

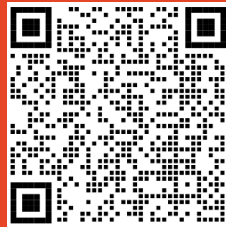
Design and Analysis of Surveys 1: An Introduction

Presented by Omar Arnaiz
Statistical Consultant
Sydney Informatics Hub
Core Research Facilities

sydney.edu.au/sydney-informatics-hub



THE UNIVERSITY OF
SYDNEY

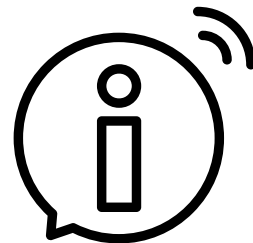


Slides available here

CRICOS 00026A TEQSA PRV12057



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.”*

During the workshop



- Ask short questions or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.



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



- Challenge questions will be encountered throughout the workshop.

After the workshop

These slides should be used after the workshop as reference material and include **workflows**

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

1on1 assistance

- You can email us about the material in these workshops at any time
- Or request a consultation for more in-depth discussion of the material as it relates to your specific project. Consults can be requested via our Webpage (link is at the end of this presentation)

Research Workflows

Why do we need a research workflow?

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

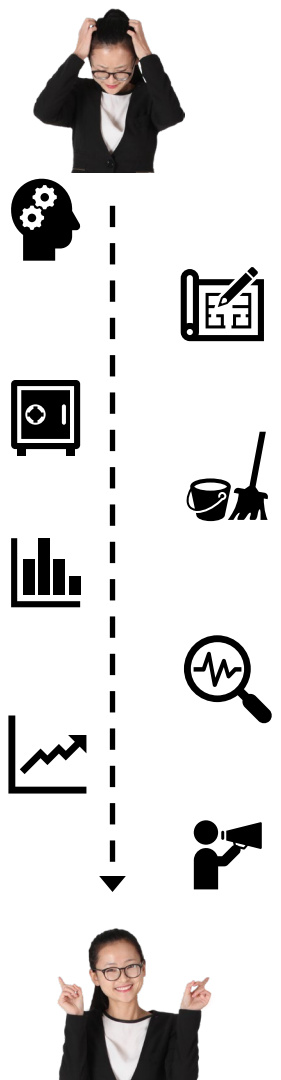
So... what is a workflow?

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



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**



Contents

1. Designing a Survey
2. Data Export and Cleaning
3. Introduction, Design Examples, Data Cleaning, EDA (Exploratory Data Analysis), Reporting and Analysis for these instruments:
 - Categorical
 - Free Text / Open Enders
 - Continuous
 - Likert
4. Tricks of the Trade

Designing a Survey

Treat your Respondents as Friends, or at least with Respect

They are doing you a favour, so return it by:

- ***Keeping it as short as possible.***
- Recognise they only have so much cognitive ability, so use it wisely and make things as simple and easy as possible i.e. avoid ***Respondent Fatigue***
 - Put harder questions up front, easier ones like demographics at the end.
- If possible, **offer an incentive**, even a single \$100 randomly picked reward shows you value their time and will get you a much higher response.
 - Ethics can sometimes take a dim view of this as they feel it may lead to people doing it only for the money. One way around this is to use lucky draws for a small amount, e.g. I know researchers who have got a lucky draw of six \$50 grocery vouchers through.
 - You will need to know who they are and possibly ask for their email to do this. If confidentiality is important than an independent 3rd party may need to collect this information and administer the incentive.

Common Sections, and order of

1. **Screenener:** Used to identify and screen out people you don't want in the survey, and/or to funnel them into appropriate sections e.g. maybe it's a survey about childhood obesity and there is 1 section for children and a different one for their parents.
2. **Main Body:** Where you ask most of the questions of interest. Can itself be split into sections.
3. **Demographics and other user information questions used for profiling respondents:** Can also be asked in the screener if required to screen people out. If some are required for screening it may be easier to group them all in 1 place i.e. at the beginning.

Lead-in / Intro

TREAT YOUR RESPONDENTS AS FRIENDS, OR AT LEAST WITH RESPECT

EXPLAIN HOW THIS HELPS THEM

In the intro explain what the purpose of the research is and how important their information is to it. Try and relate it back to them e.g. “this research is being used to develop better treatments for COVID” or “ This research will be used to build new phones with the features you want”.

TELL THEM HOW LONG IT TAKES

LET THEM SAVE AND COME BACK TO IT

Must have if a long survey (>10 min).

SHOW YOU RESPECT THEIR CONFIDENTIALITY

say something like this

Your individual responses are treated with strict confidentiality. Results will only be reported for groups with 10 or more completed respondents in a particular demographic, work area, or combination thereof, so as to ensure no individual can be identified. Data may be used by in research and benchmarking, but individual and organisational confidentiality will be protected.

Towards the end of the survey there are open-ended questions where you can give more information about your previous answers or bring things up that aren't covered in other places in the survey. Unidentified copies of your comments will be included in the final report.

How to make a great Survey: The Basics

1. **Write down your research questions.** Then work out what you need to collect and how you want to analyse it to answer them, what subgroups do you need in terms of things like demographics, etc. Consider logistics and mode of data collection i.e. will it be done on a phone, offline, etc.
2. **Write a draft in word** based on your needs, desktop research and qualitative work.
3. **Review this draft** after leaving it a few days, ideally a week. Keep doing until no new edits. Plan and budget for multiple reviews both here and during subsequent reviews.
4. **Seek feedback** from friends and colleagues.
5. **Set up in Survey Tool** e.g. REDCap or Qualtrics.
6. **Enter some test data.** Make it up by thinking of some likely but very different respondents and enter as they would. Try to break it, ensure the functionality works e.g. branching logic, invites, calculations, follow-ups.
7. **Export data** and ensure you have set it up so the data exports in an easy to analyse format.
8. **Review the Survey tool and exported data** after leaving it a few days, ideally a week. Keep doing until no new edits.
9. **Seek feedback** by sending the link to friends and colleagues and asking them to fill it out.
10. **Final Review** after leaving it a few days, ideally a week. Then send the word document and possibly link to Ethics for approval.
 - *Even small changes can be problematic as new ethics approval is often required which can take months,* which is one reason it is so important to get it right before submitting to ethics.
11. **Go Live!!!**
12. **Review the first 12-50 respondents for any problem** e.g. look for missing categories you should add by reviewing the open ender linked to 'other' to see if it has a lot of responses representing the same thing. Particularly worth keeping an eye on as the survey progresses to avoid time consuming back coding later

Step 1 & 2) Write a draft in Word

Your needs

- Spend some time writing out your research questions, an analysis plan, etc. Refer to our Research Essentials workshop for more info.

Desktop Research

- Use similar surveys as templates.
- Appropriate scales and questions.

Qualitative work

- If possible always do, even its informally i.e. ask relevant friends and colleagues
- To explore:
 - Relevant dimensions of interest – informs what questions to ask.
 - Unknown area's of interest (unknown to the researchers that is).
 - Possible sensitivities.
- Common Types are:
 - Focus groups: Typically 5-10 people to discuss the questionnaire.
 - In-depth interviews with just 1 person.
 - Cognitive Interviews to REDCap understand the thinking process that elicits the response.
 - Online Qual via:
 - Surveys.
 - Communities.

Dimensions/Factors of interest

Prior research and qualitative work often identify dimensions of interest the researcher wants to understand. Indeed, this is often the primary reason for the survey.

Even if not the focus it's still a good idea to identify possible ***dimensions*** that might impact the research prior to developing the survey. For example:

- Business: Price, Quality, Animal Welfare.
- Vaccines: Education, Previous Bad experiences.

These dimensions are then included in the survey, which is used to quantify their impact.

If one is using rating scales it is common to assign 2-5 statements per dimension and use these to quantify each one. The simplest way is to simply average them, a more complex way is to create an **index** from them (as covered in Surveys 2).

Randomise order of statements and categories/ options they can select from

The order people see statements and categories can bias the results.

- It is common to show people a list of categories/options and ask them to select from them e.g. what emotion does this music elicit. Categories/options seen first are more likely to be selected. Not all categories need to be randomised, if they are objective there is less of a need than if they are subjective. And if there is a natural order it should not be done either e.g. age.
- Statements seen first can have more attention focused on them, and their order can introduce contextual differences.

As such always consider randomising the order people answer statements and the categories they can pick from.

The overall questions do not need to be randomised.

For example:

- If showing a list of musical instruments and asking people to select the ones they use it's not that important to randomise them.
- If asking 15 questions/statements about what types of music they like, then we should randomise them.
- If asking what emotions certain music elicits, then we should randomise them.
- But keep the above 3 overall questions in the same order.

Always finish with open enders

Such as

- Some ones ***specific*** to your topic since these are easier to analyse.
- A final ***general*** “Is there anything else you would like to tell us, especially anything we could do better next time?”

Why?

- Gives ‘colour’ to your report, often worth adding in a few open enders to emphasise a point you’re making. Some people will not be impacted by facts and figures, but a good story will drive your point home.
- If you missed something you will notice it here. So keep an eye on this as the survey progresses and adjust if necessary.

“Humans are narrative beings, hardwired to understand the world around us through stories. Even if we know intellectually that the stories we see are fiction, we hold on to the emotional truth that strikes us and use that to navigate our existence.” Alice Tovey

Advanced Survey Features

- Pipes/Logic allow different questions to be asked based on answers of previous questions e.g.
 - Ask some screener questions and if they don't pass then reject e.g. have they had the experience/disease you are researching.
 - Maybe you are researching housing policy and what to ask different questions to people who rent vs own.
 - Talk to Research Data Consulting for more info and to attend their REDCap workshops (link at the end of workshop with other references).

Other Considerations

Anonymous is different to **Deidentified**

- **Anonymous** means the data to identify participants isn't collected i.e. no-one knows who they are, not even you.
- **Deidentified** means the data to know who they are is collected. But you can remove their name and other identifiable features so they can't be identified, for example when sending the data to be analysed.
- If you want to follow people up such as when doing a pre/post test then you can't use Anonymous data, as you need to identify people.

Logistics

- Do you need the form to be completed offline e.g. don't go to a farm and expect your online survey to work? If so, there are ways to complete computer surveys offline.
- Many short surveys, such as diary's, will be completed on people's phones whilst on the move. Phone surveys require a small screen, so some questionnaire setups don't work well on a small screen. If you intend it to be done on a phone **test it on a phone**.

Experimental Design: External Validity

Inferential analysis is often a key objective of surveys. Meaning you want to infer from your sample what is happening in the general population. This requires your sample to be **Externally Valid** i.e. valid outside of itself. Or in other words it needs to be a good representation of the population you want to be inferring to.

A frequently overlooked problem when using online surveys is their reliance on online panels, which rarely match the general population meaning they lack Externally Validity. This calls into question their representativeness and hence if they can be generalised to the wider population. **Refer to our Experimental Design workshop** for more info external (and internal) validity and how to ensure your sample can be generalised to the wider population.

A common way around this is to recruit one's own sample and not rely on established panels e.g. a disease database. And then email the survey link to them.

Experimental Design: Sample Size

Quotas/cells are often used to ensure sufficient sample to profile/compare groups of interest e.g. severe vs not severe disease, none vs social vs light vs heavy smoking, men vs women.

- The minimum sample size for each cell is often set as a minimum quota for the field company to achieve.

This means the total sample no longer generalises to the general population i.e. lacks external validity. *Weighting is often used to weight each cell so together they represent the total population.* This is quite complicated and beyond the scope of this workshop.

- One thing to always consider up front is that weighting requires population information. One can get this information from other surveys such as the census or even the screener if set up right and the panel is representative of the general population.
- For example, if we quoted 4 smoking cells to have $n=150$ (none, social, light, heavy smoking) we can't combine them and expect that to represent the general population. However, if we knew the actual population ratio was 70:10:10:10 then we could weight the 4 cells accordingly. If we didn't know this and we assume the survey was sent out to the general population then we can track how many people fall into each of these categories before they are screened into the cells to get this ratio (even if not all of them proceed to the main survey).

Follow ups, dropouts and differential attrition

It's quite common to use surveys to follow people up e.g. to see how they're faring 1, 6 and 12 months after their treatment. We expect some attrition over time, which is when people drop out and don't answer the survey anymore. We need to account for this when calculating the initial sample size.

For example, if we want 200 people at the end of the study and expect 10% to drop out then we need to start with 220 people.

It's also common for the ***control group to have higher attrition*** as people feel like they didn't get the treatment so why should they bother! It can be 5% lower than the treatments attrition.

Continuing the example above that means the controls expect 15% attrition and need to start with 230 people.

It's different for every domain, so best to refer to other papers to see what their attrition rate was.

