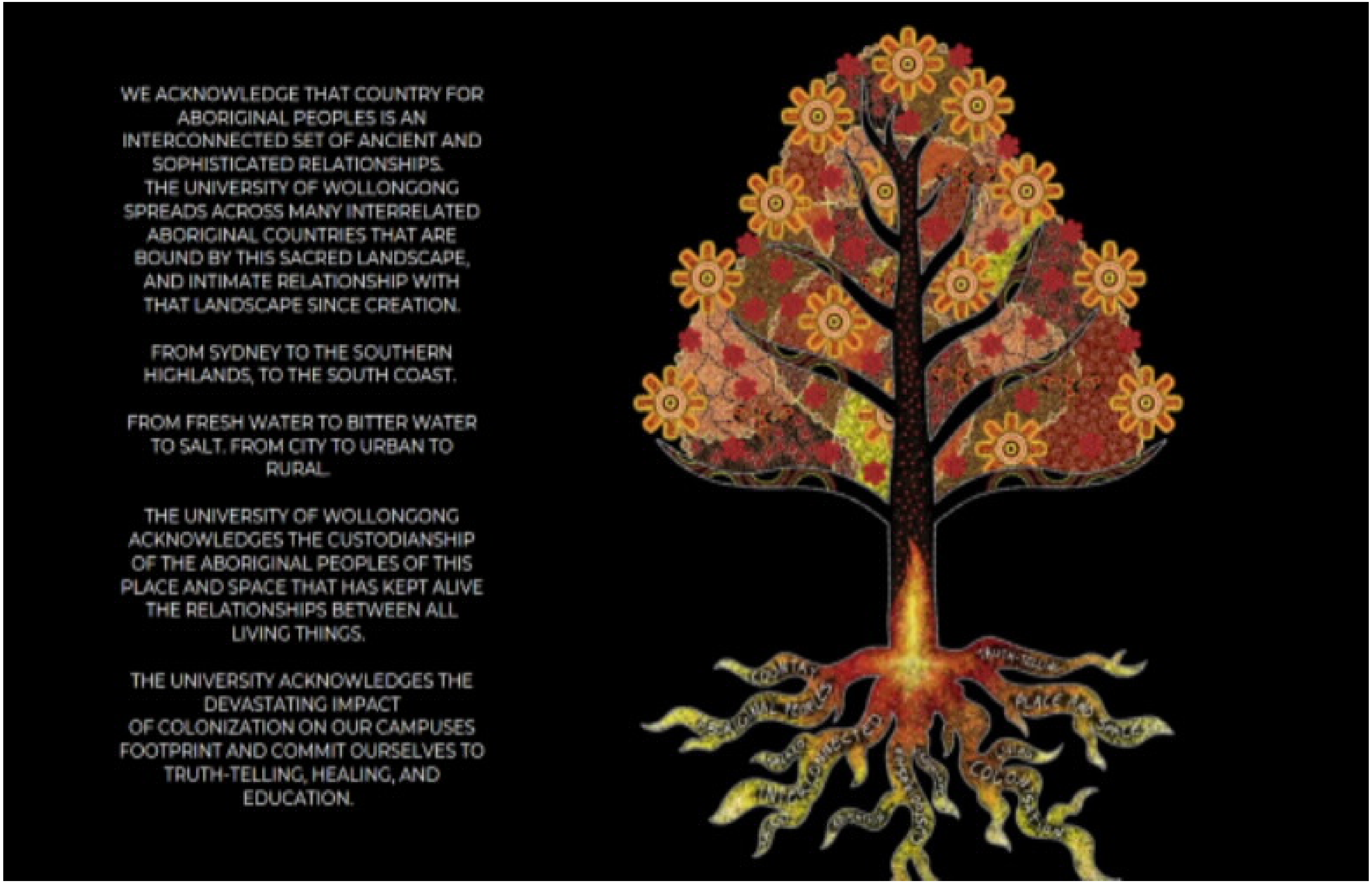
Statistical Analysis Plans and Data Management

DATA AND DECISION SCIENCE NETWORK MEETING
DATA SCIENCE AND STATISTICS COMMUNITY OF PRACTICE
NOVEMBER 2023



WE ACKNOWLEDGE THAT COUNTRY FOR ABORIGINAL PEOPLES IS AN INTERCONNECTED SET OF ANCIENT AND SOPHISTICATED RELATIONSHIPS.
THE UNIVERSITY OF WOLLONGONG SPREADS ACROSS MANY INTERRELATED ABORIGINAL COUNTRIES THAT ARE BOUND BY THIS SACRED LANDSCAPE, AND INTIMATE RELATIONSHIP WITH THAT LANDSCAPE SINCE CREATION.

FROM SYDNEY TO THE SOUTHERN HIGHLANDS, TO THE SOUTH COAST.

FROM FRESH WATER TO BITTER WATER TO SALT. FROM CITY TO URBAN TO RURAL.

THE UNIVERSITY OF WOLLONGONG ACKNOWLEDGES THE CUSTODIANSHIP OF THE ABORIGINAL PEOPLES OF THIS PLACE AND SPACE THAT HAS KEPT ALIVE THE RELATIONSHIPS BETWEEN ALL LIVING THINGS.

THE UNIVERSITY ACKNOWLEDGES THE DEVASTATING IMPACT
OF COLONIZATION ON OUR CAMPUSES FOOTPRINT AND COMMIT OURSELVES TO TRUTH-TELLING, HEALING, AND EDUCATION.

Introductions

Dr Brad Wakefield
Statistical Consultant in the Stats Consulting Centre.
Interests in data privacy, probability theory, statistical inference, and data analytics.
Passion for ethical applications of data science methods in research and industry.
Enjoys learning and collaborating with other disciplines and solving real-world problems.
Always up for a chat.

