Design and Analysis of Surveys 1: An Introduction

Presented by
Chris Howden
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

THE UNIVERSITY OF SYDNEY

Acknowledging SIH

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

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

Suggested wording:

General acknowledgement:

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

Acknowledging specific staff:

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

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

We value your feedback

- We aim to help HDR students and researchers in a wide range of fields across different faculties
- We want to hear about you and whether this workshop has helped you in your research.

- Later in this workshop there will be a link to a survey
- It only takes a few minutes to complete (*really!*)
- Completing this survey will help us create workshops that best meet the needs of researchers like you

During the workshop

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

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

Challenge Question

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

After the workshop

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

- Todays 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)

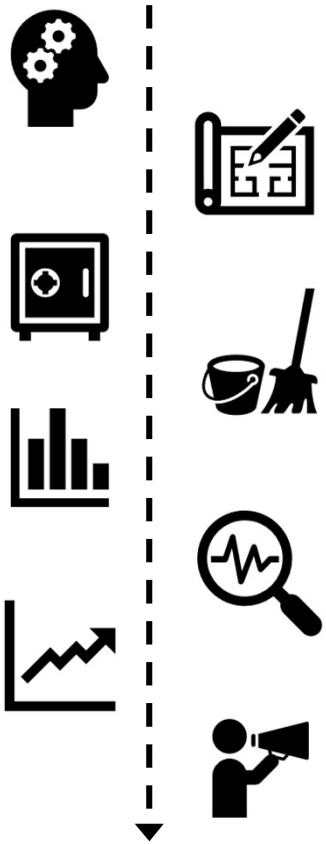
Research Workflows

- 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

A Conversation is better than a Presentation

So please speak up and ask questions!

People think differently. So I may need to explain things in 2 or 3 different ways!

Designing a Survey

THE UNIVERSITY OF SYDNEY

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 people who have got a lucky draw of six $50 grocery vouchers through.
- Keep it short, did I mention that??

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 in to it. Try and relate it back to them e.g. “this market 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

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 Voice Project 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 a draft in word based on your needs, desktop research and qualitative work.
2. Leave it at least a few days, ideally a week, then review. Keep doing until no new edits.
3. Feedback from friends and colleagues.
4. Set up in Survey Tool e.g. REDCap or Qualtrics.
5. Enter some demo data: Just think of some likely but very different respondents and enter as they would.
6. Export data and ensure you have set it up so the data exports in an easy to analyse format.
7. Leave it at least a few days, ideally a week, then review. Keep doing until no new edits.
8. Send link to friends and colleagues and get some feedback.
9. Leave a final week, review and send off to Ethics for approval.
    1. Even small changes can be problematic as new ethics approval is often required, this is one reason it is so important to get it right before submitting to ethics.
10. Go Live!!!
11. Review the first 12-50 respondents for any problems
    - Look for missing categories you should consider adding 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.)

NB: This method assumes a straight forward survey with established questions and scales. So little testing required. Surveys 2 has more detail and advanced methods on Best Practise Survey Design and Validation.

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.

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?

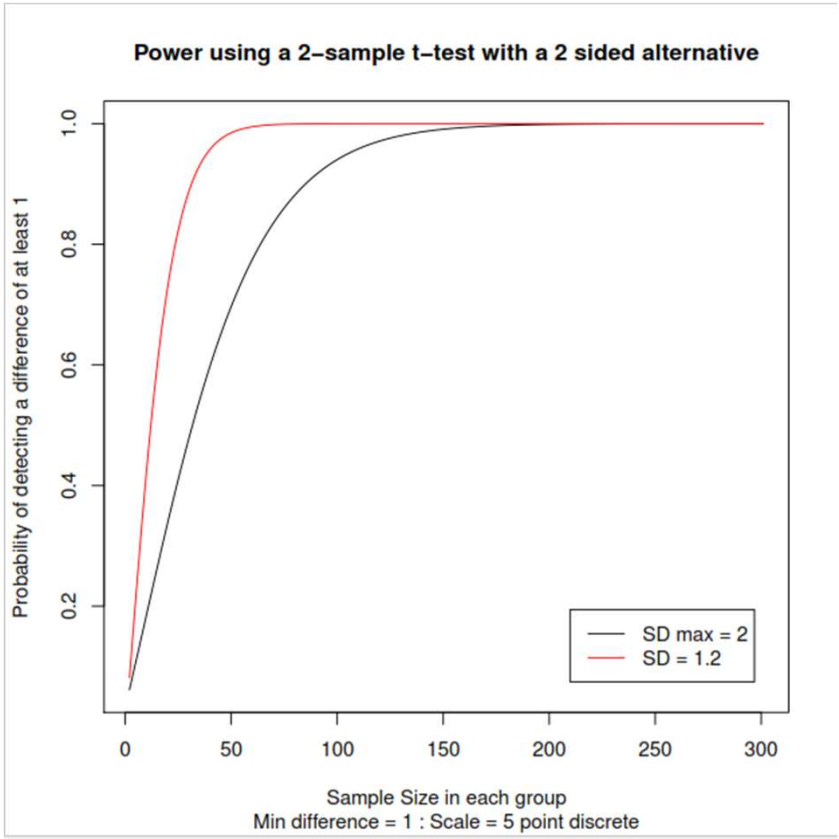
- Gives 'colour' to your report, often worth adding in a few open enders to emphasise a point you're making.
- If you missed something you will notice it here. So keep an eye on this as the survey progresses and adjust if necessary.

