Research Essentials

Presented by Dr Alexandra Green Statistical Consultant, Statistical Consulting Unit Sydney Informatics Hub Core Research Facilities

sydney.edu.au/sydney-informatics-hub







Acknowledging SIH

- All University of Sydney resources are available to 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."

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.
- To access these statistical workflows and more, visit our <u>Workshops and Workflows</u> page.



Software workflows



- 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 <u>Hacky Hour</u> where SIH staff can help you.
- Our software workflows contain:
 - R code and comments.
 - SPSS syntax as well as screenshots of the point and click procedures and written methods.
 - Screenshots of the point and click procedures and written methods for other bespoke software.

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.

S

- 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

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?

Setting up your data for most analyses:

- Step 3: Collect and store data
- Step 4: Cleaning data

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:

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

Collins English Dictionary

7. Predictive modelling – Inferential predictive statistics versus machine learning predictive analytics

Inferential predictive statistics



Predicting temperature rise in climate change.

 Machine learning/predictive analytics



Very accurately verify fingerprint to unlock a mobile phone.



Ecosystem of SIH statistical training

Workflow step	Other training
1. Hypothesis generation	Research Essentials
2. Study design	Experimental Design Power and Sample Size Design and Analysis of Surveys 1 + Design and Analysis of Surveys 2: Advanced Topics Statistical 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 to 3 + Statistical Model Building Introduction to Survival Analysis Meta-Analysis: An Introduction Design and Analysis of Surveys 1 + Design and Analysis of Surveys 2: Advanced Topics Multivariate Statistical Analysis 1: Dimension Reduction
7. Predictive modelling	[Predictive analytics: <i>Introduction to machine learning in R/Python (SIH Data Science)</i>]: <u>From 26th August</u>
8. Publication	



SIH training

64

Different offerings, in person, online, and hybrid content in a variety of formats from webinars to interactive workshops.



Attendees at all career levels, from undergraduate students to senior professors, and representation from every Faculty and School.



Partnerships with national organisations like Australian BioCommons: biocommons.org.au/trainingcooperative.



Find out more on our training calendar: <u>sydney.edu.au/informatics-</u> <u>hub/training</u>. Or stay up to date

with our newsletter.





	Statistics	Data Science	Research Computing	Bioinformatics	Events
-	Fundamentals	Machine Learning	High Performance Computing	'omics Techniques	Hacky Hour
	Modelling	Visualisation	Cloud Computing	Reproducible Pipelines	Summer Schools
	Specialist	ecialist Language Containers Processing		Data Analytics	Coding Challenges
-		Geospatial Analysis	Workflows	National Compute Infrastructure	



Ecosystem of other USyd training

Workflow step	Other training
1. Hypothesis generation	Library research support: Literature and systematic review
2. Study design	
3. Collect/store data	RedCap- <u>Various trainings for survey data from</u> introduction to advanced Research Data Management modules and techniques
4. Data cleaning	
5. Exploratory data analysis	
6. Inferential analysis	
7. Predictive modelling	
8. Publication	Library research support: Publishing

Research data management

Research Data Consulting

Research Integrity & Ethics Administration

digital.research@sydney.edu.au

Book a consultation





platforms: Features and comparisons

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.





Hacky Hour For researchers who code or analyse data



- A monthly meetup where anyone from the University students, staff and university affiliates – can collaborate and get support e.g., swap notes, get help, or learn new techniques in programming and data science.
- Experts & mentors from SIH and across the University will be available to advise and answer questions on coding, data analytics or digital tools.
- Come join us on zoom the 3rd Wednesday of every month, 2 - 3pm!
- sydney.edu.au/informatics-hub/hacky-hour





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



If statistical programs were cars...

The first question: Which car will you take?













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**) – like an automatic car Interactive, point and click and Menu driven Easier to get started

A Good GUI

Command line interface (CLI) – like a manual car Writing code Easier to handle complex and/or large data sets



Software choice as the first question: GUI versus CLI

- Which software are you more familiar or comfortable with?
- How do you record your analysis for reproducible research?
- By documenting, you should always be able to re-run your analysis from start to finish (and get the same result)!
- If using interactive processing, you should keep track of the commands you run.

Graphical User Interface (GUI) SPSS



Command line interface (CLI) R code

• • •		R	~ະ#1
< ₨ lm(Sepal.Length	~ Species, data =	iris)	
Call: Lm(formula = Sepal	.Length ~ Species,	data = iris)	
Coefficients:			
(Intercept)	Speciesversicolor	Speciesvirginica	
5.006	0.930	1.582	
~			

3. Collect/store your data

A. Research data managementB. 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?
 - What is a Research Data Management Plan (RDMP)?
 - 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
 - <u>Best practices for data management and sharing in experimental</u> biomedical research



A. Research data management

Guide to storing and managing your projects research data									
University supported and licenced platforms							Unsuitable as primary storage for research data	Prohibited for protected research data	
Platform/Tool	eNotebook	REDCap	Research Data Store (RDS)	OneDrive (Enterprise)	Teams (Enterprise)	Highly Protected SharePoint (Enterprise)	Australian Imaging Service (AIS)	Local storage, USB Drive	Other cloud tools (e.g. Google Drive, Dropbox)
function	electronic notebook	survey and data capture, including Clinical trials	networked data storage, large files, HPC access	cloud storage	chat, collaboration, cloud storage	collaboration, cloud storage	imaging repository and analytics	removable media, local storage	cloud storage
suitable for data classification	•••	•••	•••	0	0	•••	•••	•	•
stored in Australia	~	~	~	~	~	~	~	various	×
external collaborator access	~	~	~	~	~	~	~	×	×
context and commentary supported	~	×	×	~	×	×	n/a	×	×
syncing with local copy	n/a	n/a	n/a	~	~	~	n/a	×	×
available storage	unlimited	unlimited	unlimited (default 2TB)	5TB	2TB+	25TB max (default 2TB)	unlimited	×	×
backup and disaster recovery	~	~	~	~	~	~	~	×	×
audit trail/version control	~	~	~	~	~	~	~	×	×
versioning retained	~	manual	up to 60 days	7 years	7 years	7 years	×	×	×





Highly Protected data may require addi

highly protected data needs additional file encryption

protected

public

highly protected

0

Highly Protected data may require additional encryption depending on some platforms. Protected data may benefit from encryption.