- Professor Marijka Batterham
- Co-Ordinator Data & Decision Science Initiative
- Director NIASRA
- Director Stats Consulting Centre
- Passionate about data literacy
- Likes learning ML & exploring new packages

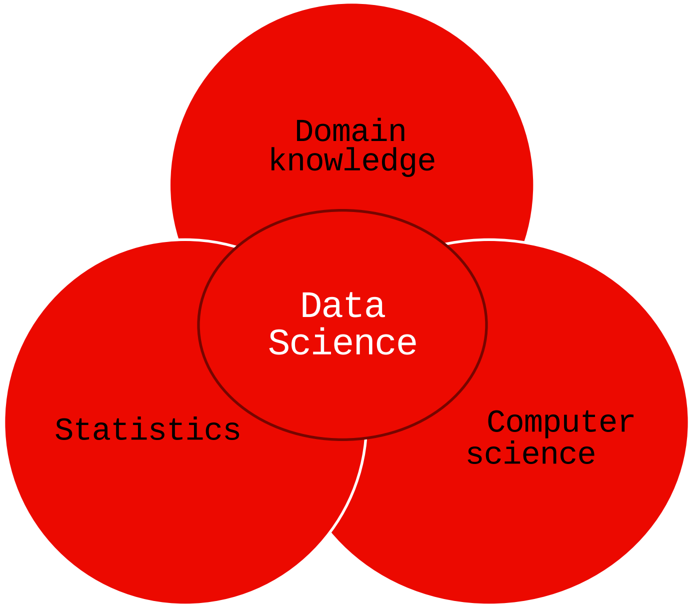
NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

UOW Data & Decision Science Initiative

- The Data and Decision Science Initiative is part of the UOW strategic Plan (2.5 Transformative technologies)
- Developed from a 2019 review and recommendations of “Big Data” and Health Informatics at UOW
- Commenced July 2021
- Led by NIASRA (Marijka Batterham Co-Ordinator)

Data Science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formation and hypothesis testing



NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Data & Decision Science Initiative

four key areas of focus

Research: virtual network and working groups of Data and Decision Science researchers

- Focal point for coordinating the development of Data Science at UOW
- Composed of researchers actively using or interested in Data Science methods
- Themed meetings emphasising translation: Data and Decision Science Network (DDSN)
- Strategically collaborations through the DDSI give a competitive advantage in translation

Education: Training in data science and reproducibility of research.

- Internal and external training and education in data science
- Upskilling research students & staff (particularly ECRs) in data & decision science methods
- Workshops (GRS, Statistical Consulting Centre)

T shaped graduates: Reviewing service subjects to refocus on data science.

- Review of service subjects in statistics and quantitative methods to give data science focus
- Graduates literate in data science and reproducible research

External/Industry engagement: Capitalising on existing links

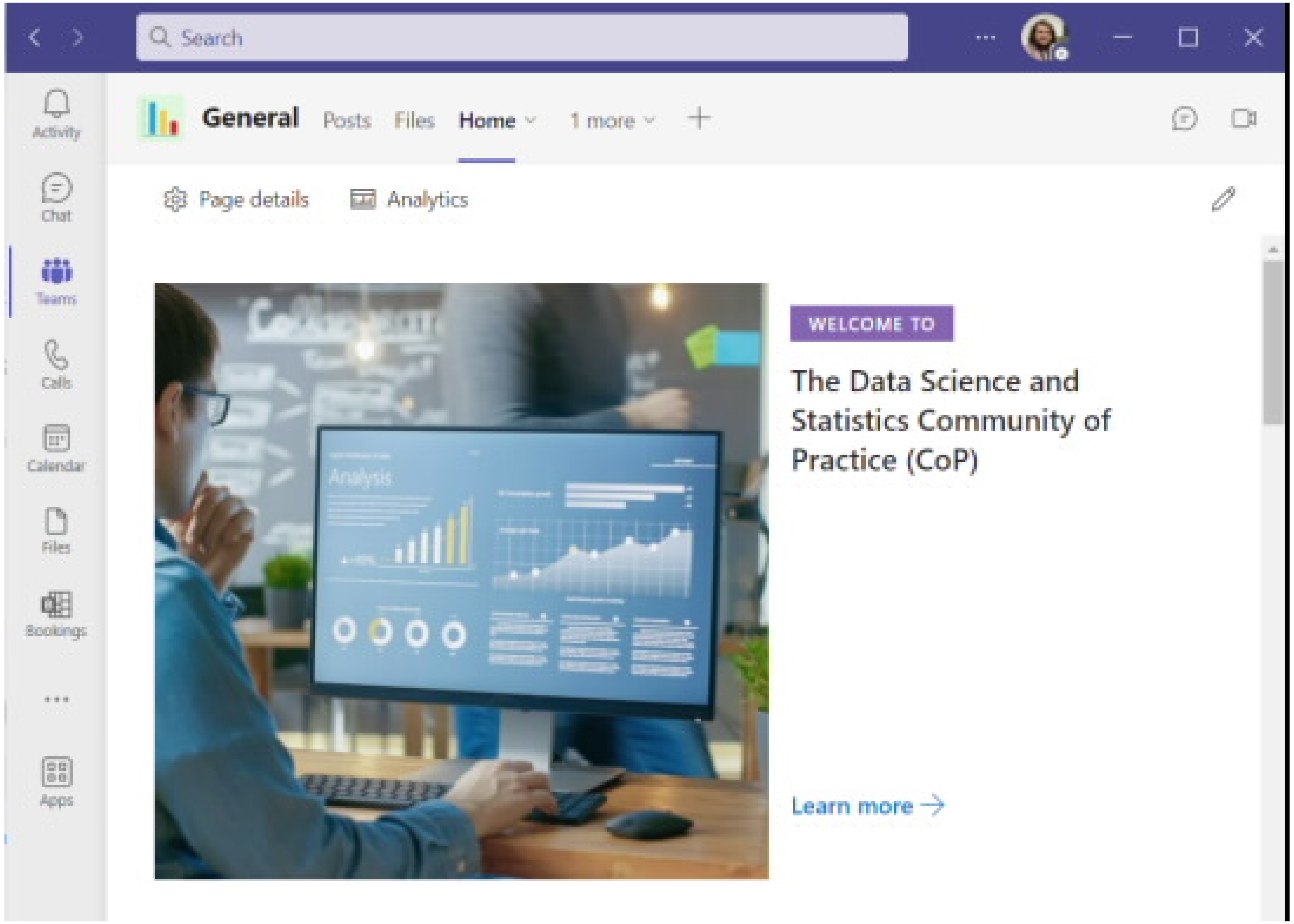
- Provide enhanced opportunities for external engagement

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG AUSTRALIA

The Data Science and Statistics Community of Practice

An online space to:
Foster Knowledge Sharing
Access Resources
Collaborate with Peers.
Get Data Science and Stats Support.
Learn about training opportunities.



WELCOME TO

The Data Science and Statistics Community of Practice (CoP)

Learn more →

Link in the Chat

UNIVERSITY OF WOLLONGONG AUSTRALIA

SEMINAR

Dr. Brad Wakefield

What is a Statistical Analysis Plan?

DEFINITION

A **Statistical Analysis Plan** (SAP) is a detailed, pre-defined blueprint outlining the methodologies and procedures for statistically analysing data in a research study.