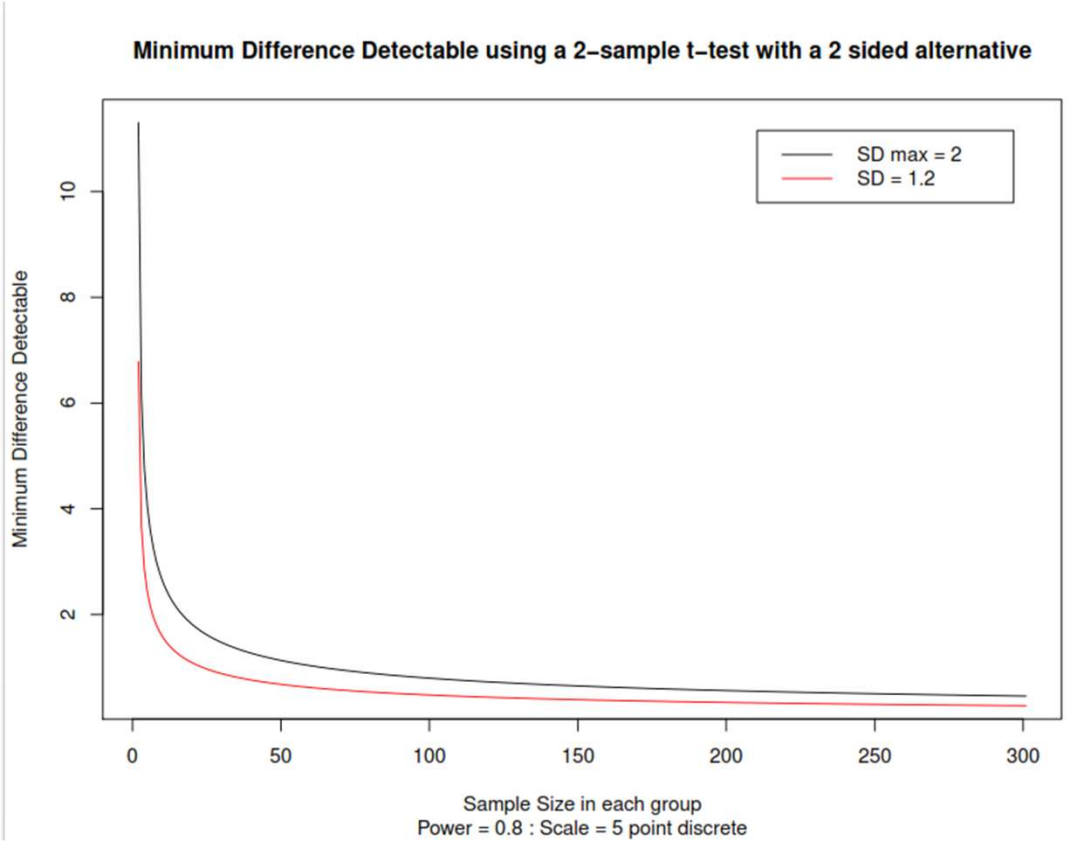
Experimental Design: Power Analysis

Can be hard to do since there are so many different types of scales. That said it can be done and 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.