> For more information about research data classifications, go to https://sydney.edu.au/research-data-classifications

> > For research data management enquiries, please contact digital.research@sydney.edu.au

The University of Sydney

How do I manage my research data?

Version control – keeping track of your files

- Use a separate directory for each discrete analysis.
- When processing data and intermediate files save with a new name.
 - Use a good <u>file naming convention</u> to keep track of files.
- 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. <u>Git</u> (free) or <u>GitHub</u>.
- 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).

Version control – keeping track of your files

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





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

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



B. Data formats – transformations

Patient ID	Time 1	Time 2	Time 3
1	50	55	60
2	47	49	50

Wide/unstacked format



Patient 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

B. Organising a dataset for statistical analysis – data coding

- Specify type of variable: ensure your analysis software knows whether a variable is numeric (continuous or discrete), or categorical/factor/string/character (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 (e.g. 999, 777, 99, 77).

4. Data cleaning



Data cleaning: data wrangling and data dictionary

			,	
	А	С	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

A simple example of a 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)

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

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 an Rmarkdown/Quarto notebook if using R.
- Also covered in Good enough practices in scientific computing.
Example R Markdown and Quarto files

Files Plots Packages Help Viewer Jobs

 $-\Box$ Se Publish •

Hello R Markdown

Awesome Me

2018-02-14

0 1 5

This is a paragraph in an R Markdown document.

Below is a code chunk:

fit = $lm(dist \sim speed, data = cars)$ b = coef(fit)plot(cars) abline(fit)



R for Data Science: Chapter 27 R Markdown

> R for Data Science: Chapter 28 Quarto

markdown.gmd >> 💷 📄 📄 🖸 Render on Save 🛛 🖑 🔍 📫 Render 🛛 - 🛞 -+Run - 🔗 *a . Source Visual 🛛 B 👖 🐢 Normal - 📋 🏣 🛷 🔛 Format - Insert - Table -

Text formatting

italic bold underline strikeout SMALL CAPS code superscript² and subscript²

Headings

1st Level Header

2nd Level Header

3rd Level Header

Lists

Bulleted list item 1

Item 2

Tables 0

Item 2a

Item 2b

1. Numbered list item 1

2. Item 2. The numbers are incremented automatically in the output.

Links and images	Image Attributes Advanced
http://example.com	Image (File or URL): bwarto.png Browse
linked phrase	Width: 320 Height: 77 px 0 Cock ratio
http://example.com 🔍 🖤 😁	Alignment Befault O Left O Center O Right
quarto	Caption: optional caption text
optional caption text	Alternative text: Quarto logo and the word quarto spelled in small case letters
w: 320 h: 77 🛛 🖛 🕈 🖉 Lock ratio 🛛 😑	Link to:
laples	OK Cancel
First Header	Second Header
Content Cell	Content Cell
	Contant Coll



Medical datasets for teaching Reproducible Medical Research with R

Data cleaning in action

A	B C	D	E	F	G	Н	1	J	К	L	М	Ν	0	Р
1 Acute Kid	ney Injury (AKI) Study		C		import	thin a	liveet	vinto	CDCC		and arun	ah aama num	hara?	
2 A. Motiva	ted Resident		U c	an we	e import	uns o	irecu	у шю	3733		k and crune	ch some num	bersr	
3 Research	Project							1			1			
-								1 (-						
5 study_id	dob Race	Sex	Hispanic		Admission GFR	Day 1 gfr	DAY2GFR'	day 3 GfR	Day 4 GFR	DAY5 GFR	diagnosis	sediment .	HD catheter	AV fistula
6	. 11/11/1967 White	Male	Hispanic		17	12	10	22	34	4	1 hypotension	many granular casts	Yes	No
7	2 11/12/1990 Caucasian	Female	Not Hispanio	C	21	. 17	19 but poor	د 12 ۱2	19	2	o sepsis	muddy brown casts	Yes	No
8	8 may 5 1970 White	Male	Hispanic				15	11		2	4 bleeding out	many epithelial cell casts	No	No
9	3-Jun-81 Dther	Female	Not Hispanio	C	11	. 9	6	5	9	1	8 norovirus diarrhea	muddy brown casts	YES	YES
10	7/06/1984 Af-Am	Male	Not Hispanio	C	15	12	9	8	11	. 1	4 sepsis	muddy brown casts	YES	No
11	9/04/1980 Black	Female	н	X4	n percent	valid_	percen	1t 26	38	4	4 liver failure	many epithelial cell casts	No	No
12	5 13/01/1985 White	Female	Hispan	F	2 0.08	0.0	952381	.0 33	39	4	3 pneumonia sepsis	many granular casts	No	No
13	7 15/11/1968 Asian	F	Hispani	E	1 0.04	0.0	1476100	38	47	5	1 line sepsis	many epithelial cell casts	No`	NO
14	3 22/04/1982 Mixed	Male	Not His	Г 3	1 0.04	0.0	14/0190	39	36	4	1 C diff diarrhea	many epithelial cell casts	No	No
15	8 12/10/2003 White	M	Not His Fe	emale	/ 0.28	0.2	5333333	3 9	11	. 1	4 NSAID tox	muddy brown casts	Yes	NO
16	1. 6/11/1982 White	F	NH	M	3 0.12	0.1	L428571	.4 29	36	4	1 burns	muddy brown casts	Yes	Yes
17	12 9/07/1979 Caucasian	M	Not His	м	1 0.04	0 (M76100	37	44	4	8 pyelo	many epithelial cell casts	No	No
18	13 5=11-1984 Black	F	NH	M-1-	C 0.04	0.0		33	39	4	4 urethral obstruction	many epi cell casts	Yes	No
19	11 7/04/2004 Af-Am	M	Hispani	мате	0.24	0.2	282/142	41	44	. 4	9 GI bleed	many granular casts	No	No
20	15 8/03/1990 Asian	M	NH	Sex	1 0.04	0.0)476190	36	39	4	3 sepsis	rare gran casts	No	No
21	15 6/09/1983 WHITE	Male	Not His	<na></na>	4 0.16		N	A 39	47	5	1 salmonella	many granular casts	No	No
22	17 11/07/1986 WHITE	Female	Not Hispann			52	57	34	48	5	2 sepsis	rare granular casts	No	NO
23	1) 9/06/1987 Black	Male	Hispanic		22	20	24	24	39	4	8 esophageal varices	many epithelial cell casts	Yes	No
24	20 29/03/1979 Caucasian	Female	Hisspanic		232	19	15	18	28	3	8 dehydration	many granular casts	No	No
25	2. 11/08/1978 Caucasian	Female	Not Hispanio	с	177	14	12		28	3	5 high ostomy output	muddy brown	Yes	No
26	· · · · · · · · · · · · · · · · · · ·													
27	Note data collected from	m March to Ju	une 2022 while	rotating th	rough the Gastonia	VA								
28	Permission from Attend	ing doctor												