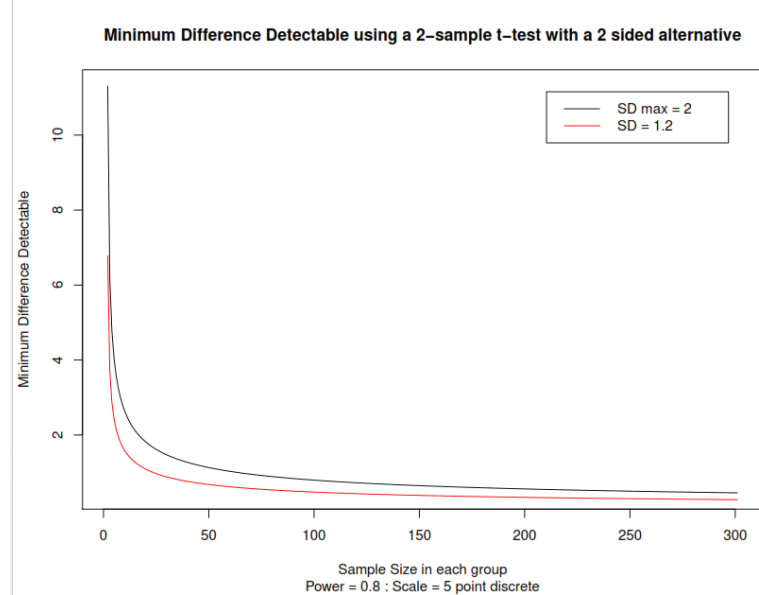
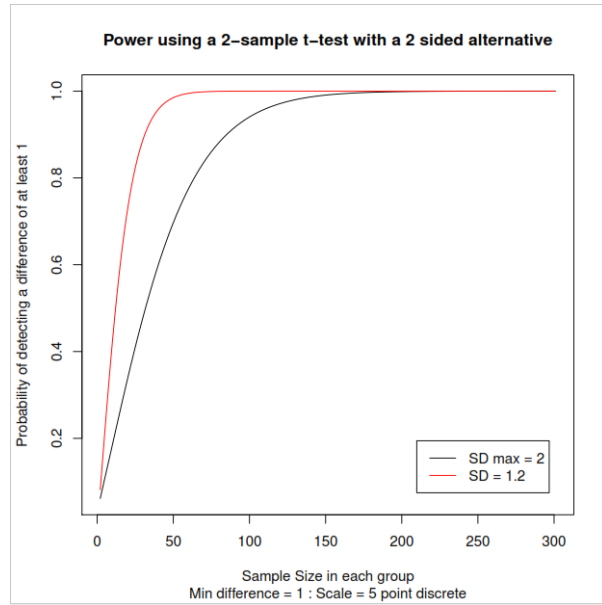
Experimental Design: Power Analysis

Can be hard to do since there are so many different types of scales. It is often done on those questions that are the main focus.

So, for example, if we had some Likert scales on attitudes we wanted to research and lots of demographics we would likely focus on the Likert scales so something like the following 2 power charts would be useful.

Our Power and Sample size workshop has more information and workflows on this topic.

Experimental Design: Power Analysis



This helps explain why so many Market Research Companies often recommend a sample size of 100-150 per cell. (Since the power curves flatten here).

Learning Outcomes

Keep it as easy as possible for respondents to give you quality answers.

- **Your survey should be short as possible** and, let them save and come back.
- Invested respondents – **let them know what the survey is for, respect confidentiality, and let them have their say via open enders.**

Do you need external validity?

- Collect information that helps you compare your sample to the population you want to make inferences on.
- **See our Experimental Design workshop for more information.**

Think about the sample size you need to collect for your study.

- **Consider your most important questions and any subgroups you want to report.**
- **See our Power and Sample size workshop for more information.**

Data Export and Cleaning

Data Export – Stacked vs Unstacked

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

Wide/unstacked format

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

Long/stacked format

Data Export – Stacked vs Unstacked

UNSTACKED

- **Don't Use It**
- Often used, incorrectly.
 - In some experimental designs seems the obvious first choice.
 - Usually requires processing before analysis which can be quite time consuming, difficult and open to error.
- **FORMAT:** Stores respondents in different sheets e.g. each treatment might have it's data in a different sheet/table. Which usually need to be merged before analysis.

STACKED

- **Use it**
- Much easier to analyse and store data.
 - MOST analysis software expect data in this format. That said there are exceptions, most notable being the SPSS > Repeated measures module.
- **FORMAT:** Stores respondents in a single sheets, with each variable in the same column. There are usually extra 'indicator variables' to define things like treatment.
- REDCap and most other survey instruments can export like this if set up correctly.

How to get your Survey Software to export as Stacked

1. Define a variable in the survey that defines each treatment or group.
 - If possible, have this filled automatically and do not ask the respondents to do so since they will often either not know or some will get it wrong e.g. our feedback survey link automatically fills in the workshop and date, so you don't need to.
 - Or you might know which emails/people were in each treatment and then you will need to code it up after exporting
2. To allow different treatments/groups to have different questions use branching logic/routing/piping. These allow respondents to see different questions based on how they answer previous ones.
3. Now you can export the data as stacked with no/minimal post export data processing.

DON'T

Setup as different surveys (or sub surveys) since then you will need to export them individually and process them into stacked.

Research Essentials has a workflow for turning unstacked data into stacked data.

Common Cleaning and Quality Control Checks

More detail follows in relevant sections – but a synopsis of common cleaning and checks are:

Non response: #/% who didn't answer i.e. missing values.

Response time: longer times may indicate a harder question and not necessarily a problem with it. What is a problem are Racers i.e. people finishing too quickly, they could even be bots – they are usually removed.

Continuous variable distributions: are they behaving as expected? Look for:

- Outliers
- Poor differentiation e.g. only “Agree” being used in a 5 point Likert scale.
- Flatliners e.g. people who answer Likert scales using the same score, they may be removed.

Categorical variables: are they behaving as expected? Look for:

- missing categories you should consider adding by reviewing the open ender linked to ‘other’. If it has a lot of responses representing the same thing consider adding them as a hard coded option.
 - Particularly worth keeping an eye on as the survey progresses to avoid time consuming back coding later.
 - But can be hard to do in an Academic setting as it often requires going back through Ethics. Which is why it is so important to do some pilot questionnaires and qualitative work before submitting to Ethics.

Flatlines \ Straightliners

- Are often removed.
- Easy way to find them is look for respondents whose answers have a $SD=0$ for such statement batteries.
- Ways to identify if flatlined data is valid is discussed in a subsequent Likert scale section.

Please indicate how much you agree or disagree with the following statements:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Qualtrics is awesome	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Chocolate is the best	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Oxygen is important	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Crime doesn't pay	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting bitten by a shark would be fun	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I dislike my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Common Processing

Transform Likert to “Top Box” binary agree/disagree data for easier reporting (as explained in the Likert section below).

Back code Free Form Text to a categorical variable for easier analysis or sentiment analysis (as explained in the in the categorical variable section below).

Learning Outcomes

Set up the survey tool so your next steps are easier. **Stacked data is usually what you want for storage and analysis.**

- Set up automatically created indicator variables for times, etc.
- If using repeated measures, make sure respondents have the same ID at each point.

Cleaning and quality control checks are important – better to take your time here rather than redo analyses or worse.

- Missing data.
- Racers.
- Continuous variable distributions i.e. outliers and asymmetric distributions.
- Likert data – poor discrimination, flatliners
- Categories – counts, missing categories in other/free text.

Consider additional processing.

- Top box from Likert data.
- Categorical variables from free text and regrouping small sample categories.

Different Types of Survey Questions

Introduction

Design Examples

Data Cleaning

EDA (Exploratory Data Analysis)

Reporting

Different Types of Survey Questions

The different types of questions one can use are often called Survey Items. And together form the questionnaire (or instrument) respondents fill out.

A useful way to describe Survey Items is by their scale. Common types are:

- Categorical
- Free Text
- Continuous
- Likert scales

These sections discuss the following workflow steps for each of these:

- Introduction
- Design i.e. examples and best practise
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Inferential Analysis
 - Reporting
 - Analysis

Exploratory Data Analysis (EDA)

Used to get a **broad understanding** of the data.

And to **look for problems** such as:

- Outliers
- Missing data
- Obviously wrong data

EDA is covered in more detail in the Research
Essentials Workshop.

Inferential Analysis: Is split into 2 stages

Reporting

- Surveys tend to simply report the results first using charts or tables.
- There is rarely any need for formal hypothesis testing, however 95% CI's should be used if possible since this tells us accuracy and factors in sample size e.g. just knowing that 75% of people agree with something might seem high. But if it's 95% CI was 50-100% than it doesn't seem so good anymore. However if it was 73%-77% this is an accurate and high estimate.
- Often split or filtered by variables of interest e.g. demographics, disease severity, etc.
 - Testing to see if there are differences between groups is often done either using p-values and/or including CI's.

Analysis

- Once the basic reporting has been done one then moves onto more complicated analysis e.g. Multivariate Maps (refer to our Multivariate workshop), Segmentations, Driver Analysis, etc.
- And then Predictive Modelling, which is only occasionally done with survey information for example:
 - Preference / Volume Estimation. Often using Choice Models.

Hypothesis testing vs Screening/Exploratory analysis

There is considerable debate about when Multiple Comparisons should be used, preferences can be quite domain specific.

One generally **always tests 'within model and/or factor' comparisons, but rarely between model comparisons** i.e. also known as correcting for multiple **testing** to distinguish it from multiple **comparisons**. For example: if we had a single model for freckles with 2 predictors: hair colour (4 options) and eye colour (4 options) we would generally correct each predictor for multiple comparisons independently i.e. assume 6 comparisons were being done for each. We wouldn't sum up the total comparison and correct for 12. Similarly if we ran 2 different models each with a different predictor we would correct each one independently.

1 useful distinction I often make is the difference between Hypothesis Testing vs Screening/Exploratory Analysis.

Hypothesis testing

- Requires corrections for Multiple Comparisons, e.g. Bonferroni, Tukey, Holmes, False Discovery Rate. For more information on correcting for multiple comparisons refer to our Linear Models 3 workshop.
- Is when we are testing apriori theories developed from previous research or modelling and are the focus of the paper. Usually only a few are made.
- Often used to make important decisions with minimal or no supporting evidence.
- EXAMPLE: Randomised clinical trials to evaluate 3 vaccines, Comparing a new formulation to the existing product, Land management trials.

Screening/Exploratory Analysis i.e. Screening lots of tests for possibly interesting pattern.

- Often doesn't correct for all multiple comparisons being done.
- Is when we do lots of tests looking for unknown associations or interesting patterns.
- Often used to suggest future research.
- If used to make decisions must be in conjunction with other information e.g. other studies, qualitative work, prior expert knowledge.
- EXAMPLE: Pharmacological study on 1000's of off the shelf medications impact on covid to identify those worth moving into better randomised clinical trials , analysing a survey with lots of questions and splits, driver analysis between numerous sensory/hedonistic variables and liking, data mining.

Surveys

Surveys often consider analyses of each question a different test, so we don't correct for multiple testing.

We also consider different splits of the same variable as different tests e.g. if comparing different medical treatments between genders, age and BMI we don't correct for all of them at the same time. **Instead of using strict hypothesis testing we take the view that these p-values are used to screen** all the different comparisons being done to see what might be worthwhile incorporating into the story and to generate hypotheses to be tested in future research.

We do however often correct for comparisons between different categories within a single variable e.g. if we had 4 age groups that's 6 different pairwise comparisons which we would usually correct for. Sometimes though we can have so many different categories to make even this problematic as correcting for multiple comparisons in the normal ways usually means nothing is worthwhile reporting.

As such we can also report both. For instance, if one was comparing some statements to a benchmark one can use colour, font and/or asterisk's to signify whether something has a p-value < 0.05 **with and without correcting for multiple comparisons (MC).**

The basic idea is **that as we are more sure of those corrected for multiple comparisons we bring more attention to them.**

Method	P<0.05 No MC correction	P<0.05 MC correction
Colour	Light Red	Dark Red
Asterisk	*	**
Bold or not	Not Bold	Bold

Significance testing, colour coding and screening

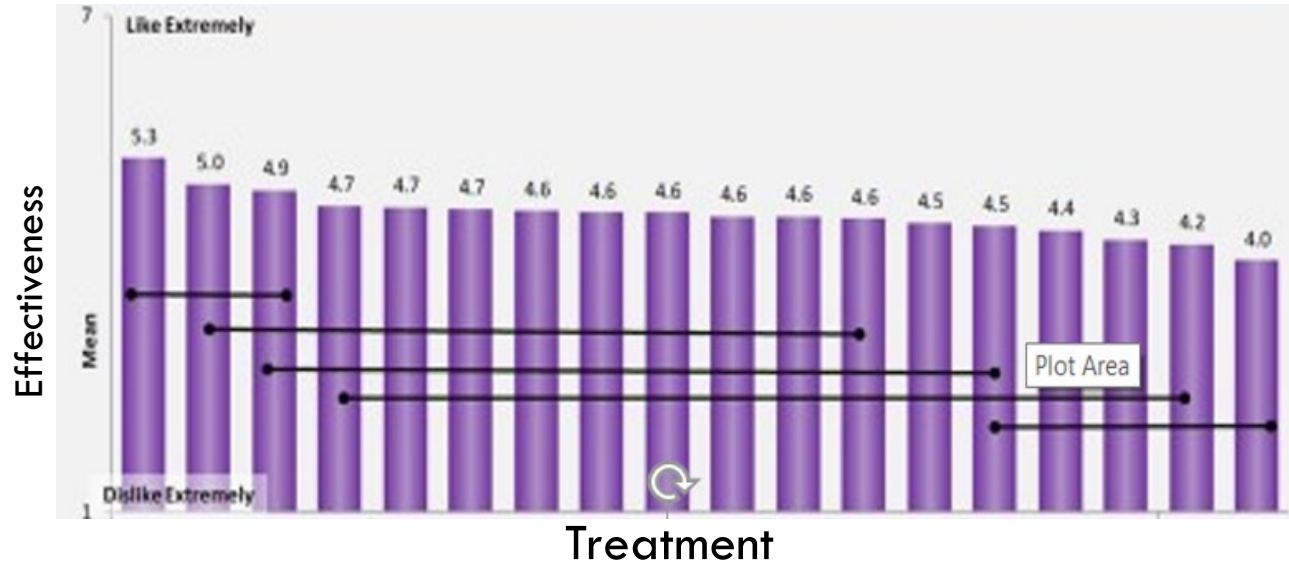
Example 1 - Colour

Importance of Animal Welfare on purchase decisions	% who agree
Australian Average (Benchmark)	50%
Vegetarian	90%
Byron Bay	60%
Low Socio Economic Band	20%
Sydney	53%

Example 2 – No colour so it can be used in more journals

Importance of Animal Welfare on purchase decisions	% who agree
AUSTRALIAN AVERAGE (BENCHMARK)	50%
Vegetarian	90% **
Byron Bay	60% *
Low Socio Economic Band	20% **
Sydney	53% *

Homogenous Subset Example



- Bars linked with a black line form a homogenous group i.e. there is no significant difference.
- Duncan's Multiple Range Test (MRT) is one way to get these.