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 flattens here).

Data Export and Cleaning

THE UNIVERSITY OF SYDNEY

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.
    1. 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.
    2. 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 see different questions based on how they answer previous ones.
3. Now you can export the data as staked 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

Look for these types of respondents and decide if they should be removed.

Racers

- Respondents that finish too quickly

Flatliners

- People who answer all questions the same

Straightliners / Flatlines

- Are often removed
- Easy way to find them is look for variables with SD=0
- Talk about ways to identify if flatlined data is valid vs respondents not caring in the later 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	●	○	○	○	○
Chocolate is the best	●	○	○	○	○
Oxygen is important	●	○	○	○	○
Crime doesn't pay	●	○	○	○	○
I like my friends	●	○	○	○	○
Getting bitten by a shark would be fun	●	○	○	○	○
I dislike my friends	●	○	○	○	○

Common Processing

Transform LIKERT to "Top Box" binary agree/disagree data for easier reporting (covered in more detail relevant section below)

Back code Free Form Text to a categorical variable for easier analysis or sentiment analysis (covered in more detail relevant section below)

Different Types of Survey Questions

Introduction

Design Examples

Data Cleaning

EDA (Exploratory Data Analysis)

Reporting

THE UNIVERSITY OF SYDNEY

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 know that 75% of people agree with something is less useful than knowing the 95% CI is
- 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. Maps, Segmentations, Driver Analysis, etc.
- And then Predictive Modelling, only occasionally done with survey information for example:
  - Preference / Volume Estimation. Often using Choice Models.

Significance testing, colour coding and screening

When **testing hypotheses identified before the study we should always account for multiple comparisons** e.g. Bonferroni, Tukey, Holmes, False Discovery Rate. For more information on correcting for multiple comparisons refer to Linear models iii.

However, surveys often have a lot of questions we calculate p-values for and correcting for multiple comparisons in the normal ways usually means nothing is worthwhile reporting. So **instead of strictly testing hypothesis these p-values are often 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.

One 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

Hypothesis testing vs Screening

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 Exploratory Analysis.

Hypothesis testing

- Requires corrections for Multiple Comparisons.
- 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.

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

- 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.