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.





DISCRETE

OBSERVATIONS CAN ONLY EXIST at limited values, often

@allison_horst



Variable types

Variable types – Example Cath dataset

- Dataset from the Duke University Disease Databank – an observational cohort study.
- Patients were referred to the clinic for chest pain and cardiac catheterisation was performed to diagnose and open blockages in the arteries, followed by stenting to keep them open.
- 3504 participants

Cath Dataset Description Document: This dataset was not published as a manuscript but has been generously provided by Frank Harrell.

Cath Dataset Description Document: This dataset was not published as a manuscript but has been generously provided by Frank Harrell.

Variable types – Example Cath dataset

Codebook

variable	position	Variable_Label	units	Codes	type
sex	1	Gender		0 = male, <mark>1</mark> = female	double
age	2	Age	years		double
cad_dur	3	duration of cadiac symptoms	days		double
choleste	4	Serum Cholesterol	milligrams per deciliter		double
sigdz	5	Significant Coronary Artery Disease Found on Cardiac Cath		0 = no, 1 = yes	double
tvdlm	6	Three Vessel or Left Main Disease Found on Cardiac Cath		0 = no, 1 = yes	double

B. Descriptive analysis for individual variables

Categorical variables

Graphical summaries

- Bar charts

Tabular summaries

- Frequency tables

Numerical variables

Graphical summaries

- Histogram
- Boxplots

Tabular summaries

 Numerical summary statistics of centre (mean, median mode) and spread (quartiles, percentiles, sd)

Cath Dataset Description Document: This dataset was not published as a manuscript but has been generously provided by Frank Harrell.

RE_wor	3 RE_workshop_data - RStudio Source Editor										
acath	×										
← ⇒	🚛 📔 🕆 Filt	er									
•	patient_id	sex	÷	^{ge} Fl ^{ge [Yea}	requency – count th	e numbe	r of males and fen	+ Nales It Main Disease by Cardiac Cath			
1				73	132	268	1	1			
2				68	85	120					
3				54	45						
4				58	86	245					
5				56		269					
6				64		NA					
7				65	76	NA					
8				41	15	247					
9				68	30						
10		1		52		NA					
11		1		48		NA					
12		1		35	44	257					
13		1		69	10						
14		1		58		168					
15		1		81		246					
16		1		58	79	221					
17		1		59	36	NA					
18		1		47	6	272					
19				66	8	257		0			
20		2		48	69	236					

- Frequency – count the number of males and females



- Serum cholesterol (mg per dl) of the participants who provided a response
- A frequency table would be long and messy. Not a great summary!

^	choleste ‡	n [‡]	percent [‡]	valid_percent [‡]
	29		0.0002853881	0.0004428698
	63		0.0002853881	0.0004428698
			0.0002853881	0.0004428698
	74		0.0002853881	0.0004428698
	89		0.0002853881	0.0004428698
	90		0.0002853881	0.0004428698
	93		0.0002853881	0.0004428698
	99		0.0002853881	0.0004428698
	105		0.0008561644	0.0013286094
	107		0.0005707763	0.0008857396
	108		0.0002853881	0.0004428698
	115		0.0005707763	0.0008857396
			0.0008561644	0.0013286094
	118		0.0002853881	0.0004428698
	119		0.0002853881	0.0004428698
	120		0.0008561644	0.0013286094
			0.0002853881	0.0004428698
	126		0.0002853881	0.0004428698
			0.0002853881	0.0004428698
	130		0.0005707763	0.0008857396
			0.0002853881	0.0004428698
	132		0.0005707763	0.0008857396
	134		0.0005707763	0.0008857396
24	135		0.0005707763	0.0008857396
25	136		0.0002853881	0.0004428698
26	137		0.0005707763	0.0008857396
	139		0.0005707763	0.0008857396
28	140		0.0011415525	0.0017714792
	142		0.0005707763	0.0008857396
30	143		0.0002853881	0.0004428698
	144		0.0002853881	0.0004428698

alid percent

000285388

0.0002853881

0.0011415525

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0005707763

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0002853881

0.0005707763

0.0002853881

0.0002853881

0.0002853881

0.0005707763

0.0002853881

0.0002853881

0.0002853881

0.0005707763

0.0002853881

0.3555936073

0.000885739

0.000442869

0.000885739

0.001771479

0.000442869

0.000442869

0.000442869

0.000442869

0.000442869

0.000442869

0.000442869

0.000442869

0.000885739

0.000442869

0.000442869

0.000442869

0.000442869

0.000442869

0 000442869

0.000885739

0.000442869

0.000442869

0.000442869

0.000885739

0.000442869

0.000442869

0.000442869

0.000885739

0.000442869

The	University	of Sydne
-----	------------	----------

- Serum cholesterol (mg per dl) of the participants who provided a response
- A bar chart isn't the best either! 😕



- Serum cholesterol (mg per dl) of the participants who provided a response
- A histogram of cholesterol, with frequency shown with equally sized bins (n = 30).





- Age (years) of the participants who provided a response

n

percent 80 0.0228310502 78 0.0222602740 70 0.0199771689 70 0.0199771689 55 0.0156963470 51 0.0145547945 37 0.0105593607 23 0.0065639269 26 0.0074200913 16 0.0045662100 20 0.0057077626 12 0.0034246575 10 0.0028538813 2 0.0005707763 2 0.0005707763

- A frequency table would be long and messy. Not a great summary!