The different types of multiple comparison tests are covered in more detail in the Linear Models III Workshop.

Learning Outcomes

Consider the type of question you need for the type of data you want to collect (more on this up next).

- **Categorical.**
- **Free text.**
- **Continuous.**
- **Likert scales.**

Exploratory Data Analysis – *the first step before any analysis*

- Plots and summary stats - check for missing data, obviously wrong data, and/or outliers.
- **See our Research Essentials Workshop for more on EDA.**

Inferential Analysis – split into Screening/Exploratory Analysis and Hypothesis Testing.

- Hypothesis testing for main question/s. Should use **multiple comparisons for different levels in factors** but usually not for different questions.
- Can also consider reporting both regular p-values (exploratory/screening) and multiple comparison correction p-values (stronger evidence if “significant”).
- **See Linear Models 3 for more on multiple comparisons.**

Categorical / Discrete Variables

What is a Categorical Scale?

Simple put, **anything where people select an option, and the basic summary is a count** i.e. the number of people who picked an option.

Nominal

- Single response e.g. please pick your favourite flavour of ice cream, gender, hair colour, etc.
- Multiple response e.g. please pick all flavours of ice cream you like.

★ Which of the following websites have you visited on a desktop or laptop computer in the past 12 months? Select all that apply.

- ☐ Amazon
- ☐ Walmart
- ☐ Target
- ☐ Walgreens
- ☐ CVS
- ☐ Wayfair
- ☐ None of the Above

Don't forget to
randomise their order

What is a Categorical Scale?

Ordinal

- A scale with **discrete** options with inherent order e.g. how much do you like this ice cream, how often do you exercise, age, income.
- They can often be answered on a continuous scale as well e.g. age, income.

How do you feel today?

- ☒ 1 – Very Unhappy
- ☐ 2 – Unhappy
- ☐ 3 – OK
- ☐ 4 – Happy
- ☐ 5 – Very Happy

How satisfied are you with our service?

- ☒ 1 – Very Unsatisfied
- ☐ 2 – Somewhat Unsatisfied
- ☐ 3 – Neutral
- ☐ 4 – Somewhat Satisfied
- ☐ 5 – Very Satisfied

Design: Code frames

- Many platform still use a **numeric code frame with a label e.g. 1 = Mumbai**.
- Historically this was because the shorter numeric code was more efficient in terms of computational speed and amount of memory required e.g. the data was actually stored as a 1, but could be displayed as “Mumbai”.
- As computers got more **efficient some people were tempted to just use the label. Which can be a bad idea, since it complicates updating the label and retaining backwards compatibility in our code.**
- Example 1) One might have a dataset on cities. If one then extracts cities by name for a city specific analysis what happens if that city changes name? The entire code now needs to be updated with the new city name. However if the original code used the factor code this is not needed.
- The Indian city now called Mumbai used to be called Bombay. Say it was originally coded as Bombay = 1, and the code referenced it as 1. Then we could simply update the name by changing the code frame to be Mumbai = 1, and the rest of the code stays the same. But if we only had the label “Bombay” we would need to update it throughout the code.
- Example 2) One can also correct spelling mistakes.
- **Conversely** using the actual names can make coding easier to read and review since the actual reference is explicitly stated i.e. Mumbai instead of 1. However, if the text string is very long than a numeric code can be much easier to use than writing out the entire sentence each time!

Design: Use existing Code Frames

Employment

- Use the ABS National Statistics Socio-Economic classifications
- Don't forget
 - Homemaker
 - Retiree

Design: Useful to have an OTHER option

What instrument do you play?

- Piano
- Guitar
- Wind instrument
- **Other**
 - If picked then have an open ender pop up that says “If other please specify” (requires routing logic in REDCap).
 - Such open enders can be a great source of new and surprising info.
 - Data Entry and Analysis.
 - They can all be reported as an ‘other’ category, or if some new ones occur often enough can be back coded and treated the same as the apriori categories.
 - Worth reviewing when in field to see if any are occurring enough to be added as new categories. Avoids lots of time consuming manual back coding later.

Design: Biological Sex vs Gender Identity vs Sexual Orientation

Transgender people and other gender and sex minorities have always been part of society, even if these identities have not been encoded in previous data collection efforts. It is a scientific reality that much research is focused on response variables that can be moderated by gender identity, sex characteristics, or both. It is time to end the erasure of GSM in our standard data collection procedures both for the sake of inclusivity and for the sake of decreasing measurement error and bias.

The idea that one needs more than male/female for biological sex, gender identity and sexual orientation is a relatively new one. I didn't see it gaining much traction before 2020. It is now necessary to consider if asking more than a Sex question with a binary *male/female answer* is required.

There are several reasons for this change:

- A wish to make gender options beyond the traditional binary female/male more acceptable in mainstream society.
- Recognition that:
 - medical studies need to know biological sex and that hormone therapy may influence and impact the safety, efficacy and impact of a drug or other treatment.
 - a better understanding in this area will lead to better policy to ensure people aren't "falling between the cracks". And delivery of different services/goods to people in a more targeted way.
 - An interesting example is clothing. In general, size and shape is determined by biological sex, but style is determined by gender identity and sexual orientation. So knowing more about the market for male clothing, for female bodies, will make it easier to make clothing for these people.

Before deciding how to ask and measure this consider

What are you interested in? This tells you if you need 1 or 4 questions:

1. current biological sex
2. biological sex at birth - often required in medical studies where genetics plays an important role.
3. gender identity
4. sexual orientation

Theoretically it's likely best to be clear whether one is asking about biological sex or gender identity. BUT we need to consider:

- If the extra information will be used, does it deliver useful extra insight?
- Is it worth the potential trauma of delving into people's private lives. Will not asking cause trauma?
- Is it worth the extra questions-remember we want to keep our surveys short!

Realistically, in many applications it doesn't really matter if a single Sex question is used and some people interpret it as biological sex and others as gender identity. Even if error or bias is introduced it will be small as most people relate to this binary framework. However, in small sample sizes it could have a large impact if sex or gender is strongly correlated with the response of interest.

Which to ask, and when both Biological Sex and Gender Identity should be included.

If asking a single 'gender' question one first needs to decide if biological sex or gender identity is required. Biological sex is often required for medical studies, while gender identity may be more important in social studies. Psychological studies often need both.

Even for medical studies it may be worthwhile asking both since it shows non binary people that sex is required for medical purposes, while their gender identity is still valued and respected. This allows the researcher to respectfully identify transgender participants, which can be important as hormonal changes may impact the treatment and be biologically important, and from the viewpoint of holistic medicine identity itself may be important.

If asking both:

- ask Sex as Sex assigned at birth
- explain why to show some respect on why you are delving into such a sensitive topic.
- it's not uncommon to use male/female for Sex, and man/women for Gender.

The minimum number of options

For both sex and gender I would suggest these **3 options as a minimum**. Statistically they will **usually** capture most people's intended meaning. Meaning the error/variance/bias will be minimal. That said they may not be sufficient for reasons of inclusion and depending on the population being sampled.

- Female
- Male
- Prefer not to say

Personally, I would usually **replace 3 with 4/5 when asking about gender and 5 for sex** since it covers all possibilities. If people prefer to not say they still can at 5, but this encourages a response. When analysing one may choose to combine them if there is insufficient sample to analyse them separately.

- Non-binary, Gender Fluid, gender non-conforming
- Other (pipe to free text)
- OR None of the above

One can also add a sexual orientation question (with an opt out).

Other considerations

- In situations where the data analyst has no control over the data collection procedure, such as a meta-analysis, they are responsible for noting any ambiguities or possible sources of bias in the data at hand.
- The privacy of personal information, especially in small studies, can leave minorities vulnerable to discovery if the information collected is too specific.

References

There is no established method yet, partly because what is culturally expected and acceptable is changing so quickly. As such it is an active area of research. There is a lot of discussion and fine tuning ahead. This is one of the sections I most frequently update.

References

- ABS standard for Sex, Gender, Variations of Sex Characteristics and Sexual Orientation Variables (2020) <https://www.abs.gov.au/statistics/standards/standard-sex-gender-variations-sex-characteristics-and-sexual-orientation-variables/latest-release>
- Australian Government Guidelines on the Recognition of Sex and Gender (2015) <https://www.ag.gov.au/rights-and-protections/publications/australian-government-guidelines-recognition-sex-and-gender>

Design: Ordinal Scales – Anchor all points with labels

Reduces noise (variance) since all respondents know exactly what each point represents and people who mean the same thing don't use different points.

Ensures **good discrimination** since respondents are biased towards selecting anchored categories.

Respondents tend to use them as if they are **equally spaced** so labels should usually match this.

Have a good range from 2 extreme points e.g. like extremely to dislike extremely.

If not every point is anchored **at the very least every other one** i.e. don't just anchor the ends.

- **Should only need to be done for 9 point scales** since it can be hard to find 9 anchors.
- If you do this **expect the ones without anchors to have less data**. Some people see this as a bad thing, others suggest that it's OK since only people who have a nuanced view and want to differentiate use the non labelled 'in between' points and we expect fewer people to do this.

How important are surveys to your startup's success?

Not at all Important	Slightly Important	Moderately Important	Very Important	Extremely Important
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Design: Ordinal Scales – Optimal # of points

Bipolar: 5, 7 or 9 points are common.

- Organised around a **clear middle point** with either side going in the opposite direction. It usually has the same number of categories either side which are labelled using opposite wording e.g. Not at all Important vs Very Important.
- More points give better differentiation, but too many become ambiguous and only add noise:
- 5 point scales are good for the general public.
- **Experts can differentiate more consistently** so if asking for expert opinion more than 5 points can be used without adding noise.

Not Extremely Important 0	Not Important 0	Neutral 0	Important 0	Extremely Important 0
------------------------------	--------------------	--------------	----------------	--------------------------

Uni polar: 5 points is common.

- Organised from low to high, without a clear middle point.

Not At All Important 0	Slightly Important 0	Moderately Important 0	Very Important 0	Extremely Important 0
---------------------------	-------------------------	---------------------------	---------------------	--------------------------

GOOD EXAMPLES

5 point scale – Bipolar Likert scale for Importance

Not Extremely Important 0	Not Important 0	Neutral 0	Important 0	Extremely Important 0
------------------------------	--------------------	--------------	----------------	--------------------------

5 point scale – Unipolar for Importance

Not At All Important 0	Slightly Important 0	Moderately Important 0	Very Important 0	Extremely Important 0
---------------------------	-------------------------	---------------------------	---------------------	--------------------------

5 point scale – Bipolar Likert scale for Agreement

Strongly Disagree	Disagree	Unsure (or neither)	Agree	Strongly Agree
------------------------------	-----------------	--------------------------------	--------------	---------------------------

7 point scale – Bipolar Likert scale for Agreement

How satisfied are you with our service?

Extremely Dissatisfied	Moderately Dissatisfied	Slightly Dissatisfied	Neutral	Slightly Satisfied	Moderately Satisfied	Extremely Satisfied
---------------------------	----------------------------	--------------------------	---------	-----------------------	-------------------------	------------------------

9 point scale – Bipolar Likert scale for Agreement

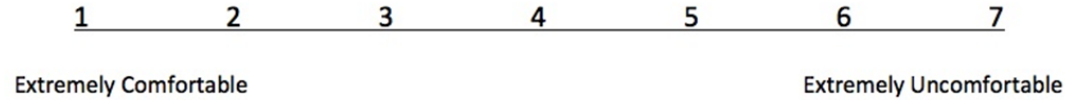
Expect the labelled points to have more data.

Strongly Disagree		Disagree		Unsure		Agree		Strongly Agree
------------------------------	--	-----------------	--	---------------	--	--------------	--	---------------------------

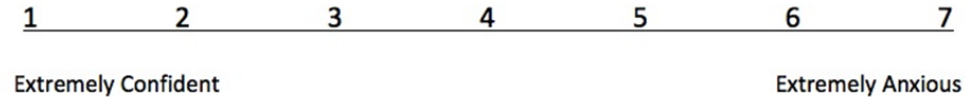
Strongly Disagree	Disagree	Moderately Disagree	Mildly Disagree	Unsure	Mildly Agree	Moderately Agree	Agree	Strongly Agree
------------------------------	-----------------	--------------------------------	----------------------------	---------------	-------------------------	-----------------------------	--------------	---------------------------

BAD EXAMPLES - Not all points are anchored

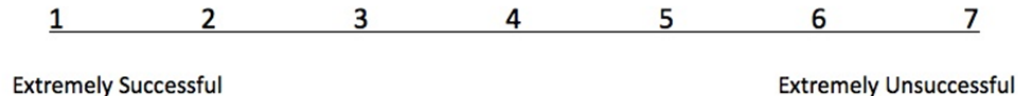
1. How comfortable were you with your talking?