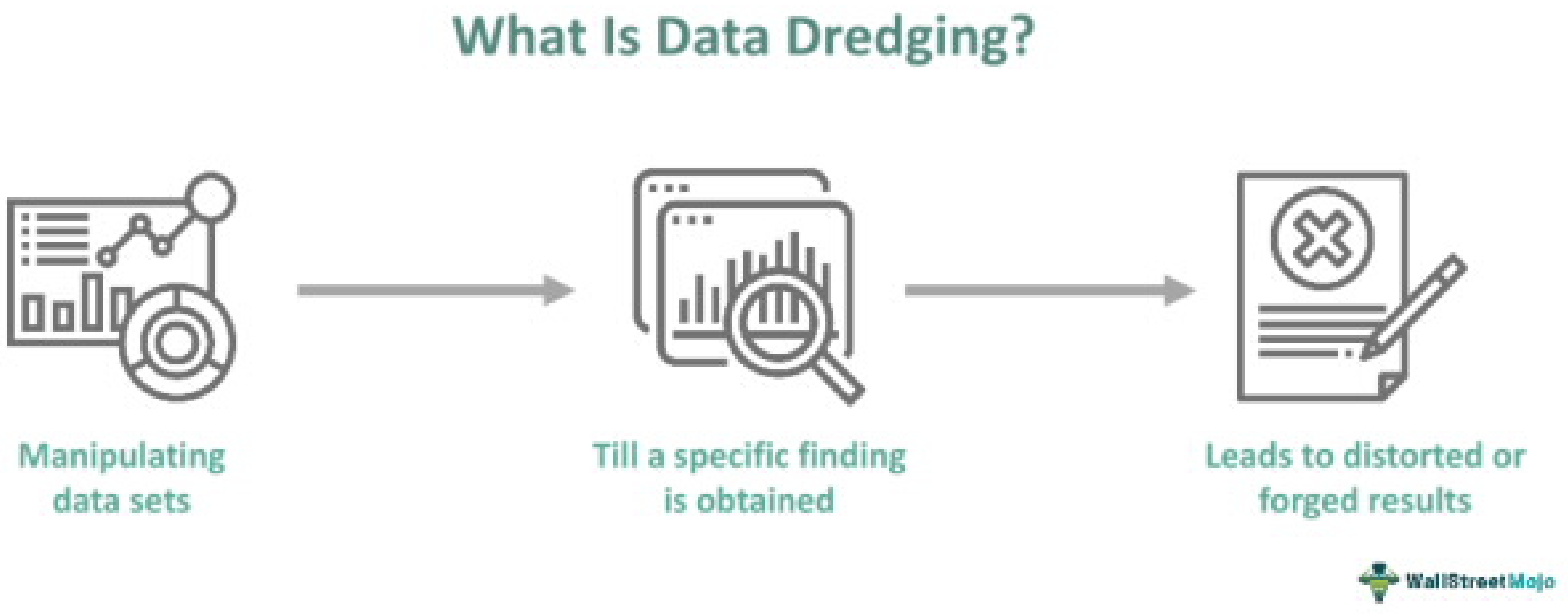
A well-considered SAP:

- Mitigates the risks of problems in the analysis
- Provides a transparent overview of the research
- Protects against **data dredging**.
- Promotes **reproducibility**.

Data Dredging

ALSO KNOWN AS DATA FISHING or P-HACKING

Data dredging refers to the practice of performing extensive and indiscriminate analysis on a dataset to find statistically significant patterns or relationships, without a prior hypothesis.


What Is Data Dredging?
Manipulating data sets
Till a specific finding is obtained
Leads to distorted or forged results
WallStreetMojo

The False Positive Bias

BUT WHAT IS SO WRONG WITH PERFORMING MULTIPLE TESTS UNTIL YOU FIND SIGNIFICANCE?

Statistical Analysis Plans

PROMOTE REPRODUCIBILITY

Statistical Analysis Plans ensure that only **pre-planned analyses** are performed regardless of their outcome.

Pre-registration of SAPs and protocols are often required for clinical trials but are still not common in non-clinical studies.

But as we a greater focus on **reproducibility** emerges, SAPs will likely become more and more common.

SAPs are considered best practice for analysis. However, the greatest benefit is usually to the research team.

DEFINITION

Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators. (Gundersen, 2021)

Reproducible research is distinct from **repeatability/replicability**, which refers to the general consistency or agreeance of measurements when the study or investigation is repeated.

If research is reproducible, then an independent researcher should be able to produce the same analysis, summaries, and visualisations obtained by the initial team.

Statistical Analysis Plans

PLANNING AHEAD REDUCES MISTAKES

- Seemingly small decisions in study design can have large impacts on the analysis.
- At the analysis phase, it is too late to change the design, or collect additional variables.
- SAPs formally outline how research objectives are measured, analysed, tested and reported from the onset.
- Potential problems can be identified early.
- Your study can also be easily and fairly reviewed a priori.

Although stated first, research questions often develop throughout the entire drafting process of the statistical analysis plan as more and more considerations are made, and the scope of the project adjusts.

.

Building a Statistical Analysis Plan

S1: Study Synopsis

Developing your research questions

Developing a Research Question

THERE ARE A LOT OF THINGS TO CONSIDER WHEN DEVELOPING A RESEARCH QUESTION

OBJECTIVES

ELIGIBILITY

SCALES

CONFOUNDERS

POPULATION

EXPOSURE

INTERVENTION

OUTCOMES

SAMPLING FRAME

INCLUSIONS/EXCLUSIONS

MODIFIERS

CONTROLS

RECRUITMENT

RELEVANCE

COST

SAMPLING FRAME

NOVELTY

BLINDING

RANDOMISATION

TIMEFRAME

FEASIBILITY

ETHICS

SAMPLE SIZE

What is the problem?

WHAT ARE THE REASONS FOR DOING THIS RESEARCH?

Questions to ask….	Corresponding Terminology
What is the purpose of the research?	**Objectives**
Why are you interested?	Motivation
Has this been previously studied?	Literature Review
What has not been addressed previously?	
What will your research contribute?	Novelty
What will the benefits of your study be?	Relevance
What are the ethical constraints?	Ethics
Do I have the resources needed to achieve this?	Feasibility
Are the objectives possible?	