← →	1	▼ F	ilter				age	÷	n		percent ‡		age		n		¢ percent	^	ag
	age		n	¢	percent [‡]	16		33		15	0.0042808219	31		48	13	34	0.0382420091	46	
		17			0.0002853881	17		34		30	0.0085616438	32		49	14	44	0.0410958904	47	
		18		1	0.0002853881	18		35		30	0.0085616438	33		50	18	36	0.0388127854	48	
		19			0.0002853881	19		36		34	0.0097031963	34		51	12	24	0.0353881279	49	
4		20		2	0.0005707763	20		37		46	0.0131278539	35		52	1	16	0.0331050228	50	
		22		2	0.0005707763	21		38		56	0.0159817352	36		53	12	28	0.0365296804	51	
		23			0.0002853881	22		39		61	0.0174086758	37		54	12	26	0.0359589041	52	
		24		2	0.0005707763	23		40		63	0.0179794521	38		55	12	26	0.0359589041	53	
8		25		4	0.0011415525	24		41		84	0.0239726027	39		56	1	56	0.0445205479	54	
		26		2	0.0005707763	25		42		76	0.0216894977	40		57	14	44	0.0410958904	55	
10		27		5	0.0014269406	26		43		90	0.0256849315	41		58	12	26	0.0359589041	56	
11		28	1		0.0031392694	27		44		102	0.0291095890	42		59	(96	0.0273972603	57	
12		29		9	0.0025684932	28		45		95	0.0271118721	43		60	ç	99	0.0282534247	58	
13		30	1	0	0.0028538813	29		46		130	0.0371004566	44		61		97	0.0276826484	59	
14		31	1	8	0.0051369863	30		47		111	0.0316780822	45		62		78	0.0222602740	60	
15		32	1	7	0.0048515982							- 45		02.			0.0222002140		

	age 🍦	n [‡]	¢ percent
61	78	5	0.0014269406
62	79	3	0.0008561644
63	80	1	0.0002853881
64	81	3	0.0008561644
65	82	1	0.0002853881



- Age (years) of the participants who provided a response
- A bar chart isn't the best either!





Age (years) of the participants who provided a response
A histogram of age, with frequency shown with equally sized bins (n = 20).



Shapes of the distribution



Examples of Skewed Distributions

Symmetric distribution

Asymmetric distribution

- Another way to summarise a numerical variable is through a boxplot.



Example Cath dataset – Summarising the boxplot

- A numerical distribution can be summarised by giving descriptions/measures of: Q2 (Median)



ade

JAMA Guide to Statistics and Methods: Nonparametric Statistical Analysis

Parametric versus Nonparametric stats

- NORMAL DISTRIBUTION
- Parametric
- Based on assumptions about data distribution and shape.
 - Normality
 - Though quite robust in large samples due to the central limit theorem.
 - Homogeneity of variance
- Mean and standard deviation reported.
- Suitable for larger samples, with a normal distribution.
- Examples:
 - 2-sample t-test
 - Generalised linear model (One-way Anova, Least squares regression)
 - Pearson correlation

With the rise of GLMs and other more flexible parametric methods, such as distributional regression, quantile regression and generalised additive models, nonparametric methods are becoming less useful (and are now less of a requirement for publication).

Nonparametric

- Not based on assumptions about data distribution and shape.
- Median and percentiles/quartiles/max and min reported.
- Suitable for smaller samples, skewed data or ordinal variables.
- Examples:
 - Mann-Whitney U
 - Kruskal-Wallis
 - Spearman correlation





Best practice in statistics: The use of log transformation

- Why check the distribution?
- Helps us understand skewness and any extreme values in the raw data.
- Common transformations: log, square root, inverse, Box-Cox.
 We cannot take a log of 0, so if the raw data contains values of 0, add a constant to the variable before taking the log.
 E.g. log(x + 1).
- Checking the raw data is an important part of EDA. However, when fitting linear models, we assess distributional assumptions using the residuals, not the raw data. See the <u>Linear Models 1</u> workshop for more information.



- UN: Blood Urea Nitrogen. Numeric. Range: 2-118. 53 missing values.



Distribution of un

The University of Sydney

<u>Thiomon Dataset Description Document</u>: This dataset is from Akbar K. Waljee and Peter D. Higgins, who de-identified data on CBC and chemistry testing at the University of Michigan for development of a machine learning algorithm to predict response to thiopurine medications in IBD patients.



- ALK: Alkaline phosphatase. Numeric, range 13-1938, 0 missing values.



Distribution of alk

The University of Sydney

<u>Thiomon Dataset Description Document</u>: This dataset is from Akbar K. Waljee and Peter D. Higgins, who de-identified data on CBC and chemistry testing at the University of Michigan for development of a machine learning algorithm to predict response to thiopurine medications in IBD patients.



- ALT: Alanine Transaminase. Numeric, range 1-10666, 18 missing values.



The University of Sydney

<u>Thiomon Dataset Description Document</u>: This dataset is from Akbar K. Waljee and Peter D. Higgins, who de-identified data on CBC and chemistry testing at the University of Michigan for development of a machine learning algorithm to predict response to thiopurine medications in IBD patients.

What is a systematic approach to conduct descriptive analyses for individual variables?

Categorical variables

Graphical summaries

- Bar charts

Tabular summaries

- Frequency tables

Don't forget to check for missing data/NAs!

For formatting tips, check your target journal's instructions to authors or recent publications in that journal!

Numerical variables

Graphical summaries

- Histogram
- Boxplots

Tabular summaries

- Symmetric?
 - Mean, standard deviation, minimum and maximum.
- Asymmetric?

•Median, quartiles, minimum and maximum.

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?

What is the research question?

Depending on who taught you statistics, your discipline, stats software or textbook, you may have come across some of these terms and more!

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

> Smoking (yes/no) Predictor, explanatory variable, risk factor, independent variable*

*This term is frequently used, but we don't promote it as predictors may be correlated and hence are not independent. Lung disease (yes/no) Response, outcome, dependent variable



Other functional classifications for variable types

- Covariate: a measured predictor (numerical predictor variable).
- Factor: (a categorical predictor variable).