2. How confident were you with your talking?



3. How successful did you feel with your talking?



INTERESTING EXAMPLES

If you have come to see the exhibition “We Don’t Need A Map”, how would you rate your experience?

Disappointing	OK	Good	Great	Awesome
----------------------	-----------	-------------	--------------	----------------

SMHAT: Sport Mental Health Assessment Tool

- Not ideal as rather subjective, could frame these more objectively as actual frequencies. As used across different time frames and sports this more generic phrasing may be required when describing the method. But when implemented still best to include relevant objective time frequencies.

- ☐ Applied to me very much, or most of the time - ALMOST ALWAYS
- ☐ Applied to me to a considerable degree, or a good part of time - OFTEN
- ☐ Applied to me to some degree, or some of the time - SOMETIMES
- ☐ Did not apply to me at all - NEVER

Design: Ordinal Scales – Natural metric anchors

Questions with a natural metric

- Frequency, Amount, Spend, Weight, Probability.
- Label with numbers not words since it's more accurate e.g. frequency would be Number of times per week, month etc rather than regularly, occasionally etc.

Questions with no natural metric

- E.g. liking, Likert, certainty, importance, happiness, satisfaction, quality.
- Use words or phrases, not numbers.
- Note that Likert scales are covered in more detail in their own section.

Design: Ordinal Scales – Natural metric anchors

How often do you feel pain when you exercise?

- **Don't use this scale, too vague**

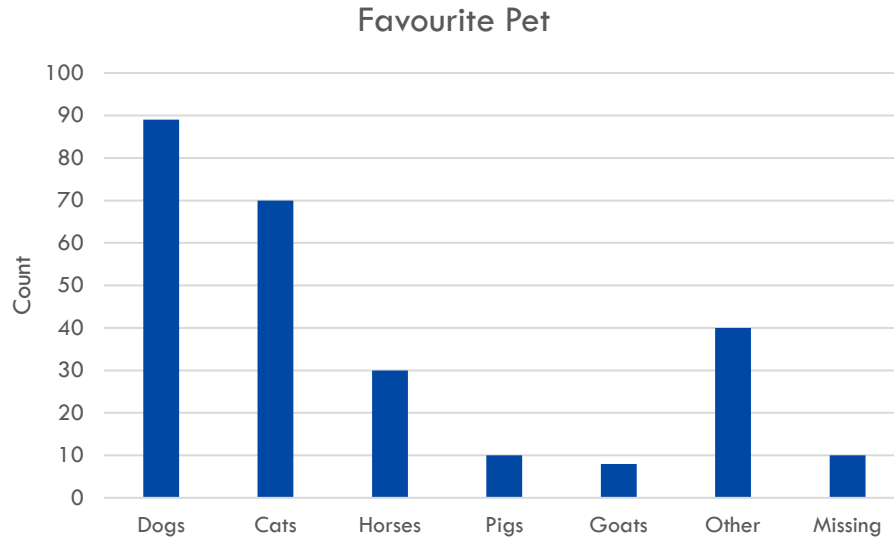
- Never
- Very little/ rarely
- Occasionally
- Sometimes
- Most of the time
- All the time

- **This Scale is better, since everyone agrees on what it means**

- Never
- 1 out of 4 times
- Half the time
- 3 out of 4 times
- All the time

EDA (Exploratory Data Analysis)

Check to ensure it makes sense, and there aren't too many missing

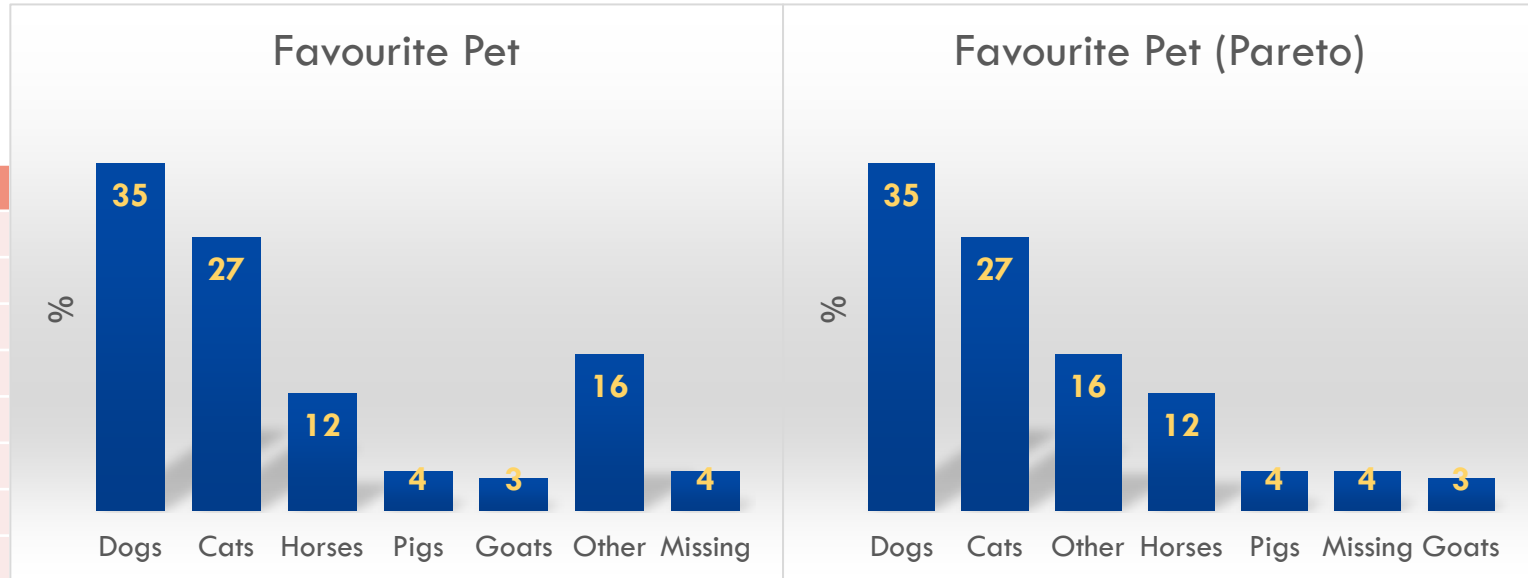


Reporting

Usually reported as tables or charts, refer to Likert section for other examples.

- % are often used as they are easier to compare between studies and easier to understand intuitively.
- Often useful to include the actual counts/% as labels within the bar.
- Pareto Charts are often easier to interpret as they sort from highest to lowest.

Pet	Count	%
Dogs	89	35
Cats	70	27
Horses	30	12
Pigs	10	4
Goats	8	3
Other	40	16
Missing	10	4
total	257	100



Reporting Splits and comparing to Benchmarks

Example 1 – Colour

Importance of Animal Welfare on purchase decisions	% who agree
Australian Average (Benchmark)	50%
Vegetarian	90%
Byron Bay	60%
Low Socio Economic Band	20%
Sydney	53%

Example 2 – Black and White (some journals require this)

Importance of Animal Welfare on purchase decisions	% who agree
AUSTRALIAN AVERAGE (BENCHMARK)	50%
Vegetarian	90% **
Byron Bay	60% *
Low Socio Economic Band	20% **
Sydney	53% *

Analysis

Nominal

- Logistic regression (covered in Linear Model II workshop)
- Poisson regression (covered in Linear Model II workshop)
- Multinomial
- Chi-squared analysis and its mapping equivalent Correspondence analysis
- Proportion (%) tests

Ordinal

- Above plus:
- Ordinal regression (Discussed in Likert section)

Assumption testing

Both Chi-squared and the proportion test make some assumptions about the frequency of an event in order to work.

Proportion test A common rule of thumb is that $np > 5$ and $n(1-p) > 5$ e.g. if the expected proportion is 2% and the sample size of the cell is 100 then $np = 2$, which is less than 5. Meaning these methods may not apply. (The same applies if it was 98% since $100 * (1 - 0.98) = 100 * 0.02 = 2$).

Chi-squared test A common rule of thumb is that all expected counts must be > 5 . Some domains may have some additional conditions too.

If these assumptions are not met

- Fishers Exact Test can be used.
- Merging categories with small samples sizes can also fix the problem.

Learning Outcomes

Categorical scales are where the respondent picks an option, and the basic summary is a count.

- Nominal – either single response (pick one) or multiple response (pick all that apply).
- Ordinal e.g. Likert scales or continuous data grouped into ranges.

Code frames – each category is a number with a label.

- Good for changes over time and if you need to correct typos only the label needs updating also less keystrokes when programming. Can be bad for sense checking e.g. what is a 6?
- Try to use existing code frames and have an Other to capture anything missed.

Biological Sex/Gender – consider what you need and treat respondents with respect.

- Options: ***Female***, ***Male*** and ***Prefer not to say***. Probably also have ***Non-binary***, ***Gender Fluid***, ***gender non-conforming***, and ***Other*** (pipe to free text).
- See ***ABS for the latest guidelines***.

Ordinal Scales – unipolar or bipolar, 5-point Likert for non-experts, anchor all points with labels.

- If there is a natural metric – use it i.e. use numbers.
- If there isn't a natural metric, use words.

EDA – bar plots (Pareto!), counts and percentages.

Analysis

- Reporting vs a Benchmark.
- Several modelling options e.g. Linear regression (continuous), Ordinal regression (Likert), Binomial regression (Agree/not Agree). **See Linear Models 1 and 2 workshops.**
- Proportions Test and Chi-squared – don't forget assumptions.

Free Text / Open Enders

What are Open Enders/Free Text?

When the respondent types in their answer e.g. “Please tell us what you found useful about today’s workshop”

Are very useful to get unexpected information. Done correctly they are a type of Qualitative research.

Use with care. If a code frame is more appropriate use them with an *Other* option. Not doing so can lead to a lot of work post processing the data and the open enders may be more ambiguous leading to missing data e.g. a survey that had gender as an open ender got 34 different answers, most of which were versions of male or female rather than non binary options e.g. malle, male, man, boy, female, women, lady, etc!!

Design

If using a lot of them try keeping each to single topic since this makes them easier to process and interpret.

Ensure respondents have enough space to write and proof read their answer.

Consider limiting the characters to ensure succinct answers.

Data Cleaning and Processing

Back code Free Form Text to a categorical variable for easier analysis

- Generally done by grouping similar statements into the same category, with those not fitting into a category lumped into an “Other” category.
 - Review the categories using Bar charts. Idea is categories should have enough answers to be useful and reduce the # in Other to a manageable level. Remember that when analysing groups an absolute minimum of 10 per group is usually required (refer to Linear Model workshops).
- Sentiment analysis
 - A simple code frame for comments which can then be used in “Sentiment analysis”:
 - Positive
 - Neutral
 - Negative
 - There are lists of positive and negative words that can help automate this, ChatGPT and other NLP’s can also be used.
- ChatGPT and other Natural Language Processing (NLPs’s) can make this back coding easier. Last I looked ChatGPT would categorise free text and also tell you why.
 - Personally, I would review the answers first myself. Even if there are a lot, say 300, one can quickly skim them to get a feeling for what is happening. This should help you give ChatGPT directions on how to categorise them, which will lead to better results.
 - This is another reason it’s worth having the open ends targeted and not general. It’s easier to skim targeted questions, partly as many of them will be empty. Then a general one with lots of unrelated answers.
 - I would also check the results. You don’t need to review all of them, but at least some from each category. You’re looking for things like:
 - Are there other subgroups that ChatGPT missed?
 - Has it categorised them correctly?
 - A simple code frame for comments which can then be used in “Sentiment analysis”:
<https://www.researchnewslive.com.au/2023/01/16/chatgpt-ai-a-market-researchers-best-friend/>

Reporting

If complicated they are often read and a short report prepared.

If processed to categorical data then all the methods shown for that data type can be used.

Word Clouds: There are lots of online tools for this.



Learning Outcomes

Open Enders/Free Text questions are where respondents type in an answer.

- Useful to get unexpected information.
- Consider if **a code frame with an Other option that pipes to free text would save you back coding time.**

Keep them to a theme as it easier to understand/process answers.

Can **back code to a categorical variable** e.g. positive negative sentiment analysis.

- ChatGPT and other platforms may be useful.


Continuous Variables

What is a Continuous variable?

Examples of Continuous variables are:

- Line scales aka Visual Analogue Scales (VAS)
- Age and income if asked as such
- Likert scales are sometimes treated as continuous variables

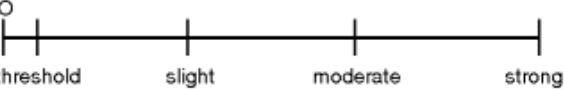
Line Scales aka Visual Analogue Scales (VAS)

a) 

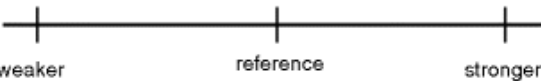
AROMA

b) 

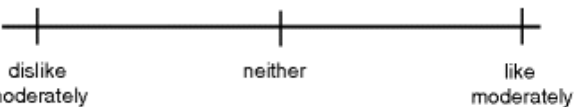
Pepper Heat

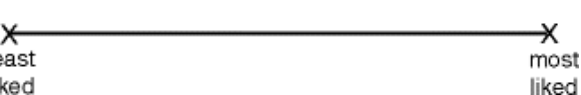
c) 

Sweetness

d) 

overall opinion

e) 

f) 

Avoid Line Scales, where possible use Categorical scales instead.

Adds measurement error e.g. ask 100 people to mark the dead middle on a 100mm line scale and I guarantee they will be between 45 to 55.

- But if you must use them at the very least show a halfway mark and even a $\frac{1}{4}$ and $\frac{3}{4}$ marks too.

