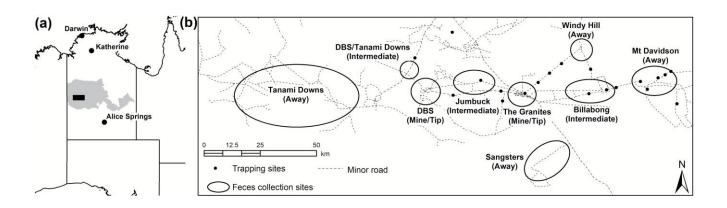
There is a significant difference in survival between males and females (by log-rank test)

Median survival for males: 2160 days [95%CI: not calc]

Median survival for females: 1317 days [95% CI 970-1664]

Example C: Extended Linear Model – Generalised Linear Model (GLM) / Poisson regression
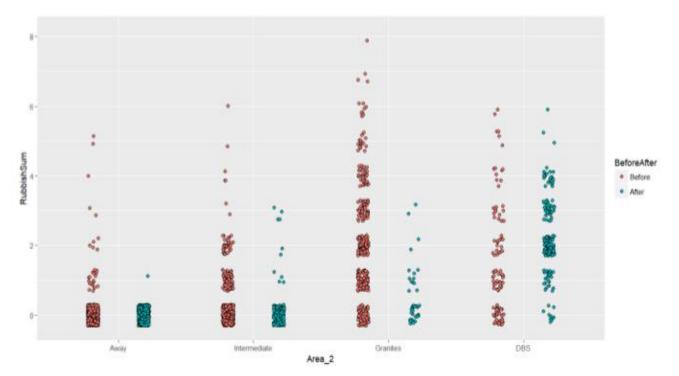
Scenario: Quasi-before-after-control-impact experiment to assess how dingoes respond to a decline in anthropogenic foods.

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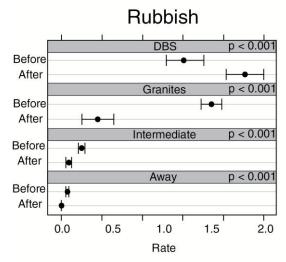
Step 5: EDA – plot count data by area and treatment

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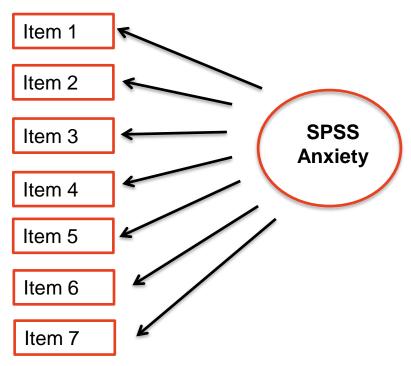
➔ **To learn more about logistic and Poisson regression attend our SIH "Linear Models 2" Workshop!**

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Example D: Multivariate analysis – Confirmatory Factor Analysis

Scenario: To test if a factor model '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/

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Example D: Multivariate analysis – Confirmatory Factor Analysis

Scenario design - Confirm SPSS Anxiety as a factor explaining the common variance among the 7 items

Example D: Multivariate analysis – Confirmatory Factor Analysis
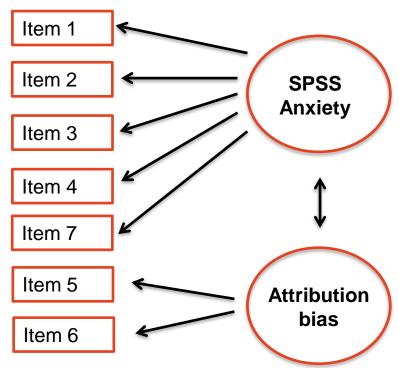
Step 5: EDA – scatter plots + Pearson's correlation coefficient r; correlation matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Q1	1						
Q2	-0.34	1					
Q3	0.44	-0.38	1				
Q4	0.40	-0.31	0.40	1			
Q5	0.22	-0.23	0.28	0.26	1		
Q6	0.31	-0.38	0.41	0.34	**0.51**	1	
Q7	0.33	-0.26	0.35	0.27	0.22	0.30	1

Example D: Multivariate analysis – Confirmatory Factor Analysis

Step 6: Inferential analysis- To test if a 2 - factor model 'SPSS statistical software Anxiety' and 'Attribution bias' explains the common variance among 7 questionnaire items

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

Further R resources

– University of Sydney OLE units of study
– 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:

– R Essential Training
– Learning the R Tidyverse
– Learning R markdown
– The tidyverse style guide

Further Assistance at Sydney University

SIH

– [Statistical Consulting website](): containing our workshop slides and our favourite external resources (including links for learning R and SPSS)
– [Hacky Hour]() an informal monthly meetup for getting help with coding or using statistics software
– 1on1 Consults can be requested [on our website]() (click on the big red 'contact us' link)

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.]()
– [Other SIH workshops]()
– [Sign up to our mailing list]() to be notified of upcoming training

Other

– Open Learning Environment (OLE) courses
– [Linkedin Learning]()

Request support via our webpage

https://sydney.edu.au/research/facilities/sydney-informatics-hub.html
(google "Sydney university SIH statistical consulting")

THE UNIVERSITY OF SYDNEY

Study   **Research**   Engage with us   About us   News & opinion

Home / Research / Facilities / Sydney Informatics Hub

← Home

← Research

← Facilities

Research and prototype foundry

Sydney Analytical

Sydney Cytometry

Sydney Imaging

Sydney Informatics

Sydney Informatics Hub

Enabling excellence in data and compute intensive research

We provide support, training, and expertise in research data management, statistics, data science, software engineering, simulation, visualisation, bioinformatics, and research computing.

Share

Contact Us

Request a consult or advice for your research project

→

Click here for support

Request a Quote

How to use our workshops

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

We recommend our Experimental Design and Sample Size Workshops

Experimental Design Workshop

– Far too many researchers think they know all they need to in this area. We commonly see designs that could be substantially improved for stronger causal inference and improved results which leads to publication in higher impact journals (amongst other benefits).

– Even if you have already collected your data it is well worth attending since it may improve your write up and analysis e.g. we had a client who didn't realise they had a very strong Before/After Control/Impact (BACI) design.

Sample and Power Workshop

– Shows the steps and decisions researchers need to make when designing an experiments to ensure sufficient sample e.g. Power, minimum required to fit the necessary model, etc.

– Also how much Power the study has i.e. does it have sufficient power to detect the effects you expect to see, or is your study a complete waste of time and resources.

A reminder: 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 for use of workshops and workflows:

"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

We value your feedback

We want to hear about you and whether this workshop has helped you in your research. What **worked** and what **didn't work.**

We actively use the feedback to improve our workshops.

Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (*promise!*)

You will receive a link to the anonymous survey by email

The University of Sydney – Sydney Informatics Hub – Statistical Consulting

Appendix: Data Analysis – some terminology:

- Univariate – involving one variable, e.g. one outcome per analysis; analysis with one predictor variable
- 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)