Experimental design variables:

- <u>Design variables</u>: Based on the physical design of the experiment. They are often included in the analysis even if not 'significant' e.g., correctly partition the variance e.g., Block (batch of reagent, source of lab mice), subject ID, etc.
- <u>Treatment:</u> 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 numerical).
 - cumulative morbidity (sum numerical).
 - peak morbidity (maximum numerical).
 - 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).
- Numerical predictors: handle as numerical 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

Your variable types will dictate the type of statistical analysis you perform

- The type of outcome and explanatory variables you have will dictate the type of EDA and inferential analysis you can do, so it is crucial to think about this and your research question before you collect your data!
 - Is your outcome numerical or categorical?
 - Are your explanatory variables numerical or categorical?
 - Do you have one or multiple outcomes and/or explanatory variables?
- Choosing the wrong analysis can violate statistical assumptions and often won't run properly in your statistical software.



Step 5: Exploratory data analysis (EDA)

It will depend on the analysis and variables involved.

Basic EDA is where you plot the relationship of each predictor with the outcome, and you may need to reconsider data processing. E.g. Do I need to merge more categories together?

Two categorical variables

- Contingency table
- Side-by-side bar chart

A categorical and a numerical variable

- Tabulate summary statistics by groups
- Box-and-whisker plot by groups

Two numerical variables

 Scatter plot and correlation coefficient r

XY scatterplot and Pearson's correlation coefficient (r)

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


Examples for reporting descriptive analyses:

- Plotting gives a quick visual summary during EDA and highlights any issues.
- Tables are more publication friendly as they save space!
 - This step of the statistical analysis process is often seen in Table 1 of your manuscript.
 - See this paper for more information: Who is in this study, anyway? Guidelines for a useful Table 1.

Examples for reporting descriptive analyses: Cath dataset

Table 1. Summary statistics for the variables in a study of patientsundergoing cardiac catheterisation, split according to sex.

	Male (N=2405)	Female (N=1099)	Overall (N=3504)
Age [Year]			
Mean (SD)	51.6 (9.83)	53.8 (9.99)	52.3 (9.93)
Median [Min, Max]	52.0 [17.0, 82.0]	54.0 [20.0, 81.0]	52.0 [17.0, 82.0]
Duration of Symptoms of Coronary Artery Disease			
Mean (SD)	43.0 (58.1)	43.0 (58.3)	43.0 (58.2)
Median [Min, Max]	17.0 [0, 416]	20.0 [0, 404]	18.0 [0, 416]
Cholesterol [mg %]			
Mean (SD)	227 (47.3)	237 (56.9)	230 (50.6)
Median [Min, Max]	223 [63.0, 500]	230 [29.0, 576]	225 [29.0, 576]
Missing	836 (34.8%)	410 (37.3%)	1246 (35.6%)
Significant Coronary Artery Disease			
No	533 (22.2%)	637 (58.0%)	1170 (33.4%)
Yes	1872 (77.8%)	462 (42.0%)	2334 (66.6%)
Three Vessel or Left Main Disease			
No	1463 (60.8%)	909 (82.7%)	2372 (67.7%)
Yes	941 (39.1%)	188 (17.1%)	1129 (32.2%)
Missing	1 (0.0%)	2 (0.2%)	3 (0.1%)



Cath Dataset Description Document: This dataset was not published as a manuscript but has been generously provided by Frank Harrell.

Other examples for reporting descriptive analyses:

<u>Midurethral Sling vs OnabotulinumtoxinA in Females</u> With Urinary Incontinence

Table 1. Demographics and Baseline Characteristics No. (%) Characteristic OnabotulinumtoxinA (n = 71)^a Midurethral sling (n = 66)^a Age, mean (SD) [range], y 59.1 (11.4) [27-78] 59.0 (11.7) [33-87] Raceb Asian 2 (2.8) Black/African American 10 (14.1) 10(15.2) Native Hawaijan or Other Pacific Islander 1(1.4)1(1.5)White 55 (77.5) 54 (81.8) Unknown/not reported 3 (4.2) 1(1.5) Hispanic/Latina ethnicity, No./total (%) 15/71 (21.1) 6/65 (9.1) Education, highest level obtained was greater than high school, 46/68 (64.8) 32/63 (48.5) No./total (%) Currently smoking 7 (9.9) 9(13.6) No. of vaginal deliveries, median (IQR) 2 (1-3) 2(1-3) Total No. of deliveries, median (IQR) 2 (1-3) 2(2-3) Menopausal status Pre 7 (9.9) 14(21.2) Post 56 (78.9) 49 (74.2) Not sure 8 (11.3) 3 (4.5) Currently using estrogen by prescription 21 (29.6) 18 (27.3) BMI, mean (SD)^d 34.3 (8.3) 35.0 (7.6) Type of urinary incontinence^e Stress predominant 3 (4.2) 4 (6.1) Urge predominant 7 (9.9) 10 (15.2) Balanced 61 (85.9) 52 (78.8) 7.4 (4.0) Baseline incontinence episode daily frequency, mean (SD) 7.1 (4.1) Time from baseline visit to treatment, mean (SD), d⁴ 59.0 (38.5) 56.8 (43.4) Median (IOR) 49 (36-75) 46 (27-76) Treatment >90 d after baseline 7 (9.9) 10(15.2) Baseline UDI scores, mean (SD)9 Total 187.6 (38.2) 180.9 (36.5) Irritative 75.7 (16.6) 77.3 (15.4) Stress 86.6 (18.4) 80.3 (21.9)

Abbreviations: BMI, body mass index; UDI, Urogenital Distress Inventory. ^a The primary analysis population is defined as all participants who received any

 The primary analysis population is defined as all participants who received any treatment and have postbaseline efficacy data, regardless of randomized treatment.

^b Race categories were self-reported using check all that apply and specific closed options, including an unknown/not reported selection.

^c Ethnicity categories were self-reported using select only one, specific closed option, including an unknown/not reported selection.

^d BMI calculated as weight in kilograms divided by height in square meters.