Lawless H., Heymann H. (2010) Scaling. In: Sensory Evaluation of Food. Food Science Text Series. Springer, New York, NY.
https://doi.org/10.1007/978-1-4419-6488-5_7

Tricky use of a Line scale for ranking lots of things

It can be hard to rank lots of different things since people usually rank the top and bottom 3-5 consistently but do quite poorly with the middle options.

Say we had 300 types of food we wanted to rank on “nutritional value”. We could create a line scale anchored by certain meals and ask people to Drag and Drop the other food types onto it.



Poor
Nutrition

Good
Nutrition

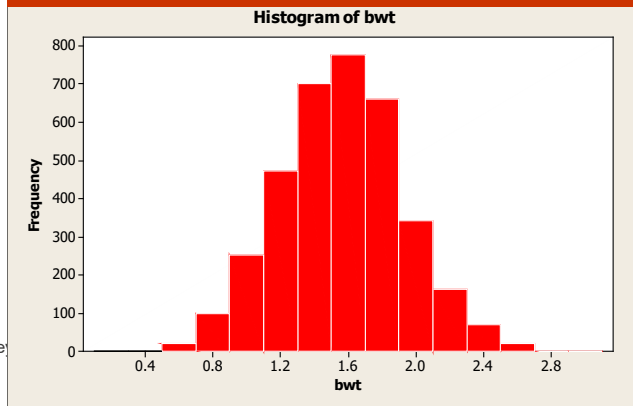
EDA (Exploratory Data Analysis) and data cleaning

A histogram or density plot should be used to understand the distribution and look for any problems such as:

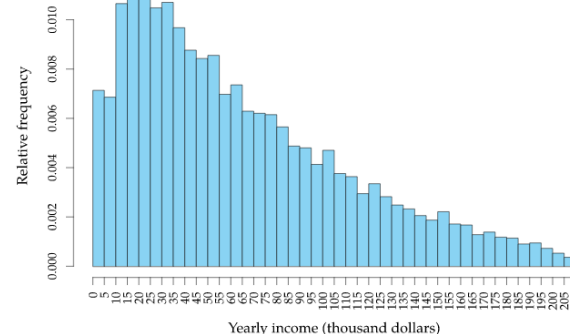
- Outliers (which might be Data Entry mistakes and require removing e.g. someone who says they are 230 years old).
- Unexpected distributions.
- If symmetric report using the average, if not and highly skewed decide if the median would be better e.g. house prices, income.

Also look at the # of actual and missing

Symmetric Distribution



Asymmetric Distribution



Reporting

Tables or plots of **Averages, ideally with Confidence Intervals** are usually reported. If highly skewed medians may be used instead e.g. house prices and income. Refer to Likert for some examples.

Characteristic ^b	Males				Females			
	n	Mean	95% CI	Range	n	Mean	95% CI	Range
Body mass (kg)	8	16.9	15.2–18.7	14.0–20.5	8	16.8	5.4–18.1	13.8–19.5
HFL (cm)	8	47.1	46.3–47.8	45.5–48.5	8	46.8	45.8–47.7	45.0–49.0
WBC ($\times 10^3/\mu\text{L}$)	6	6.2	5.0–7.5	3.9–7.8	7	4.8	4.3–5.2	3.9–5.6
Monocytes (% [$\times 10^3/\mu\text{L}$])	6	4.8	2.5–7.1	0.9–9.1	7	2.3	0.34–4.19	0.0–7.9
Fibrinogen (g/dL)	6	0.53	0.45–0.62	0.4–0.7	7	0.4	0.31–0.49	0.3–0.6
Ca (mg/dL)	8	9.7	9.3–10.2	8.8–11.1	8	10.5	10.1–10.9	9.8–11.4
TP (g/dL)	8	3.8	3.6–4.0	3.3–4.1	8	4.6	4.0–5.1	3.7–5.9
Globulin (g/dL)	8	1.64	1.37–1.90	1.0–2.1	8	2.5	1.98–2.95	1.3–3.7
Bicarbonate (mEq/L)	8	24.4	21.7–27.1	20.4–32.6	8	20.3	17.9–22.6	15.2–24.0
ALP (U/L)	8	243	204–282	170–350	8	345	290–399	211–436
GGT (U/L)	8	34.8	26.9–42.6	21–51	8	69.3	52.8–85.7	35–101

^aHandlers were able to approach and handle neonates with minimal excitement, no nets or chemicals (Severud et al. 2015a).

Analysis

Is very research dependant and there are too many options to list here. Common models would be regression, ANOVA, etc (which are covered in our Linear Model I and III workshops). Also refer to the Likert section below for some examples.

Learning Outcomes

Examples of continuous variables include:

- **Line scales** i.e. Visual Analogue Scales. Beware of measurement error!
- **Age, Income, etc. if asked this way.**
- **Likert scales** are sometimes treated as continuous.

EDA – look at the distributions.

- Outliers as well as data entry mistakes.
- Unexpected distributions.
- Symmetric – report average, asymmetric/skewed – report median.

Analysis

Research dependent e.g. regression, ANOVA, etc.

See our Linear Models 1 and 3 workshops for more information.

Likert Scales

What is a Likert scale?

Are a bit weird. Their survey item and analysis can be either discrete or continuous.

Even weirder they are often asked as a discrete variable, but analysed as a continuous one!

Named after their inventor psychologist Rensis Likert (pronounced lick-ert not like-art or like-ert)

Fundamentally, they ask people their level of Agreement or Disagreement to a question.

**Strongly
Disagree**

Disagree

**Unsure
(or neither)**

Agree

**Strongly
Agree**

Data Collection: Avoid Double Barrel Questions

Double Barrel questions ask about more than 1 thing. The problem is that you don't know exactly to which the respondents are responding to.

For example: "Please agree or disagree with the following statement: My internet should be faster and more stable".

Should instead should be split into 2.

Please agree or disagree with each of the following statements:

1. My internet should be faster.
2. My internet should be more stable.

Statement Batteries / Matrices / Grids

Please indicate how much you agree or disagree with the following statements:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Qualtrics is awesome	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Chocolate is the best	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Oxygen is important	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Crime doesn't pay	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting bitten by a shark would be fun	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I dislike my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

>>

Don't forget to
randomise their order

Statement Batteries – ALWAYS make all Positive (I love swimming) or Negative (I hate swimming) i.e. NO REVERSE CODING

That way 'agree' on the Likert scale means the same thing. Either you like it or you hate it.

- Having a couple statements reversed is occasionally touted as a way to find bots and respondents not paying attention. I do not recommend!

As we usually have this huge list with most of them usually being 'positive' what are we effectively training respondents to do?

- We're training them that the agree part of the scale (often on the right) is 'good' and hence the left is 'bad'.

So what do you think happens when we purposely try to confuse them by making some negative so now the other side of the scale should be used?

- Some people get confused and mark it wrong.

PROBLEM

- We don't know who marked it 'right' vs who got confused.
- We have purposely introduced noise.
- We haven't treated our respondent as a friend.
- We don't trust the data with the reversed scale and often therefore don't use it.

A better way to find bots and disinterested respondents

Ask at least 2 questions where **answers on opposite ends of the scale make sense**, while keeping them both positive (or negative) e.g. *Oxygen is Important vs Getting bitten by a shark would be fun*.

Please indicate how much you agree or disagree with the following statements:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Qualtrics is awesome	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Chocolate is the best	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Oxygen is important	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Crime doesn't pay	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Getting bitten by a shark would be fun	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I dislike my friends	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Although the *I like my friends vs I dislike my friends* statements are breaking the “keep it either positive or negative” rule they are such easy to answer questions it is unlikely to confuse people.

Don't forget to randomise their order

A better way to find bots and disinterested respondents

- Here's a slightly less silly example, where we expect the answers to be ***on opposite ends of the scale make sense.***
- Having them on the ***same topic lets us check they are being consistent.***

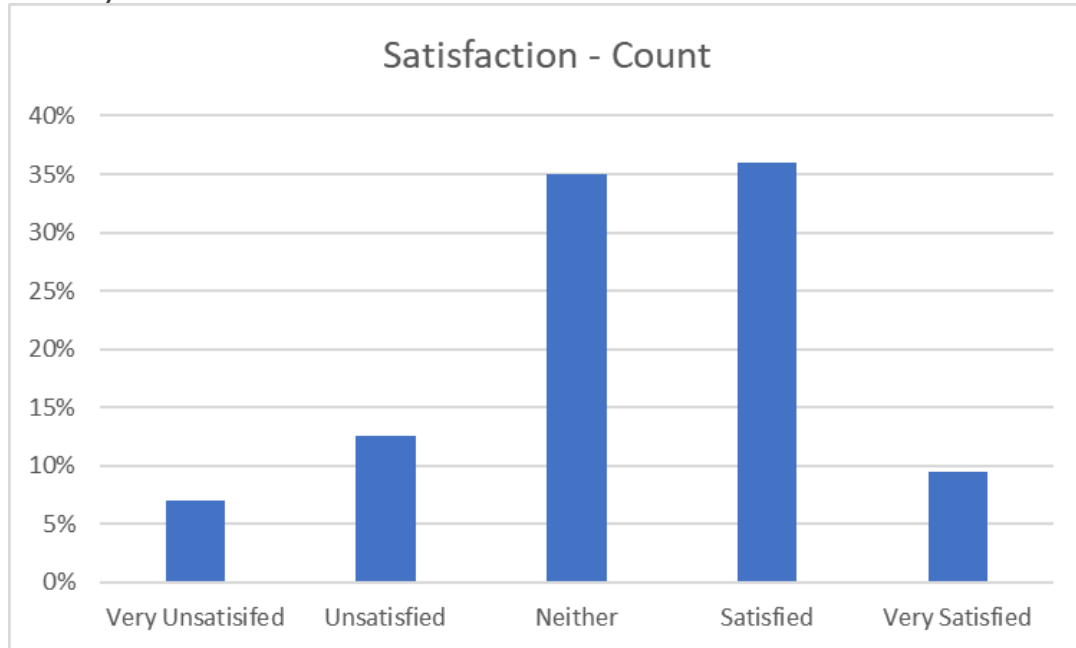
Please rate the extent to which you agree or disagree with the following statements:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not sure
It is hard to afford the lifestyle I want	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, I am satisfied that my income covers my living expenses	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

EDA (Exploratory Data Analysis)

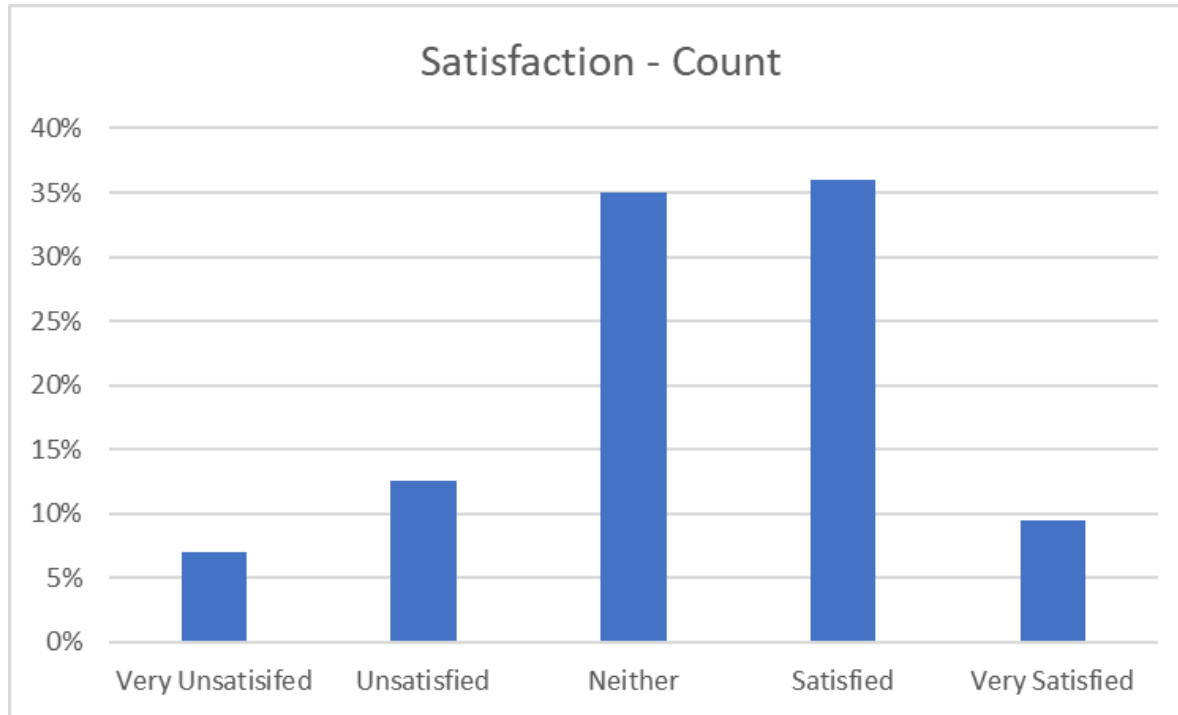
Bar Charts are great for exploratory Data Analysis, and should always be skimmed for problems. Ideally use counts for this, not %, since we want to identify low categories with low sample sizes. Common problems are:

- Poor Discrimination i.e. most people in a single category.
- Missing Data (**which often need to be checked in a different table as software often won't plot it**).