Consider FINER (Feasible, Interesting, Novel, Ethical, Relevant) Research

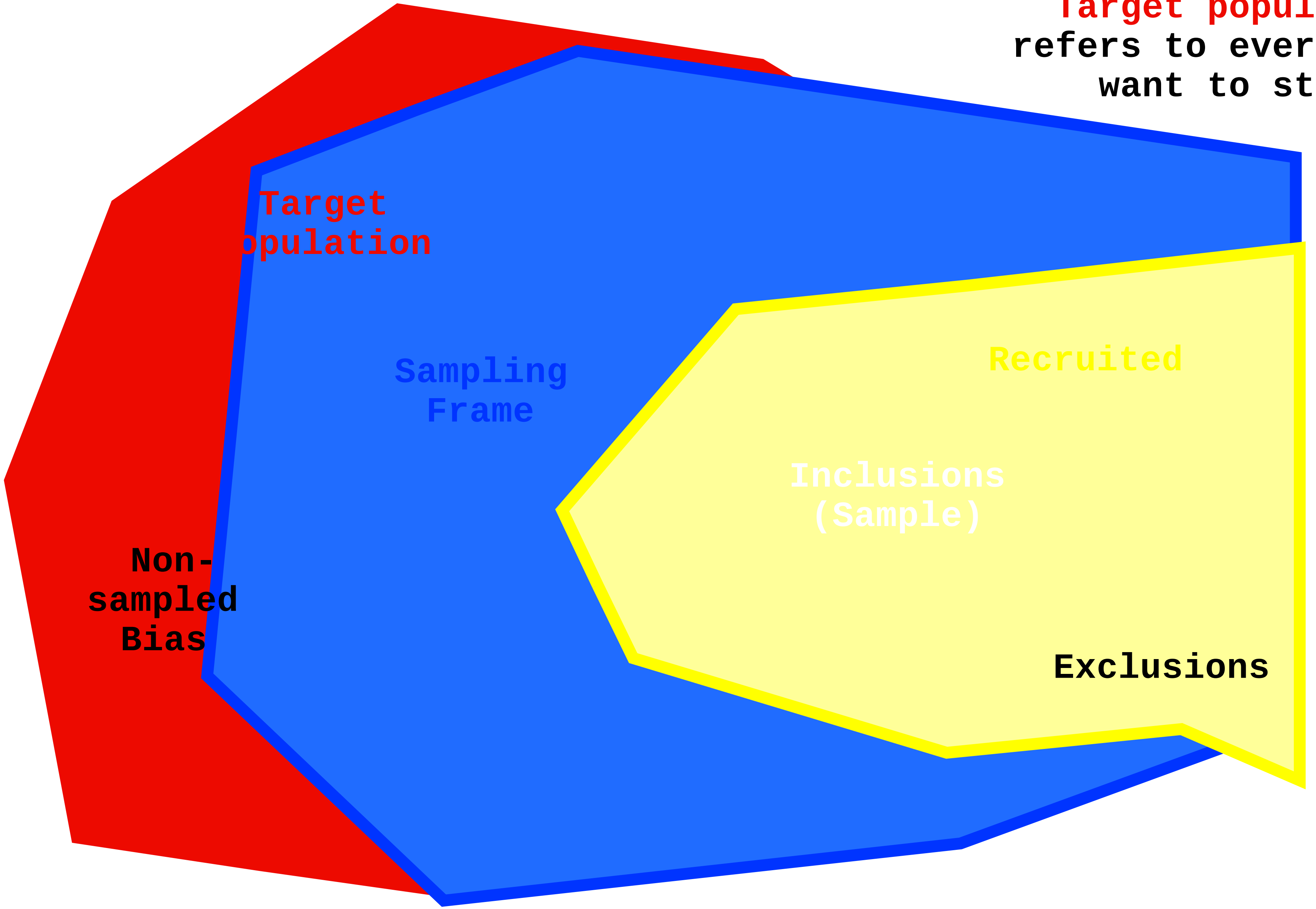
-Cummings SR, Browner WS, Hulley SB. *Designing Clinical Research.* 4th ed. Philadelphia: Lippincott Williams and Wilkins; 2013. Conceiving the research question and developing the study plan; pp. 14–22

Who (or what) are my subjects?

ENSURE YOU ONLY CONSIDER WHAT IS FEASIBLE

Questions to ask….	Corresponding Terminology
• **Who (or what) do I want to study?**	**Population**
• Who could I possibly study?	Sampling Frame
• Is representativity important?	Descriptive or Inferential
• How do I recruit / select my subject?	Sampling Method
• Will the selection process over/under represent any subjects?	Selection Bias

Population vs Sampling Frame



Target population refers to everyone we want to study

Sampling Frame refers to everyone who can be selected

Sample refers to everyone who will be studied

Selection Bias refers to when some subjects are more / less likely to be selected.

Who (or what) are my subjects?

ENSURE YOU ONLY CONSIDER WHAT IS FEASIBLE

Questions to ask….	Corresponding Terminology
• **Who (or what) do I want to study?**	**Population**
• Who could I possibly study?	Sampling Frame
• Is representativity important?	Descriptive or Inferential
• How do I recruit / select my subject?	Sampling Method
• Will the selection process over/under represent any subjects?	Selection Bias
• How do I assess the level of bias?	Demographics / Characteristics
• Are there any characteristics that may interfere with my study?	Confounders
• Can I restrict my sample to avoid these confounders?	Inclusions / Exclusions

GENERALISABILITY IS OFTEN CALLED EXTERNAL VALIDITY

What am I measuring?

WHAT DATA WILL I BE COLLECTING?

Variable - a quantity or characteristic that can take on different values or categories. Variables are used to represent and measure different attributes or features of a population or sample.

Quantitative (Numeric) - Variables which have a numeric meaning.

Qualitative (Categorical) - Variables which describe a categorisation of an attribute.

Dependent variables are variables we want to predict or explain

- **Outcome variables** - variables that relate to a particular outcome or measure the success or failure of a study.
- **Response variables** - variables that respond to a particular change in independent variable.
- **Endogenous variables** - variables that are determined by other variables in a model.

What am I measuring?

DATA THAT WE WANT TO DETECT THE EFFECT OF ….

Independent variables describe the variables that are being studied to control or assess their effect on the dependent variable.

An **exposure** variable is a variable of interest that participants may encounter or possess, which researchers examine for its potential influence on the dependent variables(s).

Used for both observational and experimental studies

An **intervention** variable refers to a specific condition or component that researchers manipulate in an experimental study to determine its effect on the dependent variable(s).

The relationship between the dependent variable(s) and the intervention/exposure variable(s) is the key comparison you would like to perform in the study.

A **control** is a group consists of subjects not exposed to the intervention and is needed to evaluate the magnitude of the intervention.

What am I measuring?

OTHER TYPES OF VARIABLES DO WE COLLECT

Confounding variables account for changes in the dependent variable not associated with the intervention or exposure.

Effect Modifiers are variables that result in differences in how a dependent variable may be associated with an intervention or exposure when present.

Control variables are variables that are held constant or monitored to minimise their effect on the dependent variable.

Grouping variables are variables that denote association structures within the data.

Characteristic / demographic variables are variables that we measure to describe the cohort and assess representativity.

How am I measuring the variables?

FOR EACH VARIABLE YOU COLLECT YOU SHOULD ALSO CONSIDER

Do your variables adequately match your **objectives**?