^e The type of urinary incontinence is defined by responses at baseline on the UDI to the urgency urinary incontinence (UUI) item "Do you experience urine leakage related to a feeling of urgency? If yes, how much does it bother you?" and stress urinary incontinence (SUI) tem "Do you experience urine leakage related to physical activity, coughing or sneezing? If yes, how much does it bother you?" Creater bother reported on the UII item is classified as urge predominant, greater bother reported on the SUI item is classified as stress predominant, and equal bother reported is classified as barroed.

^f Baseline refers to the time of the first UDI assessment completion, which is used to determine eligibility. Participants were expected to receive treatment within 91 days of completing their baseline UDI.

^a The UDI total score ranges from 0 to 300, and the UDI irritative and stress scores range from 0 to 100, with higher scores indicating greater symptom severity.

Immunogenicity and Safety of Influenza and COVID-19 Multicomponent Vaccine in Adults ≥50 Years

Table 1. Participant Demographics and Baseline Characteristics (Safety Population)²

	Age ≥65 y		Age 50-64 y		
	mRNA-1083 (n = 2011)	HD-IIV4 + mRNA-1273 (n = 2006)	mRNA-1083 (n = 1993)	SD-IIV4 + mRNA-1273 (n = 2005)	
Age, y					
Mean (SD)	70.9 (5.0)	70.7 (4.7)	57.5 (4.3)	57.4 (4.2)	
Median (IQR)	70 (67-74)	70 (67-74)	58 (54-61)	58 (54-61)	
Age group, No. (%)					
50-64 y	2 (<0.1) ^b	0	1991 (99.9)	2004 (99.9)	
65-74 y	1593 (79.2)	1591 (79.3)	1 (<0.1) ^b	0	
≥75 y	416 (20.7)	415 (20.7)	1 (<0.1) ^b	1 (<0.1) ^b	
Sex, No. (%)					
Male	933 (46.4)	908 (45.3)	837 (42.0)	811 (40.4)	
Female	1078 (53.6)	1098 (54.7)	1156 (58.0)	1194 (59.6)	
Race, No. (%) ^c	n = 2000	n = 1994	n = 1978	n = 1981	
American Indian or Alaska Native	9 (0.5)	12 (0.6)	12 (0.6)	13 (0.7)	
Asian	25 (1.3)	36 (1.8)	50 (2.5)	39 (2.0)	
Black or African American	370 (18.5)	370 (18.6)	517 (26.1)	552 (27.9)	
Native Hawaiian or Other Pacific Islander	1 (0.1)	4 (0.2)	3 (0.2)	5 (0.3)	
White	1577 (78.9)	1565 (78.5)	1373 (69.4)	1343 (67.8)	
Multiple	14 (0.7)	4 (0.2)	19 (1.0)	21 (1.1)	
Other	4 (0.2)	3 (0.2)	4 (0.2)	8 (0.4)	
Ethnicity, No. (%)					
Hispanic or Latino	283 (14.1)	275 (13.7)	392 (19.7)	381 (19.0)	
Not Hispanic or Latino	1688 (83.9)	1689 (84.2)	1576 (79.1)	1603 (80.0)	
Unknown or not reported	40 (2.0)	42 (2.1)	25 (1.3)	21 (1.0)	
BMI					
Mean (SD)	30.3 (6.2)	30.1 (6.1)	31.1 (7.1)	31.6 (7.3)	
Median (IQR)	29.4 (25.9-33.8)	29.2 (25.7-33.6)	30.2 (26.2-34.9)	30.5 (26.6-35.4)	
BMI group, No. (%) ^d	n = 1995	n = 1984	n = 1963	n = 1980	
<30	1077 (54.0)	1096 (55.2)	951 (48.4)	927 (46.8)	
≥30	918 (46.0)	888 (44.8)	1012 (51.6)	1053 (53.2)	
Comorbidity group, No. (%)					
High risk ^e	1312 (65.2)	1292 (64.4)	1233 (61.9)	1264 (63.0)	
Low risk	699 (34.8)	714 (35.6)	760 (38.1)	741 (37.0)	
Influenza vaccine received since Sept 2022, No. (%)	1019 (50.7)	1016 (50.6)	783 (39.3)	784 (39.1)	
COVID-19 vaccine received since Sept 2022, No. (%)	853 (42.4)	854 (42.6)	632 (31.7)	611 (30.5)	

Abbreviations: BMI, body mass index calculated as weight in kilograms divided by height in meters squared); Ho IIV4, high-dose quadrivalent nactivated influenza vaccine; 5D-IIV4, standard-dose quadrivalent nactivated influenza vaccine.

^a The safety population included all participants who were randomized and received the study vaccination. Participants were included in the vaccine group corresponding to the vaccine they received. ^b Due to misrandomization. 2

participants aged 50-64 years were enrolled in the ≥65 years cohort and 3 participants aged ≥65 years were enrolled in the 50-64 years cohort.

Race and ethnicity of participants was self-reported according to multiple categories. The total number of participants with a reported race (excluding participants with race unknown or unreported): percentages are based on this total.

^d The total number of participants with a reported BMI group (excluding participants with unknown BMI group); percentages are based on this total.

High-risk comorbidity included having any medical history of autoimmune/immune-mediated autoimmune/immune-mediated disorders, hovod system disorders, hopatic disorders, mental impairment disorders, pulmonary disorders, or metabolism and nutritional disorders. Comorbidites based on self-report or review of medical encorder.

Tips from the consulting room



- Have your research question in mind, as well as a clear idea of what your publication goals are. Think about your 'story'.



- If you're planning to publish your work, be sure to review the journal's "Instructions for authors" and browse recent articles to guide your layout and presentation!



- Consider who will be reading your work and present your findings in a way that's easy for them to understand.
 - Is it clinicians? Is it industry? Is it other researchers? Is it a broad audience/lay people?
 - Try and make things as readable and accessible as possible.