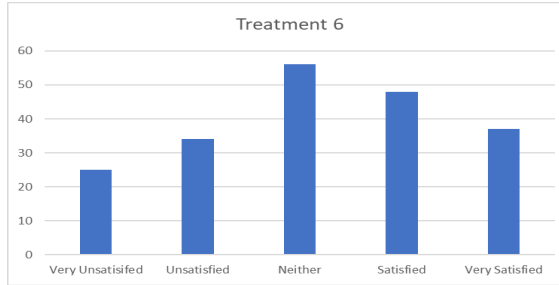
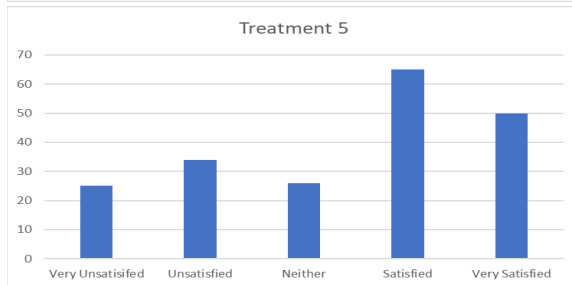
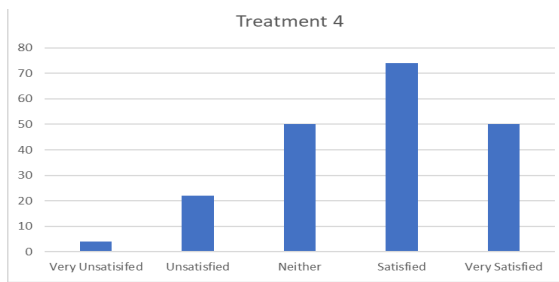
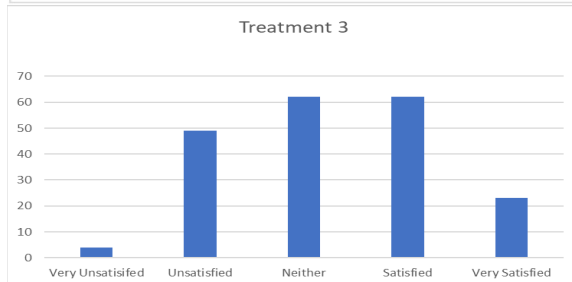
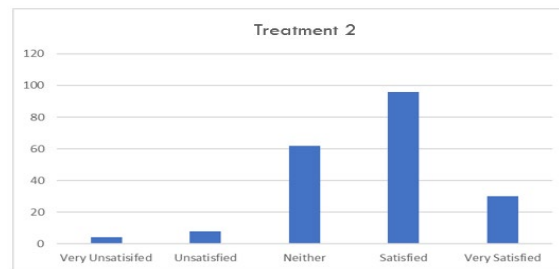
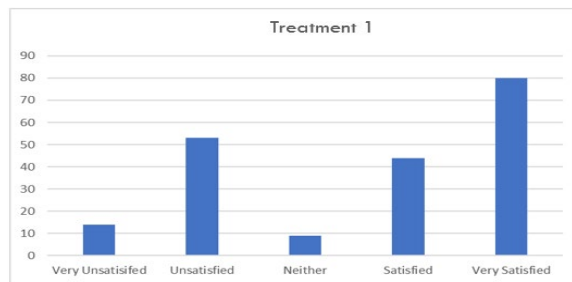
Reporting: Bar Charts of entire scale

Bar Charts can also be used for reporting.



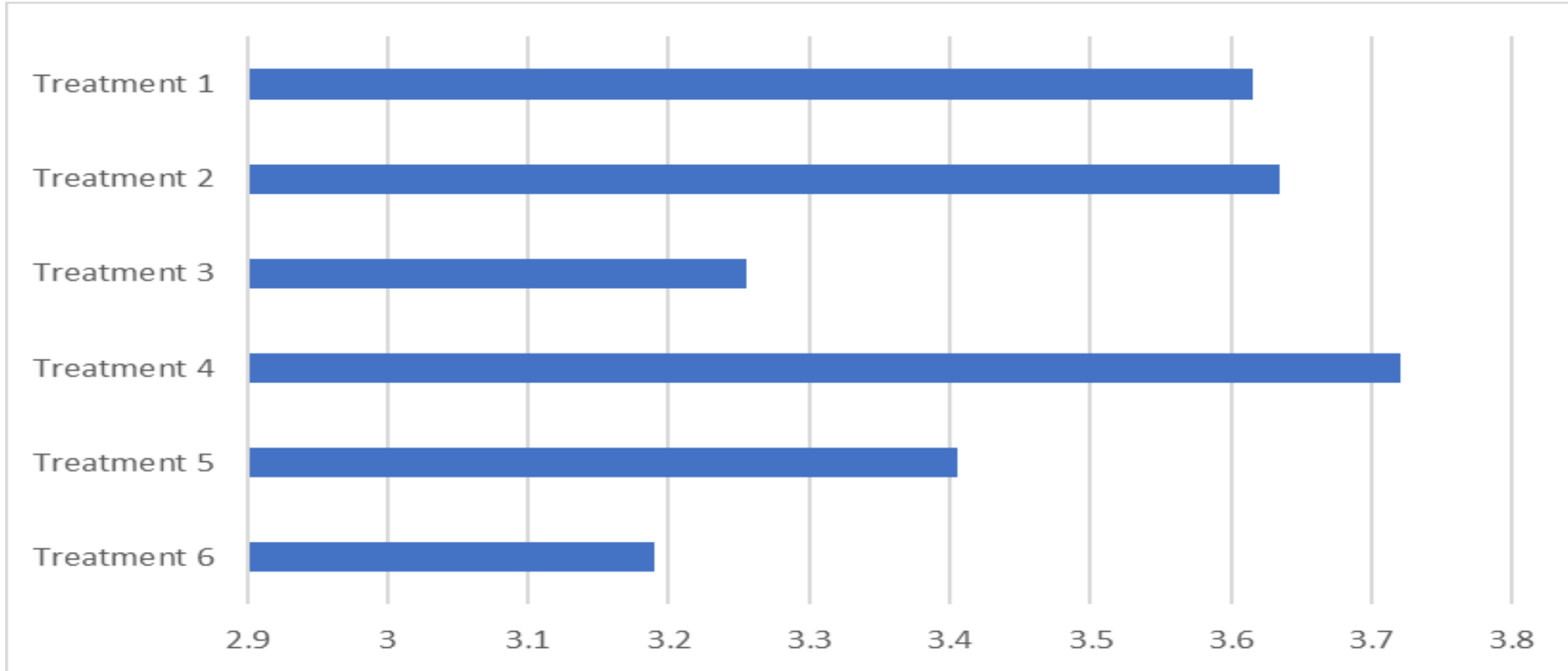
Reporting: Bar Charts of entire scale

But not so good when reporting say 5+ treatments, statements, products, etc!



Reporting: using Averages

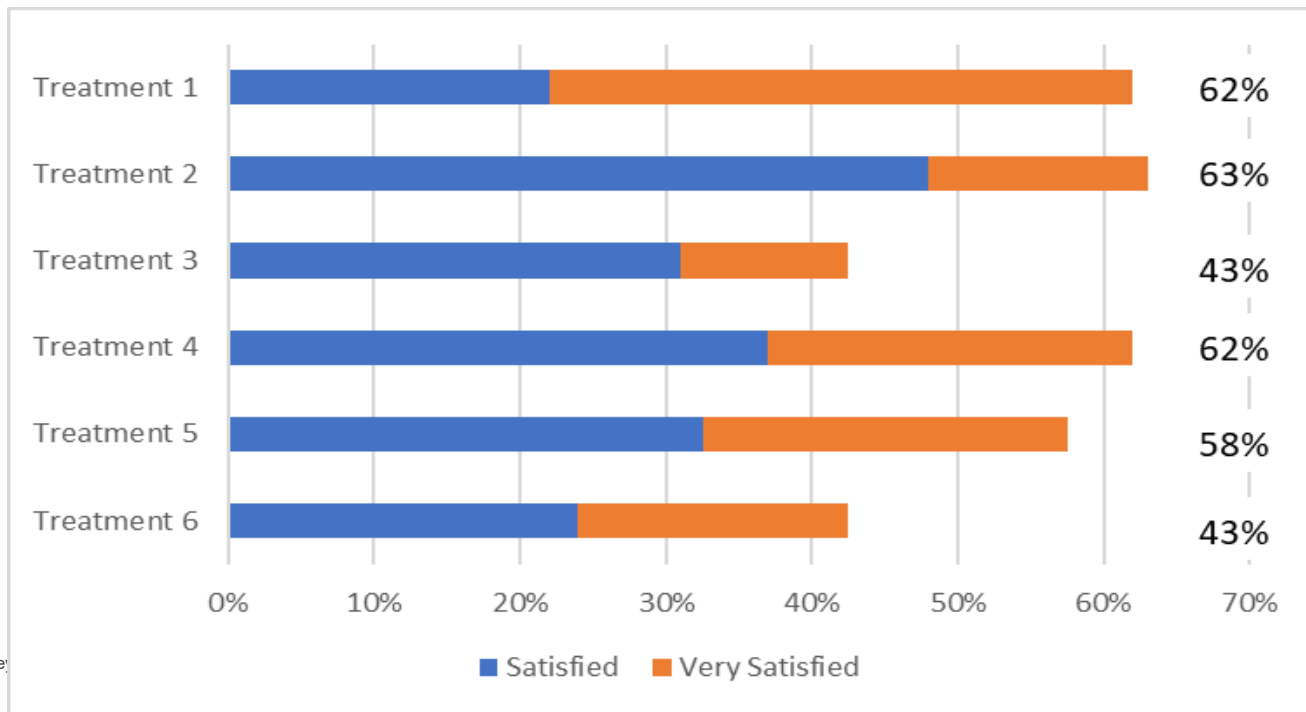
Averages allow you to report lots of items in a succinct way.



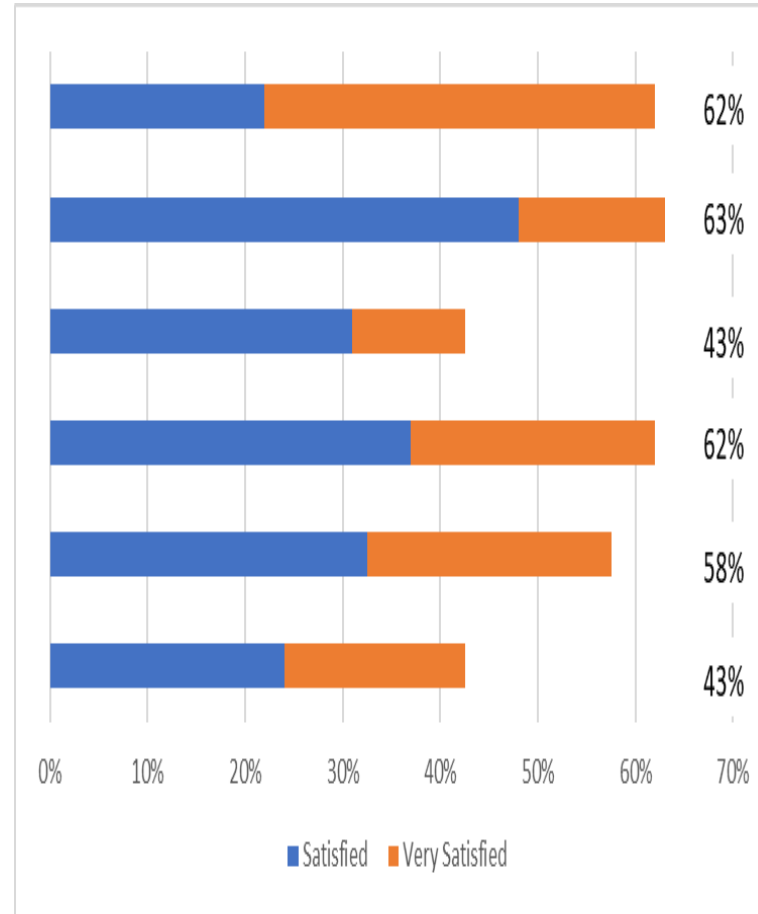
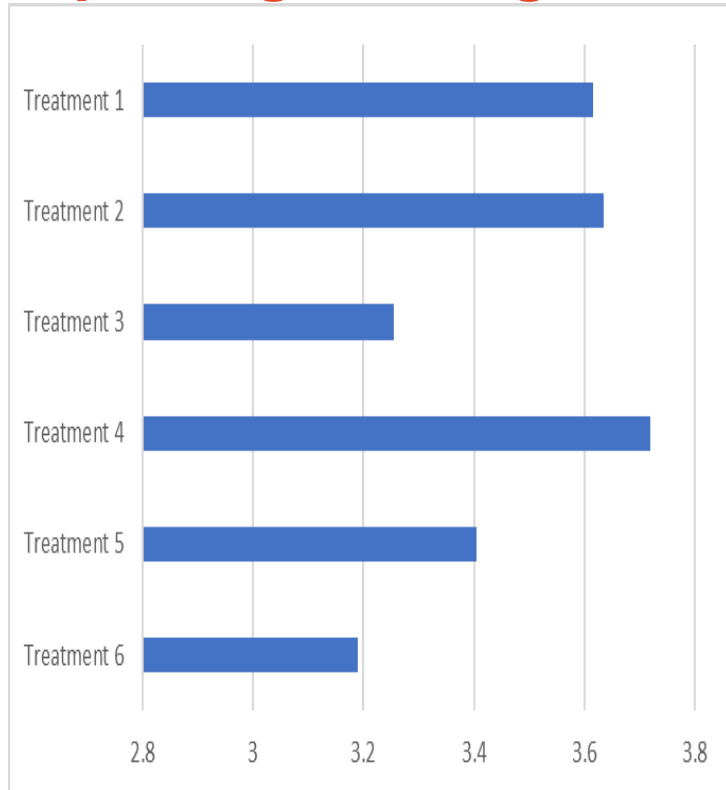
Reporting: using Top Box

Top Box also allows you to report lots of items in a succinct way.