When will you perform the measurement? Is there a **time effect**?

Will you have **repeated measures**?

What **units** are you measuring, are they comparable?

Are you dealing with **independent** or **paired** data?

If paired, how will you **link** observations?

Can you assume **independence** between observations?

Will I have **missing data**? How will that change my sample?

Variable Data Types

What type of data will my variables be?

Variable Data Types

What type of data will my variables be?

Qualitative Variables

Variables which describe a categorisation of an attribute.

- Nominal - Categorical with no ordering structure.
- Ordinal - Categorical with an ordering structure.

Confusingly, quantitative and qualitative data are both used in **quantitative analysis**.

Sometimes we must mine qualitative variables from **unstructured text data**.

EXAMPLE: QUALITATIVE VARIABLES FROM TEXT

Name	Age	Gender	Symptoms	Pain in Arm?	Has Fever?
John	52	Male	Patient experienced shortness of breath, pain in arm	Yes	No
Joe	47	Male	Pain was experienced by the patient in arm	Yes	No
Jane	32	Female	Patient exhibted a fever on arrival	No	Yes
Jack	19	Male	Coughing, sore throat, runny nose, fever	No	Yes

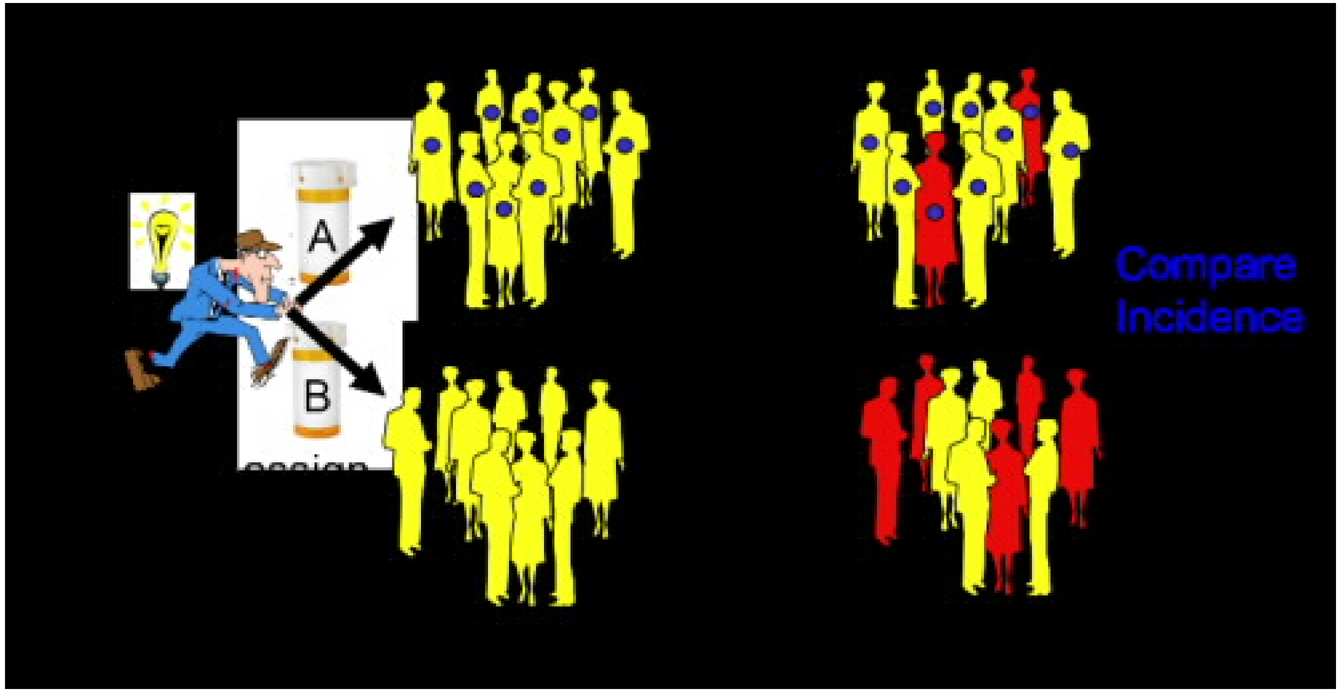
Randomisation

RANDOMISATION IS ONLY POSSIBLE IN EXPERIMENTAL STUDIES

Randomisation is the process by which subjects are randomly assigned to a particular treatment in experimental studies.

In large studies, randomisation ensures that a representative sample is used for all treatments being studied.

By randomising treatments, all known and unknown confounders should be equally likely to appear in all treatment groups meaning any difference between the groups is likely the result of the treatment.



Blinding

BLINDING IS ONLY POSSIBLE IN EXPERIMENTAL STUDIES

Blinding refers to the practice of concealing group assignments (e.g., treatment vs. control) from one or more individuals involved in an experiment to prevent bias.

Blinding ensures that participants and/or researchers can participate, administer, and/or analyse a study without undue influence.

Single-blind: Participants are blinded.

Double blind: Participants and administering researchers are blinded.

Triple blind: Participants, administering researchers, and those analysing the data are all blinded.

Sample Size

INCREASING SAMPLE SIZE REDUCES SAMPLING ERROR BUT NOT BIAS

How many observations can I feasibly get?

- What is the size of your sampling frame?
- What will your response rate likely be?

How many observations are needed to achieve the power and significance required to test my objectives?

- What is a clinically significant effect size?

Will I need to perform subgroup analysis?

Will I achieve the required heterogeneity in the sample?

Statistical Power

INCREASING SAMPLE SIZE REDUCES SAMPLING ERROR BUT NOT BIAS

Statistical Power is the probability that a test will correctly reject a false null hypothesis (i.e., detect an effect when there is one).

It is a measure of a study's ability to detect an effect that actually exists.

Factors affecting power:

Sample Size: Larger samples increase power.

Effect Size: Larger effects are easier to detect.

Significance Level (α): Setting a higher α increases power but also the risk of Type I error.

Variability: Less noise in the data increases power.

Usually, the only lever we can use to control power is **sample size.**

Problems with Power

UNDER POWERED STUDIES …

Risk not detecting meaningful differences or relationships.

Potentially waste time and resources on inconclusive studies.

Reduce the contribution of the research.

STATISTICAL POWER SHOULD BE CONSIDERED PRIOR TO CONDUCTING ANY STUDY

Power is calculated for a specific hypothesis test.

Power calculations should be based on domain expertise, past results, and reasonable approximations.

Usually, a power of 0.8 is considered the minimum sufficient power for a study.

GPower is a useful free tool to conduct Power analyses.