How to present your results in an accessible way

- Ensure that each figure or table you include can be understood independently, without needing to refer to the main text.
- Don't overload the reader remember any additional information can be provided in your Appendix/Supplementary Materials and there will be maximum word counts for most journals.
- Group and order your information logically.
- Be consistent in your terminology throughout the manuscript and always provide the units of analysis. E.g. always call it "Intervention group" not "Intervention group" in one part and "Treatment group" in another, or "numerical predictor" in one part and "continuous predictor" in another.
- Consider colourblind friendly colour palettes and ensure high contrast between graph elements (e.g. lines, points).

Best-practice reporting guidelines



- <u>The Equator Network</u> (Enhancing the QUAlity and Transparency Of health Research) is a centralised platform for researchers to access a wide range of reporting guidelines.
 - Seeks to improve the reliability and value of published health research by promoting transparent, standardised and accurate reporting.
- There are guidelines for different study types to help you with your manuscript writing.
 - Provides you with a check-list of information required so that your manuscript can be easily understood by the reader and reproduced by other researchers.



Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions			
Observational					
studies	STROBE	Extensions			
Systematic reviews	PRISMA	Extensions			
Study protocols	<u>SPIRIT</u>	PRISMA-P			
Diagnostic/prognosti					
<u>c studies</u>	STARD	TRIPOD			
Case reports	CARE	Extensions			
Clinical practice					
<u>guidelines</u>	AGREE	<u>RIGHT</u>			
Qualitative research	SRQR	COREQ			
Animal pre-clinical					
studies	ARRIVE				
Quality improvement					
studies	<u>SQUIRE</u>	Extensions			
Economic					
evaluations	CHEERS	Extensions			



Best-practice reporting guidelines

Key Reasons for Adopting Reporting Guidelines STANDARDIZED REPORTING Reporting guidelines provide a standardized framework for documenting and reporting experimental procedures, methods, and results. They ensure consistent capture of essential information. ENHANCING REPRODUCIBILITY Complete and transparent reporting enables other researchers to replicate and verify study findings, minimizing ambiguity and increasing reproducibility of results. TRANSPARENCY AND TRUSTWORTHINESS Transparent reporting instills confidence, allowing readers to critically evaluate research methodology and identify limitations, biases, or sources of error. FACILITATING DATA SHARING AND REUSE Reporting guidelines promote data sharing and facilitate the integration of existing research, accelerating scientific progress and fostering collaboration. **IMPROVING DATA INTERPRETATION** Detailed reporting helps readers accurately interpret and understand presented data by providing essential contextual information and avoiding misinterpretation. CONSISTENCY AND COMPARABILITY Reporting guidelines ensure consistency and comparability within a field, aligning experiments with accepted practices and facilitating meta-analyses and literature reviews. **IDENTIFICATION OF METHODOLOGICAL BIASES** Transparent reporting enables identification of biases or methodological flaws that may impact reliability and validity, aiding critical assessment of the study's limitations. COMPLIANCE WITH ETHICAL AND REGULATORY STANDARDS Reporting guidelines incorporate ethical and regulatory considerations, ensuring adherence to legal, ethical, and safety

obligations in biomedical experimentation.

(g)

Best practices for data management and sharing in experimental biomedical research



Statistical analysis plans (SAPs)



- SAPs as well as study protocols are a great tool to improve communication and collaboration between you and other researchers.
- They should include information on:
 - Objectives and hypotheses
 - Primary and secondary outcomes
 - Study design
 - Data collection methods
 - Data management
 - Statistical analysis including handling of missing data and sensitivity analyses
 - Reporting of findings
- Australian Clinical Trials Alliance: Statistical analysis plan
- Jama: Guidelines for statistical analysis plans
- A template for the authoring of statistical analysis plans
- Developing a Quantitative Data Analysis Plan for Observational Studies

The University of Sydney

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 (Binary logistic regression)
- D. Multivariate analysis Confirmatory factor analysis (CFA)

Your variable types will dictate the type of statistical analysis you perform

- Examples using the generalised linear model framework:
 - 2-sample t-test = binary explanatory variable and numerical outcome.
 - ANOVA = multi-categorical explanatory variable and numerical outcome.
 - Simple linear regression = numerical explanatory variable and numerical outcome.
 - Binary logistic regression = numerical or categorical explanatory variable and binary outcome.
- For more information on the type of statistical test to run, see slide 98!

Example A: Linear models examples

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

- 1. Simple linear regression one numerical predictor variable.
- 2. ANOVA (control vs. treatment) for 2 groups = 2 sample ttest = simple linear regression with one binary predictor variable.
- 3. ANCOVA ANOVA with a covariate predictor.

For more detail on how to do these analyses including R code, attend our SIH <u>Linear Models 1: Linear</u> <u>Regression, ANOVA, ANCOVA and Repeated Measures (a Simple Mixed Model)</u> workshop!

Example A1: Linear models – Simple linear regression

Step 5: EDA – Plot the data in a scatterplot.

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



Example A2: Linear models – Control vs. treatment experiment

Step 5: EDA – Plot the data with grouped boxplots.

Step 6: Inferential analysis – ANOVA/2-sample t-test; P < 0.001; report predicted means and 95% CIs.



Example A3: Linear models – ANCOVA (ANOVA with a numerical covariate)

Step 5: EDA – Plot the data (differentiate categories of the treatment variable).

Step 6: Inferential analysis – ANCOVA or multivariable regression; P < 0.001; report predicted means with or without interaction and 95% Cls.



Example B: Survival analysis

- Scenario: <u>Worcester Heart Attack Community Surveillance</u> <u>Study</u> (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 time to event.
- **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.



The University of Sydney

Example B: Survival analysis

Step 6: Inferential analysis – <u>The logrank test</u> to compare the survival curves of two or more groups.

- 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: Binary logistic regression

- **Scenario:** Survey responses of 32 long-term smokers to determine the association between smoking status and lung-cancer diagnosis.
- Outcome: Lung cancer (no/yes)
- Predictors: Years smoking, BMI
- Primer on binary logistic regression

For more detail on how to do these analyses including R code, attend our SIH <u>Linear Models 2: Logistic</u> and Poisson/Count Regression - An Introduction to Generalised Linear Models workshop!

Example C: Binary logistic regression

Step 5: EDA – Log odds of the outcome (lung cancer) vs. numerical predictor (years of smoking) to check the linearity assumption.



Figure 2

Checking the linearity assumption graphically.