Top Box simply reports the % who picked the top few boxes and can be interpreted as “% who agree”.



Reporting: Averages vs Top Box



Reporting: Individual scores vs Top Box

Property	Average	Top Box
Discrimination	Poorer. Treatment 1 vs 2 have same average.	Better. Treatment 1 vs 2 have different “Very Satisfied”.
Intuitive meaning, which feeds into how easy it is to create an interesting story	No, there is little difference between Treatment 3 and 5 (3.3 vs 3.4).	Yes. 43% vs 58% are satisfied is a much more interesting story.
Can distinguish Polarising Views	No.	Sometimes.

Converting Likert Scales to Top Box

Basically, we want to have 1's to be the thing of interest, and 0's otherwise. Once we have a column of 0's and 1's, we can simply take the average to get the % of times we see a 1 (the thing of interest).

Its safest to work on the labels (Agree, Strongly Agree, etc) rather than the underlying numeric code frame as less can go wrong.

- If using the numeric code frame one needs to be very careful that we don't overwrite the wrong things. Assuming we want Strongly Disagree (1), Disagree (2) and Neither (3) to be 0 and Agree (4), Strongly Agree (5) to be 1 then:
 - Change 1 to 0, and then 2 to 0 and then 3 to 0. Then we change 4 and 5 to 1 in the same way.
 - What we can't do is start in the other direction. Since if we changed 4 and 5 to 1 we would now have the original 1's, and the 4's and 5's as 1's! So we wouldn't be able to change the original 1 to 0.

There are a number of ways for doing this:

- EXCEL
 - Method 1) Find/replace
 - Method 2) use the formula =if(cell = "agree", 1, 0)
- Programming in R, SPSS, etc: Use an elseif function.
- SPSS also has a menu driven Recode into different variables option

Analysis: 2 main options

Strongly Disagree	Disagree	Moderately Disagree	Mildly Disagree	Undecided	Mildly Agree	Moderately Agree	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

Actual Likert Scores

- Continuous vs ordinal
 - **A lot** of debate about this which is very domain specific. Some say it's OK to treat the data as continuous and use normal linear regression. Others that this is a cardinal sin and one must use ordinal regression.
 - Find out what is acceptable in your domain and the journal you want to publish in.
 - If we treat the data as **continuous** we use linear regression.
 - If we treat it as **ordinal** we try to use ordinal regression or if the proportional assumption fails we use logistic or multinomial.
- Reporting as a mean (if continuous) or as counts of the categories or Top Box (if categorical).
 - Again, some say it's OK to treat as continuous and report as a mean, and others not.

Top Box

- Use Logistic regression (Binomial General Linear Model, refer to Linear Models 2 workshop for workflow).
- One benefit is that it avoids the continuous vs ordinal debate.

Normal linear regression vs ordinal regression

The case for Normal linear regression (and means)

I've been working in Market Research since 2012, across North America, Europe and Asia. Market Research works and everyone I have ever worked with does it like this.

Likely works since most people treat the labels as if they were equally spaced.

BUT If you do choose to do it this way consider that it:

- Makes the underlying assumption that the Likert categories are positioned on a roughly linear equidistant scale. If this isn't the case then it won't work.
- Won't work very well with less than 5 Likert categories.
- Assumes linearity. So **always** plot the data to see if the relationship is approximately linear. It may not be!
- Scatter plots won't work without a jitter.
- Set up in the survey instrument so 1 is bad and higher is good. So when you analyse a high score and average is 'good'.
- Don't show the numbers to respondents.

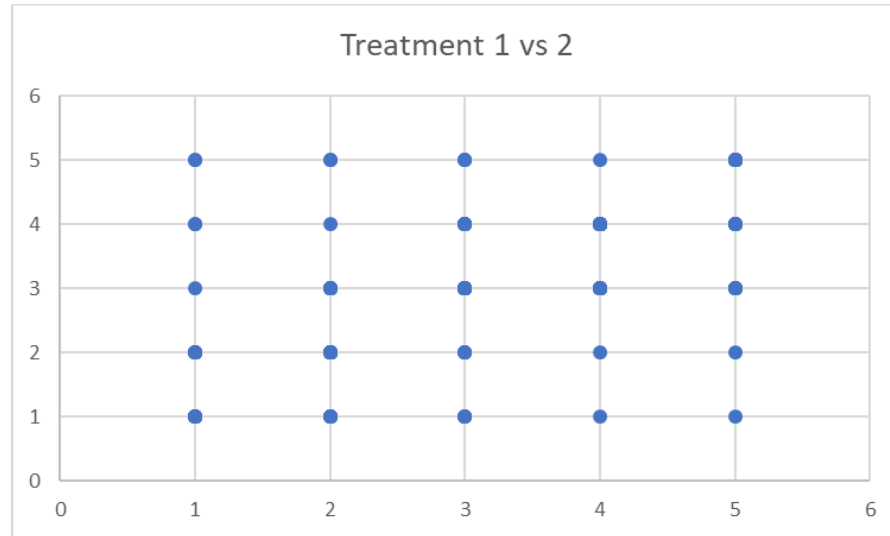
Refer to our Linear Model workshops for regression workflows.

Always look at the scatterplot to test linearity assumption

PROBLEM: Anyone want to guess?

As the number of scores are limited it often comes out as a grid! Which doesn't help us much since we don't know how many times each combination actually occurs

Treatment A
Satisfaction



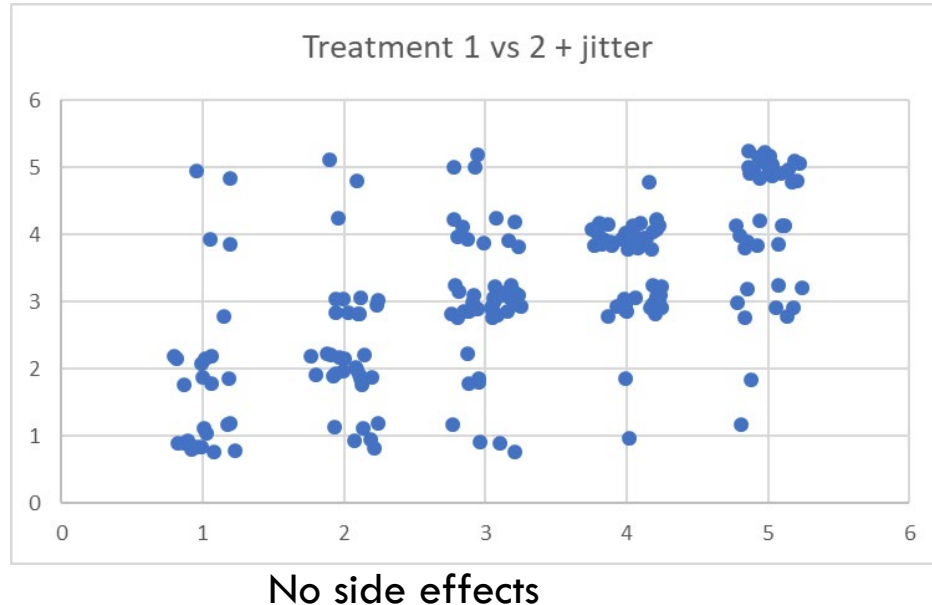
No side effects

Always look at the scatterplot to test linearity assumption

SOLUTION: add a jitter, to the chart only not analysis (jitter = a little bit of randomness).

As roughly linear we can use linear regression.

Treatment A
Satisfaction

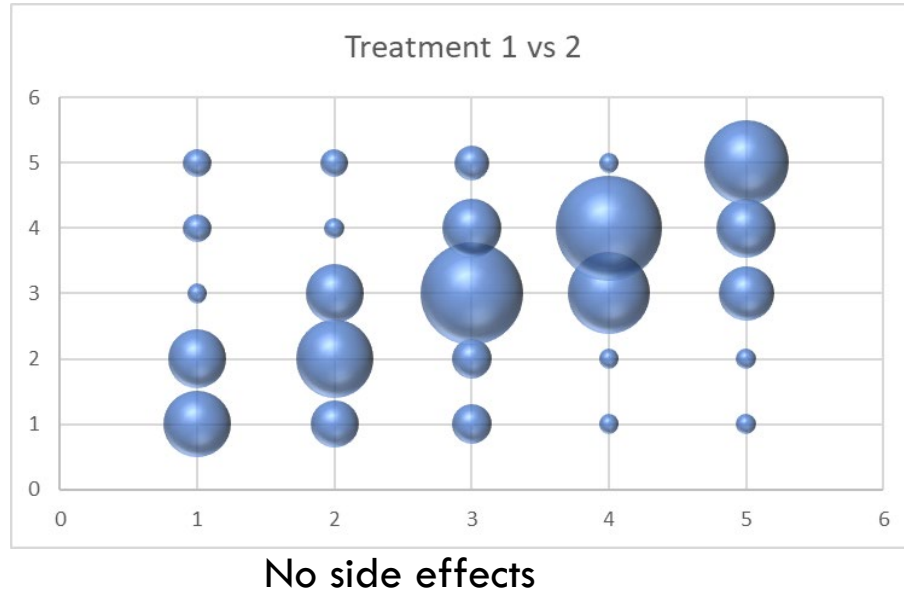


Always look at the scatterplot to test linearity assumption

SOLUTION: use a bubble plot

As roughly linear we can use linear regression.

Treatment A
Satisfaction



Normal linear regression vs ordinal regression

The case for Ordinal Regression

Some researchers are uncomfortable making the assumption that Likert scales are positioned on a linear equidistant scale. If so, then use ordinal regression.

If you do choose to do it this way look out for:

- Underlying assumptions of ordinal regression
 - **Proportional Odds** - i.e. that each independent variable has an identical effect at each cumulative split of the ordinal dependent variable. Or in other words if separate binary logistic regressions were fit to the different categories with the same explanatory variable, they would all have the same odds ratio.
 - So, in a way, similar to a linearity assumption (with some important differences). And still needs to be checked!
 - <http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod5/3/index.html>.
- Many people find them much harder to interpret.

Expert Tip - Low vs High Raters

Some people tend to use the lower part of the scale, others the higher part. This is known as the High vs Low rater effect.

It's not usually worth reporting and can cause problems when analysing. For example:

- If segmenting it can dominate the segmentation, which isn't very interesting.
- If doing Factor analysis is often the first factor.

Solution?

- Standardise each respondent to have zero average (row standardisation, not the more common column/variable standardisation). This means you are now looking at which things were over or under 'benchmark' for each person.
- However, be careful since sometimes this is relevant e.g. satisfaction surveys, where it may be useful to know some people aren't satisfied and others are.

Expert Tip - Low vs High Raters may be confounded with Cultural effects

Different Cultures use Likert differently. Similar to the High Low rater effect some will use the higher part of the scale more than the lower. This is often associated with manners or not wanting to get someone in trouble.

Makes is hard to directly compare Likert scores between cultures or suggest there is a cultural effect.

One solution is to use the same row/respondent standardisation used to remove the High vs Low rater effect. However, be careful since that removes the absolute differences between countries, which may be real and not entirely cultural.

- A better solution is to use a 'metric free' instrument such as ranking, picking all that apply, and Best Worst/Conjoint methods that require a forced trade off (covered in Surveys 2).

Learning Outcomes

Likert scales ask the level of agreement/disagreement to a question.

- Asked as discrete but may be treated as ***ordinal or continuous for analysis depending on your domain.***
- Many questions can be shown as a battery or matrix – all positive or negative (no reverse coding!).
- Avoid double barrel questions.

EDA – bar plots!

- Look for poor discrimination.
- Consider showing averages or top box for many statements.

Analysis:

- **Continuous (linear regression).** Not used in some domains (check). **See Linear Models 1 workshop for more on linear regression.**
- **Ordinal (ordinal regression).** Make sure to check proportional odds assumption.
- **Top box (agree/not agree – logistic regression).** Avoids the above, requires recoding into top box. **See Linear Models 2 workshop for more on logistic regression.**

Tricks of the Trade

Uses for ChatGPT, and other Natural Language Processing (NLP) platforms

Think of ChatGPT as a “mediocre research assistant”. You can give it tasks, but you’ll need to review and often amend them. Some uses are:

- Back coding verbatims e.g. into Positive vs Negative for sentiment analysis (as discussed in the Free Text section).
- Creating a first draft for code frames e.g. “please give me a list of common musical instruments”. You will need to review and usually amend.

WARNING. ChatGPT and other public NLP may own the data and ideas you give them. They may use this to train and help answer other people’s queries. So be careful. You may have just put your hard-earned IP and research into the public domain for all to see!!

This is a constantly evolving field so please refer to university guidelines.

For instance, the Uni may have a private NLP you can use that keeps your data confidential.

Opt Outs: are they needed?

Allowing people to not answer a question or Opt out effectively makes them optional and adds missingness. Which when:

- reporting individual questions may not be a problem.
- but if analysing data when all predictors need to have data it can lead to drastic drop in sample e.g. regression, ANOVA, etc.

For this reason, I do not encourage Opt out's, since it can dramatically reduce sample size.

People often try to fix this by imputing the missing data. Unfortunately, what this really means is that "I'm going to wave my magical statistical wand about and get some guestimates that I hope work OK". So it's best to avoid this option!

Some say a well designed study with the correct screener (target sample), questions and scales shouldn't need opt out's. Particularly if they have been formally validated via a pilot. I tend to lean this way, however there will be exceptions! Such as forced trade off style scales such as Choice Models, since some of the scenarios may not include an option they would pick and we want to know that.