S2: Data Management

Preparing for your data

Data Governance

WHAT IS DATA GOVERNANCE AND WHY DO I CARE?

Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an organisation

AT UOW it is **required** that all data that contains **personal information** is covered by a **Research Data Management Plan.**

There are policies relating to the access, de-identification, storage, security, sharing, and disposal of data.

Research Data Management Plan

WHAT DO I NEED TO KNOW?

Data Sources - the data that you’ll need.

> **Note:** Data can be in many different forms besides tables.

Data Storage – check out approved storage options

Data Usage – how will the data be used?

Data Access – who will need access to my data?

Data Ownership – who will own the data?

Data Sharing – who will I share the data with?

Data Disposal – how and when will my data be destroyed?

Data Privacy

A FEW IMPORTANT POINTS ABOUT DATA PRIVACY

Data (Informational) Privacy encompasses an individual’s freedom from excessive intrusion and the ability to choose the extent and circumstances under which one’s personal information will be shared with or withheld from others.
-Statement on Data Access and Personal Privacy from the American Statistical Association. Dec 6, 2008.
*Note: Privacy is about **control**, not about secrecy.*

We have a legal responsibility to protect **personal information** and make sure it is **not reasonably identifiable**.

When assessing risks to data privacy you should consider the **data situation.**

Data Privacy

A FEW IMPORTANT POINTS ABOUT DATA PRIVACY

De-identified should mean not identifiable, but it doesn’t

We use the term **de-identified** when we remove **direct identifiers** from a data set.
A **direct identifier** is a variable that uniquely identifies an individual e.g. names, addresses, date of births, etc.
An **indirect identifier** is a variable that in combination with other variables, can be used to uniquely identify and individual e.g. sex, age, occupation, postcode, etc.
**A student at MIT showed that 97% of names and addresses on the 1997 voting list for Cambridge Massachusetts were unique using only zip code and date of birth.*
When considering indirect identifiers, you should consider what **existing data** is already available.
Data with indirect identifiers are potentially re-identifiable in some contexts.

Data Privacy

A FEW IMPORTANT POINTS ABOUT DATA PRIVACY

Risk of **disclosure** changes based on **who** has access to the data.

Removing unnecessary information reduces the risk of re-identification.

The more data variables, the more chance of re-identification.

Consider: It may not be rare to be 28, male, a statistician, or to work at UOW, but how many people are all four?

Not all people who attempt to re-identify have malicious intent.

Concerned loved ones are often the most motivated to re-identify.

Risks to data privacy should be weighed against the benefits of the data usage.

There are different ways disclosure can occur.

Participants should be told how you will protect their data. Do not provide assurances you can’t keep.

ALWAYS CHECK WITH AN EXPERT BEFORE PUBLICLY SHARING ANY DATA.

What influences disclosure risk?

Factor	**Effect on disclosure risk**
Data age	Older data is generally less risky
Sample data (e.g. a survey)	Decreases risk
Population data (e.g. a census)	Increases risk
Administrative data	Increases risk
Longitudinal data	Increases risk
Hierarchical data	Increases risk
Sensitive data	Increases risk (sensitive data may be a more attractive target)
Data quality	Poor quality data may offer some protection
Microdata	Main risk: re-identification
Aggregate data	Main risks: attribute disclosure and disclosure from differencing
Key variables	The variables of most interest to users are usually the most disclosive

Data Sources

WHAT DATA IS TO BE COLLECTED AND ANALYSED?

Consider what data sources are needed.

What data are you collecting / producing?

What data will you get from somewhere else?

- What is the process to obtain access?
- Are there any **restrictions** on data usage?

What **form** will the data come in?

Will I need to **link** data?

- How will linkage occur?

Data Sources

WHAT DATA IS TO BE COLLECTED AND ANALYSED?

Always maintain a raw copy of all data sources.

Will your data require manipulation?

e.g. converting from wide to long format

Are you dealing with Big Data?

- What hardware/computational requirements may be needed?

If using external data, cite your sources.

- Ensure you always obey the terms of use of any external data set.

Selecting a Statistical Software

WHICH SOFTWARE IS BEST?

CHECK OUT OUR TALK – Which stats package is best?

Package	Good for
Jamovi	Teaching, infrequent use of stats (easy to pick up again if you have a break), basic analysis some advanced methods, easy to learn, good default outputs
Python	Machine learning, AI, in demand skill, regular users, good for research collaboration and integration to web platforms, regular user
R & Rstudio	Data manipulation, visualisation, advanced analysis, in demand skill, reproducible research, advanced missing data options, regular users
SAS	Good overall package for most standard and many advanced methods, regular user, big data, good for pharma and govt
SPSS	Good overall package for most standard and many advanced methods, easy to learn, infrequent use
STATA	Good overall package, has many useful advanced procedures, Used regularly in some professions, particularly good for survey analysis, meta analysis, SEM

Data Cleaning

AVOIDANCE IS THE BEST TYPE OF DATA CLEANING

Strategies to avoid data cleaning:

1. **Avoid free text entry**: Provide pre-chosen options for people to select.
2. **Make data entry easy:** Provide clear instructions, only collect necessary information.
3. **Use Data Validation**: Apply rules that automatically check data as it is entered and notify users if their response is not correct to prompt immediate correction.
4. **Standardise Data Entry**: Use consistent variables, labels, data types in all data sources.
5. **Automate Entry**: Where possible, use technology to automatically input data and autofill common values in data fields.
6. **Use mandatory fields to reduce missing data**: Mandatory fields force a response.

   Warning: People may abandon a survey if they don't know how to respond and can't skip the question

   If you want people to be able to skip, provide a *Prefer not to say*

Data Cleaning

SOME CLEANING MAY BE UNAVOIDABLE

Plan your protocol for cleaning **BEFORE** collection.

- Considering what might go wrong will help you think of strategies to limit it before you collect data.
- Having a protocol designed before you see any results ensures you aren’t **data dredging**.

Use **reproducible software** to perform data cleaning.

Take note of exactly what steps you did to clean the data.

Justify your data cleaning decisions. Consider how this may impact the quality of your data.

If you identify issues with your data, report them as a limitation.

Data Coding

THE PROCESS OF CONVERTING DATA INTO A STANDARDISED ANALYSABLE FORMAT

Match your data columns to the variables of interest

Consider what data will need to be computed

- What information is needed to compute these variables?
- Avoid computing manually.
- Use a reproducible software.
- Compute for all cases at the same time!

Use numeric codes with labels for categorical data.

Consider if any variables need to be summarised – will variables need to be merged into broader categories?

Provide clear **definitions** for each variable and its coding.

Create a **data dictionary** and a README file.