Example C: Binary logistic regression

Step 6: Inferential analysis – binary logistic regression to examine association between lung cancer (y/n) and years smoking (numerical), if linearity assumption is met.

For every additional year of smoking, the odds of lung cancer are approximately 1.98 times higher (95% CI 1.36 to 3.86).

- 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 <u>A Practical Introduction to Factor Analysis: Confirmatory Factor Analysis</u> from the UCLA: Statistical Consulting Group.

- **Scenario:** Confirm SPSS Anxiety as a factor explaining the common variance among the 7 items.



Example adapted from <u>A Practical Introduction to Factor Analysis: Confirmatory Factor Analysis</u> from the UCLA: Statistical Consulting Group.

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

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.



The University of Sydney



Final notes on Step 6: Inferential analysis

- We only showed some common examples of statistical analyses there are many different types of analyses you can consider:
 - Other Linear Models extensions such as Poisson regression and more complex mixed models - see our SIH <u>Linear Models 2:</u> <u>Logistic and Poisson/Count Regression - An Introduction to</u> <u>Generalised Linear Models</u> workshop.
 - Survival Analysis for 'time-to-event' outcome data see our SIH Introduction to Survival Analysis training!
 - Survey Data analysis see our <u>Design and Analysis of Surveys 1</u> and <u>Design and Analysis of Surveys 2: Advanced Topics</u>.
 - Other Multivariate Analyses for example PCA, Factor Analysis see our <u>Multivariate Statistical Analysis 1: Dimension Reduction</u>.
 - And more more workshops!



Final notes on Step 6: Inferential analysis

- Start simple and increase complexity step-by-step.
- Always consider/check the test/model assumptions.
- Report 95% CIs for estimates, e.g., predicted means/ probabilities/rates.
- For basic analyses consider more powerful (parametric) analyses first and use less powerful tests if assumptions are violated. e.g.:
 - Use a 2-sample t-test with equal or unequal variance for means, before a Mann-Whitney test.
 - Use a chi-squared test to compare proportions before a Fisher's exact test.
- Use knowledge of variable types to guide you through the systematic roadmap on the next slide!

Statistical inferential analysis roadmap



1

Where possible, if the assumptions are met, use the parametric test as it has more statistical power to detect differences than the nonparametric test equivalent.

*If you are unsure of the units in your study, see our <u>Experimental</u> <u>Design</u> workshop for more information!



Final notes on Step 6: Inferential analysis

- Univariate/univariable

 involving one variable,
 e.g. one outcome per analysis; analysis with one predictor variable.
- **Multivariate** multiple outcomes in the same analysis/model.
- **Multivariable** multiple explanatory variables in the same analysis/model.

- **Linear models** (LM numerical outcome).
- Generalised linear models (GLM – categorical outcomes, e.g. binary, ordinal, multinomial (for nominal outcome data) or Poisson/negative binomial regression (for count/rate outcome data).
- Mixed models (i.e. LM or GLM with a random effect = LMM or GLMM) where data are clustered in space or time, e.g. repeated measures/longitudinal data.

Further R resources

- There is a large online community of R users contributing to free 'packages' with data analysis functions, which leads to many ways of coding your analysis in R. This can be confusing. We recommend using tidyverse packages and tidy-centric code.
- See our SIH helpful links for guides on using <u>R and Rstudio</u>.
- LinkedIn Learning: R courses
 - Including Learning the R Tidyverse (2024), Complete Guide to R: Wrangling, Visualizing, and Modelling Data, and Cleaning Bad Data in R.

- RLadiesSydney: RYouWithMe

Other resources



Books on R

- <u>R for Data Science</u> by Hadley Wickham.

Statistical blogs and websites

- The Analysis Factor
 - Seven Steps for Data Cleaning
 - Best Practices for Organizing your Data Analysis
 - Best Practices for Data Preparation
 - Preparing Data for Analysis is (more than) Half the Battle
 - Four Weeds of Data Analysis That are Easy to Get Lost In

Further assistance at The University of Sydney



SIH

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

SIH Workshops

- Create your own custom programs tailored to your research needs by attending more of our Statistical Consulting workshops. Look for the statistics workshops on our training page or on our <u>Training</u> <u>calendar</u>.
- Sign up to our mailing list to be notified of upcoming training.

Other

- Open Learning Environment (OLE) courses
- Linkedin Learning

The University of Sydney



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 programs 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, found below:

Workshops and training



Online statistical consulting resources



sydney.edu.au/informatics-hub

Useful links

Use the links below to access your tool or platform quickly.



SIH Statistical Resources Collapse All | Expand All

Sydney Informatics Hub

Sites

Workshops and Workflows
 About our Workshops

- Research Essentials
- Experimental Design
- Power and Sample Size
- Statistical Modelling
- Linear Models 1
- Linear Models 2
- Linear Models 3
- Statistical Model Building
- Design and Analysis of Surveys 1
- Design and Analysis of Surveys 2
- Multivariate Dimension Reduction
- Meta-Analysis
- Survival Analysis

Sydney Informatics Hub

SIH Statistical Resources Collapse All | Expand All

- Sites
 - Welcome
 - Workshops and Workflows
 - The Statistical Consulting Unit
 - About Us
 - Contact Us
 - What to Expect in a Consult
 - C Staff
 - Acknowledgements
 - Helpful Links
 - Basic Statistical Theory
 - Linear Models
 - Logistic Regression
 - Meta Analysis
 - Power Analysis
 - Research Essentials
 - Statistical Inference
 - Statistical Software
 - Survival Analysis

- **Workshops and Workflows:** Download all our workshops with their step-by-step workflows on how to do statistical analysis.
- Helpful Links: 1-3 curated links on a variety of common statistical concepts and tests
 e.g. Basic theory, meta-analysis, software, etc.
 A great place to start looking for help!
- The Statistical Consulting Unit/What to expect in a consult: For tips on how to get the most out of your consult.
- You may need to click "Expand all" on the Sites sidebar to get an overview of the available pages.

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.

Power and sample size workshop

- Shows the steps and decisions researchers need to make when designing an experiments to ensure sufficient sample e.g., power, minimum sample 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 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.