Others claim mistakes happen so it's good to always include one otherwise we are forcing answers from respondents who either don't know or who the question doesn't apply to which corrupts the data.

Rather than an opt out it may be better to include explicit opt out options such as:

- Prefer not to say
- Doesn't apply to me
- Don't know
- Not Applicable

And remember that for Likert and other ordinal rating scales the "neither" option is a type of Opt Out. And is better since then they aren't missing data.

Opt Outs: can be a sign better survey design is needed

- If you think Opt Outs are needed this may indicate a more sophisticated survey design is required.
- If some questions don't apply to some people then don't ask them! Use pipes/logic to tailor the survey.
- For example. Say your goal was to understand how important animal welfare is on people's purchase decision. To do this you directly ask people how important a list of things are when buying various categories from the supermarket on a 5 point Likert scale.
- How fresh something is would be relevant to vegetables, meat and seafood. But not talcum powder!
- Including a "not applicable" is one way of dealing with this. But it annoys respondents when you ask them things that aren't applicable, and can degrade the rest of the data they give you.
- So instead. Use survey pipes/logic to only ask the fresh question to categories where it applies.

Sensitive Q's

For example: asking about drug use or stealing at work.

Mode has an effect e.g. online or paper is often more accurate than F2F or CATI since people don't want to admit to nefarious behaviour to another person.

Ideally the survey will be deidentified, and ensure they know this.

There are also methodological methods

- Item Count Method
 - Ask **how many** (not which) things people have done from a list of slightly dodgy answers. $X = \% \text{ done}$
 - Ask same list + the dodgy activity you want to measure. $Y = \% \text{ done}$
 - $Y - X = \% \text{ who are doing the dodgy thing}$
 - Since they never actually select the dodgy option they are more likely to answer truthfully.

Categorical: Ancestry

Australia ABS

The ancestry variables provide a self-assessed measure of ethnicity and cultural background, which, when used in conjunction with the person's and their parents' countries of birth provides a good indication of the ethnic background of first and second generation Australians. ***Ancestry in the Australian context is complex as there are many Australians with origins and heritage that do not, in practice, relate to their current ethnic identity.*** When ancestry data is used alone, it should only be done to represent a broad measure of cultural diversity. ***Ancestry is particularly useful to identify distinct ethnic or cultural groups within Australia such as Māori's or Australian South Sea Islanders, and groups which are spread across countries such as Kurds.*** Surrogate measures of ethnicity such as country of birth or languages other than English spoken at home, alone cannot identify these groups. ***This information is useful in developing policies which reflect the needs of our society and for the effective delivery of services to particular ethnic communities.***

Measurement issues

The ancestry question records all claims of association with ancestries, ethnic origins and cultures. Whilst some people may respond according to how they may identify with a particular cultural group (subjectively), the intent of the question is to capture the cultural context in which they were raised (objectively). Multiple responses are encouraged. Responses to the ancestry question are coded to the ASCCEG. The classification is not intended to classify people, but rather all claims of association with an ethnic origin or cultural group, i.e. one ancestry response is not equal to one person. ***Many people do not relate to a single ethnic origin or cultural group and will give multiple responses to a question on ancestry, ethnicity or cultural identity.*** The ABS has developed guidelines for the coding, storage and presentation of multiple responses to questions on ancestry, ethnicity or cultural identity data. These guidelines are included in the ASCCEG publication.

<https://www.abs.gov.au/census/guide-census-data/census-dictionary/2021/variables-topic/cultural-diversity/ancestry-1st-response-anc1p>

Categorical: Ancestry

Canada

The Canadian census determines generation through 3 questions:

- Birth place of the respondent
- Birth place of mother
- Birth place of father
-

Canadian Definitions:

- First generation: individual was born outside of Canada
- Second generation: individual born in Canada, and at least one parent was born outside of Canada
- Third generation: individual born in Canada, and both parents born in Canada

https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-010-x/99-010-x2011003_2-eng.cfm

Continuous variables like Age should rarely use interval brackets and instead get the actual value

Interval Brackets are really categorical scales. Examples:

- How old are you?
 - Less than 18 years old
 - 18-21
 - Older than 21 years

Problems with interval brackets

- May not match other data sets you want to merge e.g. census
- Restricts your analysis to discrete when you may prefer to look for numeric correlations or use regression.
- Can convert to brackets based on data, but can't go back the other way e.g. may see a drop drinking at age 25 so you may split into 0-18, 18-25, 25+. Not possible if you used 21-30 age bracket.

Common items that use interval brackets are Age, Income.

Alternatives for

- Age
 - Actual Age
 - DOB and auto calculate age

Exceptions

- Income
 - Ensure you use the same or smaller intervals than data you want to merge with e.g. census

The problem with asking the actual amount: Data Entry Mistakes

Asking for the actual amount can lead to Data entry Mistakes e.g. someone accidentally enters their income with an extra 0 i.e. \$100,000 instead of \$10,000.

1 way around this is to use a very fine ordinal categorical scale

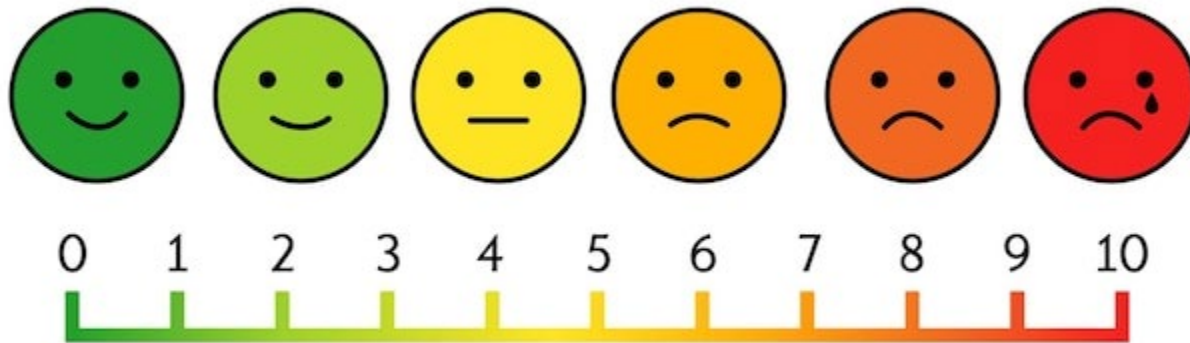
Don't randomise the order for these!!

- | | | |
|--------------------------|---|--------------------------------|
| <input type="checkbox"/> | 1 | Less than \$20,000 per year |
| <input type="checkbox"/> | 2 | \$20,001 to \$30,000 per year |
| <input type="checkbox"/> | 3 | \$30,001 to \$40,000 per year |
| <input type="checkbox"/> | 4 | \$40,001 to \$50,000 per year |
| <input type="checkbox"/> | 5 | \$50,001 to \$60,000 per year |
| <input type="checkbox"/> | 6 | \$60,001 to \$70,000 per year |
| <input type="checkbox"/> | 7 | \$70,001 to \$80,000 per year |
| <input type="checkbox"/> | 8 | \$80,001 to \$90,000 per year |
| <input type="checkbox"/> | 9 | \$90,001 to \$100,000 per year |

Kids

Some people say not to use smiley faces since they may respond to the face that most closely matches their mood rather than the answer to the question.

So, in general, don't use.



How to bias the answers – frame them to guide a response

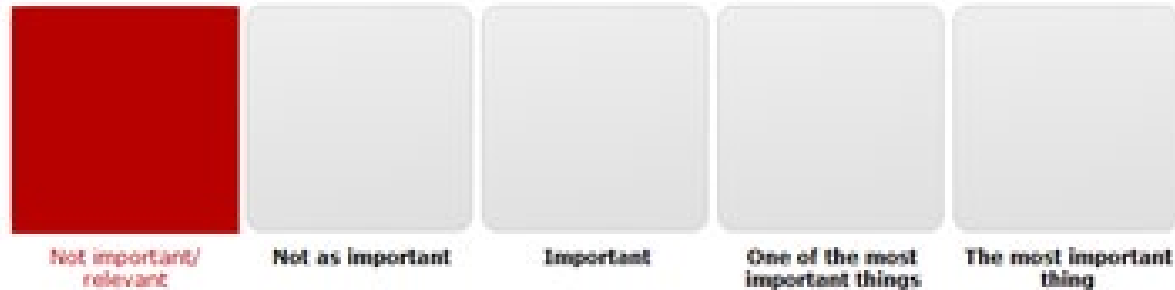
Q) Who do you think should have more influence on important policy decisions [pick all that apply]

- The people instead of politicians should make our most important policy decisions.
- Politicians should have little influence as they make decisions that harm the interests of ordinary people and can't be trusted.
- The people, as politicians are corrupt and can't be trusted to make important decisions.
- The ordinary people should have more influence than big companies that only want to make profits.

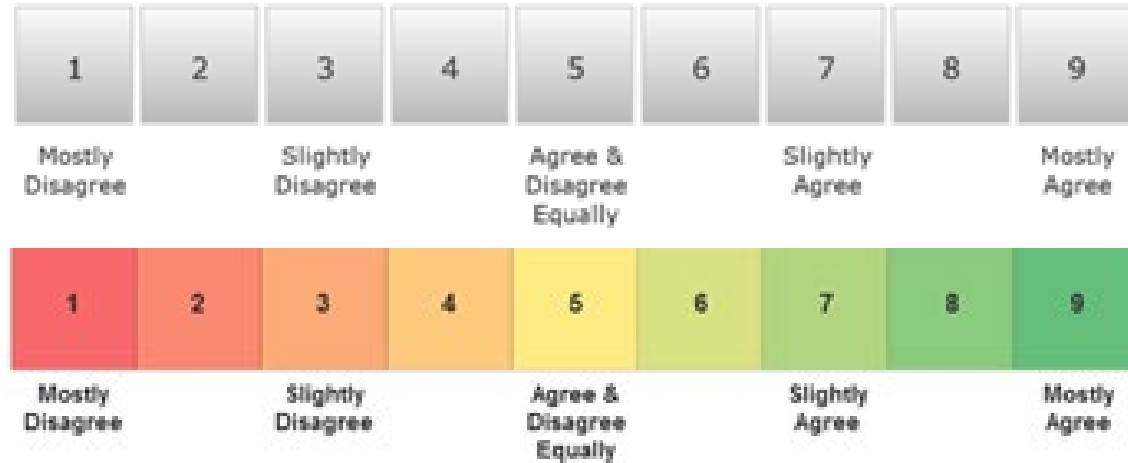
**Don't forget to
randomise their order**

Don't use Colour like this

- It *may* turn a 5 point scale into a 4 point one.



Colour can increase differentiation and cause bias



- This is a real-life example. I had a client with a multi year tracking study change a scale from the B&W scale to the colour one.
 - The new colour one had much more differentiation. Which was a serious problem since it introduced a jump in the data which had nothing to do with reality.
- To fix it they had to pay me to develop a correction factor, so they could still track through time.
- Highlights an important concept. **Don't change tracking studies.**

Learning Outcomes

ChatGPT and other platforms can be useful.

- **Back coding free text** into a categorical variable e.g. for sentiment analysis.
- Can give you **lists of categories, code frames, etc.**
- ***Don't put your IP online!***

Opt outs – are they needed or does your survey design need a rethink?

- Missingness can be problematic, particularly with smaller sample sizes.
- Use pipes/logic instead and/or consider explicit opt out options.

Sensitive questions

- **Consider mode** (face to face, etc. might not be a good choice) and **anonymity or at least deidentifying**.
- Can sometimes be indirectly estimated.

Asking **ancestry is tricky – see ABS for latest as recommendations** are regularly updated.

Asking continuous variables

- **Get the actual value if you can, intervals can be hard to match to other studies.**
- If you need to use them to avoid data entry mistakes (e.g. \$), ask as a fine ordinal categorical scale.

Biasing answers – don't frame your questions to guide a response!

- Beware of colour as it affects selection – can be different for different cultures too.
- Kids – avoid smiley faces, etc.

Survey Platforms



The University provides access to **REDCap**, **Qualtrics** and **MS Forms**. These are the preferred platforms, using others may cause researchers to not meet their legal obligations on criteria such as data security and respondent confidentiality. For more info on suitable survey research platforms please review: https://sydneyuni.service-now.com/sm?id=kb_article_view&sysparm_article=KB0019511.



Please contact **Research Data Consulting** for help with **REDCap**.

<https://REDCap.sydney.edu.au/surveys/?s=3W48H9833H>

Further assistance at The University of Sydney



SIH

- [Statistical Resources](#) website: containing our workshop slides and our favourite external resources (including links for learning R and SPSS).
- [Hacky Hour](#): an informal monthly meetup for getting help with coding or using statistics software.
- 1on1 Consults can be requested on our website 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 [Training calendar](#).
- Sign up to our mailing list to be notified of upcoming training.

Other

- Open Learning Environment (OLE) courses
- LinkedIn Learning

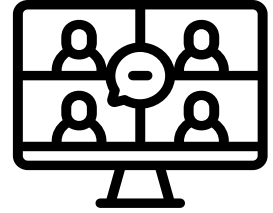
Further assistance



REDCap & QUALTRICS TEMPLATES

- Have a lot of validated survey instrument templates on file for your use. Worth looking there before you set up your own!

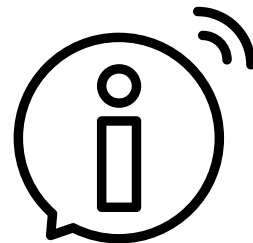
How to use our workshops



- 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 applicable 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](#)



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.