Data Checking

ALWAYS ALWAYS ALWAYS, PLOT YOUR DATA

Double Check

Each variable has values that **make sense** (no negative distances etc)

The number of unique responses matches the number of categories.

That hierarchical relationships are preserved.

That proper heterogeneity is present across all expected values. *e.g. if you expect to have people of all ages but everyone is over 80.*

That the summary statistics make sense and are in the correct units.

That no subgroups are missing that should be present.

That all cases present in the raw data can be accounted for.

S3: Planning Analyses

Statistical analyses methods

Data Description

HOW WILL YOU DESCRIBE YOUR DATA?

To contextualise data, we often provide a description of the data.

- How many observations were achieved?
- What demographics / characteristics were present?
- What was the demographic / characteristic composition?

Descriptive statistics can be compared with benchmarks to assess representativeness in **non-probability samples.**

- *Warning: Consistency with benchmarks is an indicator, not a guarantee of representativeness.*

For each subgroup we usually provide the **number of observations**, a measure of **centre**, and a measure of **spread**.

Data Description

HOW WILL YOU DESCRIBE YOUR DATA?

Measures of **centre** provide an indicator of a **typical** observation.

Measures of **spread** provide an indicator of the level of differences in data.

- For symmetric continuous data (without outliers), we usually use mean and standard deviations.
- For skewed continuous data, medians are often preferred.
- When outliers are present, medians and IQRs are often preferred.
- For categorical data, the mode, proportions, or frequencies are used.

Consider also what previous studies have reported to allow for comparison.

Exploratory Data Analysis

ALWAYS PLOT YOUR DATA TO IDENTIFY ANY PATTERNS AND MISTAKES!

Exploratory data analysis is a critical early step in data analysis and aims to understand the structure, patterns, and anomalies in data.

EDA can uncover underlying structures and associations, detect outliers and problematic points, test distribution assumptions, and provide context for later results.

Techniques include:

- **Descriptive Statistics**: Mean, Median, Mode, Range, Variance, Standard Deviation
- **Visualisation**: Histograms, Box Plots, Scatter Plots, Heatmaps, Pair Plots
- **Multivariate Analysis**: Correlation matrices, PCA
- **Non-graphical EDA**: Clustering, Dimensionality Reduction

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 1: Outcome Diagnostics

- What checks will be employed to ensure the outcome measure is a reliable measure?

  For Example:

  - For questionnaire scales, a reliability analysis checks for internal consistency.

- Consider what alterations / simplifications may need to be made to ensure the outcome variable is suitable for analysis.

  For Example:

  - What if too few observations fall in a particular category?

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 2: Statistical Model

Consider what **statistical model** should be used to predict/explain the outcome variable.

What variables are in this model?

What exposure/intervention variables need to be included?

What modifiers need to be included?

Are there any correlation / hierarchical structures that need to be modelled?

Are there any subgroups that need to be modelled differently?

What is a Statistical Model?

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 3: Model Inference

- What analysis can be used to estimate this statistical model?
- What specific **hypotheses** will be tested?
- What inferences can be drawn from these tests?
- How do these inferences relate to my study objectives?
- What are the assumptions of the hypothesis tests?
- How will these assumptions be tested?

Y **Response**	**X** **Explanatory**	**Specific question(s)**	**Displays**	**Statistical method**
Categorical	Categorical	How do proportions in response depend on **the levels of the explanatory variable?**	Tables	Chi-squared statistic
Categorical	Continuous	How does the proportion in response depend on the **values of the explanatory variable**?	Tables (X groups)	*Logistic regression* *Correlation (for a binary response only)*
Continuous	Categorical	How does mean level in response change with **the levels of the explanatory variable? If so how does it vary?**	Box plots Mean plots CI plots	t test (2 groups) ANOVA (3 or more groups)
Continuous	Continuous	How does mean level of response change with **values of the explanatory variable**.	Scatter plots	Correlation Regression
Dependent (or paired) samples				
Categorical	Categorical	Is there agreement between the matched levels? Is lack of agreement biased? Is there a difference in proportions?	Tables	Kappa (2x2 Agreement) McNemar’s test (2x2 - bias in agreement or difference in proportions
Continuous	Categorical	How does mean level in response change with the levels of the explanatory variable WITHIN e.g. subject	Box plots Mean plots Within CI plots	Paired t test Repeated measures ANOVA

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 4: Adjusted Analysis

* What happens if the assumptions of the model are violated?

* What alternative testing can be undertaken?

    (e.g. Non-parametric testing)

STEP 5: Treatment of Missing Data

* How will missing data be treated in this analysis?

* Will this affect my inference?

* What checks can I perform to ensure the planned treatment of missing data is appropriate?

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 6: Other Sensitivity Analyses

- Describe any pre-planned additional sensitivity analyses of the primary outcome.
- These could include:
  - additional covariate adjustments,
  - causal or mediation analyses,
  - additional methods for handling missing data (e.g. best/worst case imputations, tipping point analyses),
  - using a different censoring method,
  - using a GEE model instead of a mixed effects model,
  - or checking the impact of outliers.

Please note that this is not an exhaustive list.

Analysing Outcomes

STEPS TO ANALYSING EACH OUTCOME VARIABLE

STEP 7: Outputs

- What outputs will I obtain from my analysis?
  - Statistics, p-values, confidence intervals, etc
- How will they be presented in my report?
- What **data visualisation** will appear?
  - How will they be produced?
  - Have I selected the correct visualisation?

STEP 8: Repeat this process for each outcome of interest

Best Practices

FOR STATISTICAL ANALYSIS

1. Encourage the use of open source tools and software

By using open source tools and software, researchers can ensure that their work is easily reproducible by others.

Open source software is:

- available to everyone for collaboration,
- able to be backed up and archived,
- custamisable and flexible,
- transparent and open to external review,
- allows for the quick adoption and trialing of new developments,
- cost effective.

Programming languages such as R and python are great open source software that are routinely used in academic, government, and industry.

2. Document everything

Researchers should document every aspect of their work, **including data sources, code, and analysis methods**. This documentation should be clear, concise, and easily accessible to others.

Using programming scripts rather than spreadsheet-based software (such as excel) allow for every step in the data cleaning process to be captured. Commenting scripts allows explanation of code in a more readable format.

All code should be accompanied by README files and vignettes that capture the analysis methods and explain how to use the code is essential.

README files are essential in detailing the process of any analysis.

Data Descriptions detail the sources and history of data sets.

Data Dictionaries ensure others can understand the variables in your data set.

Licences provide clear usage instructions for further research

3. Use defensive coding

Defensive coding is a programming technique that involves anticipating and preventing potential errors or bugs in code by implementing safeguards and error-handling mechanisms. The goal of defensive coding is to minimise the impact of errors or bugs on the overall functionality and stability of the software.

Important functions for handling errors in R are:

1. **stop()** - This function allows you to stop the execution of your code if a certain condition is met. For example, you can use **stop("Error: Invalid input")** to halt the code if the input is not valid.
2. **tryCatch()** - This function allows you to catch and handle errors that may occur in your code. You can specify different actions to take depending on the type of error that occurs.

Avoid hard coding paths and direct indexes.

Example: Non-defensive Code

Example of code that is not defensive.

```
data <- readRDS(

'C:/Users/bradleyw/Dropbox/GitHub/DSAA311/lectures/reproducib
le_research/reproducible_research_files/data/NSW_income.rds')

max_income <- data$income[8360]
mean_income <- mean(unlist(data[,2]))

avg_occ <- data %>% group_by(occupation) %>%
  summarise(avg=mean(income))
```

Paths are hard coded. No comments.

Hard coded indexes. Package libraries not loaded.

File Structure is Key

Setting up a logical, **consistent**, file structure is essential for the navigability and portability of your analysis.

```
C:/Documents/□   └── repos/□       └── proj/□           ├── data/□          ├─
─ docs/□          ├── figs/□             ├── funs/□           ├── out/□
├── cleaning.R□           └── analysis.R
```

Make sure all dependencies are in one place.

Use relative paths. e.g. /data/file.csv rather than C:/Users/bradleyw/Documents/file.csv

4. Use version control

Version control systems such as **git** can be used to track changes to code and data over time, making it easier to reproduce research results.

Using git has some very important advantages:

1. **Version control**: allows you to keep track of changes to your code over time, making it easy to revert to previous versions if necessary. This is especially useful when working on large, complex projects with many contributors.
2. **Collaboration**: allows multiple people to work on the same project simultaneously, without the risk of overwriting each other’s work. It also makes it easy to share code and collaborate with others, even if they are located in different parts of the world.
3. **Branching and merging**: allows you to create branches in your code, which can be used for testing new features or making experimental changes without affecting the main codebase. Branches can then be merged back into the main codebase when they are ready.

What is Git?

Git is a free open-source command line program used for version control.

You can install git by downloading it from https://git-scm.com/downloads

Git can be used in conjunction with a git GUI meaning your don’t have to learn git syntax.

Note, there are a range of other git GUIs such as GitKraken, SourceTree, and GitHub Desktop.

The version control happens locally and so you can work through a project offline before sharing online.

You can connect RStudio with git by going to Tools > Global Options > Git/SVN and selecting the git executable.

5. Share data, code, and meta-data

Researchers should make their data and code available to others in a well-organised and accessible format (compliant with data security considerations).

The ARDC outline principles to improve data sharing and maximise the impact of your data

FAIR Principles

- **Findable** - Data has sufficiently rich metadata and a unique and persistent identifier to be easily discovered by others.
- **Accessible** - Data is retrievable by humans and machines through a standardised communication protocol, with authentication and authorisation Interoperable
- **Interoperable -** Data and metadata use a ‘formal, accessible, shared, and broadly applicable language for knowledge representation’ where necessary.
- **Reusable** - The associated metadata provides rich and accurate information, and the data comes with a clear usage licence and detailed provenance information.

GitHub

GitHub is a free web-based platform to store and share your git repository.

GitHub allows your git repository to be either public or private, and allows you to work in tandem with multiple collaborators on the same repository.

To update your GitHub repo, you must **push** any comitted local changes to your git repository.

To download updates from a GitHub repo, you must **pull** and **merge** the changes with your local files.

Features of GitHub:

- Create, store, and manage code repositories.
- Track and manage issues and bugs in your code.
- Submit and review changes to code.
- Review and comment on code changes.
- Collaborate with others on code projects.
- Automate code testing and deployment.

6. Practice good dependency management

Dependency management is the process of identifying, resolving, and managing the software dependencies of a software project.

Record what software, and what version was used to perform the analyses.

Analyses often require external packages or versions of software that can constantly be updating and changing, and there is no guarantee that the package you use now will still be available in 10 or 20 years time.

7. Make use of virtualisation and containerisation

Virtualisation - the creation of a virtual version of an operating system, server, or network within another system. It allows multiple operating systems to run on a single physical machine, each with its own set of resources such as CPU, RAM, and storage.

Containerisation - a lightweight form of virtualisation that allows multiple isolated user-space instances or “containers” to run on a single operating system kernel. Containers share the underlying operating system and hardware resources, but they are isolated from one another, which means that they can have different software versions and configurations.

Docker is an open-source tool that uses containerisation to create lightweight, self-contained environments for running applications. It packages all the necessary components of a project, such as code, libraries, and dependencies, into a single container, which can then be run on any system that supports Docker.

8. Use literate programming

Literate programming is a programming methodology introduced by computer scientist Donald Knuth in the 1980s. The approach involves writing programs in a way that emphasises their readability and comprehensibility to humans, rather than just to machines.

In literate programming, a program is organised into a series of “chunks” of code and text, each of which can be independently executed or viewed as part of a larger narrative. These chunks are arranged in a way that tells a story about how the program works and why it was designed that way.

The goal of literate programming is to make it easier for programmers to write, read, and understand code, as well as to encourage good documentation practices. By presenting code in a more human-readable format, it can be easier to spot errors, identify patterns, and reason about complex algorithms.

Literate Programming Software

Examples of commonly used software for literate programming:

- **Markdown** - A lightweight markup language that supports code blocks and inline code for literate programming.
- **Jupyter Notebook** - An interactive notebook environment that allows literate programming with code blocks, markdown, and rich media. Used with Python, R, and Julia.
- **Google Colab** - A cloud-based platform that allows users to create and share Jupyter notebooks for literate programming with Python, R, and other programming languages.
- **RMarkdown** - A variant of Markdown that supports literate programming with R code blocks and inline documentation.
- **Quarto** - An open-source document processor that supports literate programming with code blocks, markdown, and a variety of programming languages including
  R, Python, and Julia.

9. Seek advice

THE STATISTICAL CONSULTING CENTRE

Marijka Batterham
Director

Brad Wakefield
Statistical Consultant

Aim

The service aims to improve the statistical content of research carried out by members of the University. Researchers from all disciplines may use the Centre. Priority is currently given to staff members and postgraduate students undertaking research for Doctor of Philosophy or Masters' degrees.

How we can help

Currently the Statistical Consulting Centre provides each academic or post-graduate student with a free initial consultation. Up to ten hours per calendar year of consulting time is provided without charge if research funding is not available. When researchers require more consulting time, or receive external funding, a service charge may be necessary.

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

To book an appointment, simply go to our website and select MAKE AN APPOINTMENT