Significance testing, colour coding and screening

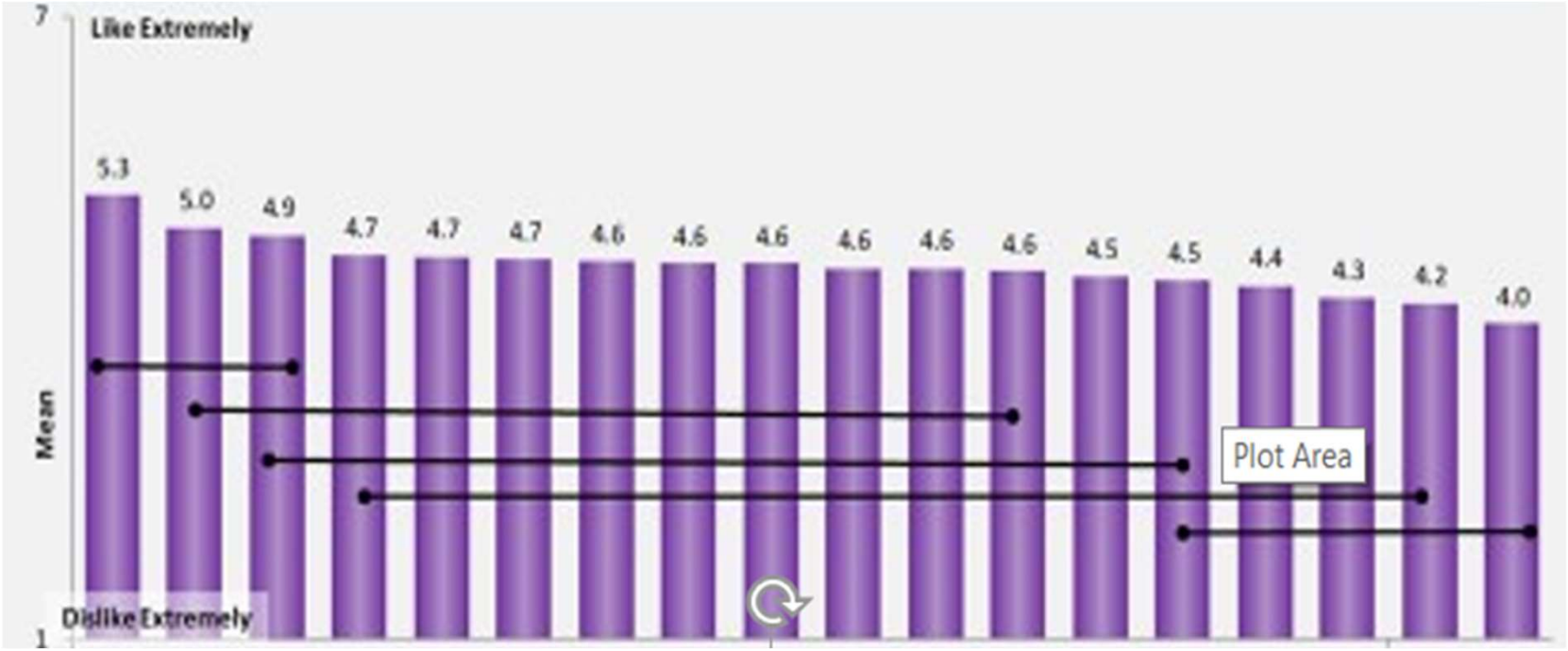
Example 1 - Colour

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 -

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 from a homogenous group i.e. there is no significant difference.
- Duncan's new multiple range test (MRT) is one way to get these.

NB: Forgive the lack of a horizontal axis, this was products from a real world consumer test that were removed to retain commercial confidentiality.

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

Categorical / Discrete Variables

THE UNIVERSITY OF SYDNEY

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.

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.

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 prevents us from 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.

Design: Use existing Code Frames

Employment

- Use the 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: Gender and Gender Identity

The idea that one needs more than male/female for sex, gender identity and sexual preference is a relatively new one. I didn't see it gaining much traction before 2020. However I have been seeing more and more surveys with more than the binary Male/Female options.

There are a number of reasons for this change:

- A wish to make other gender options more acceptable in mainstream society.
- Recognition that 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 preference. So knowing more about the market for male clothing, for female bodies, will make it easier to make clothing for these people.

Design: Gender and Gender Identity

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 important in medical studies where genetics plays an important role.
  3. gender identity
  4. Sexual orientation
- Is it worth the extra questions-remember we want to keep our surveys short! Will the extra information be used, does it deliver useful extra insight?
- Is it worth the potential trauma of delving into peoples private lives. Will not asking cause trauma?

There is no established method yet, there is a lot of discussion and fine tuning ahead.

If one decides to ask a single 'gender' question I would **suggest these 3 options as a minimum.** From a statistical point of view they will ***usually*** capture most people's intended meaning. Meaning the error/variance will be minimal. That said they may not be sufficient for reasons of inclusion and depending on the population being sampled.

1. Female
2. Male
3. Prefer not to say

Personally I would also add these since it covers all possibilities. When analysing one may choice to combine them if there is not sufficient sample to treat them separately.

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

Another common method is to then add a sexual preference question (with an opt out).

Design: Gender and Gender Identity

An active area of research.

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.

Ensure **good discrimination** i.e. more highly spread out the better (i.e. should go from like extremely to dislike extremely).

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

Optimal # of points: the more points on the scale the better, but too many points become ambiguous:

- Bipolar: 5, 7 or 9 points are common
- Uni polar: 5 points is common

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.

BAD EXAMPLES - Not all points are anchored

1. How comfortable were you with your talking?

1 2 3 4 5 6 7

Extremely Comfortable Extremely Uncomfortable

2. How confident were you with your talking?

1 2 3 4 5 6 7

Extremely Confident Extremely Anxious

3. How successful did you feel with your talking?

1 2 3 4 5 6 7

Extremely Successful Extremely Unsuccessful

GOOD EXAMPLES

5 point scale – Unipolar for Importance

How important are surveys to your startup's success?

Not at all Important	Slightly Important	Moderately Important	Very Important	Extremely Important
☐	☐	☐	☐	☐

5 point scale – Bipolar LIKERT scale for Importance

	Not At All Important	Low Importance	Neutral	Important	Very Important
Product Quality	○	○	○	○	○

5 point scale – Bipolar LIKERT scale for Agreement

Strongly Disagree	Disagree	Unsure (or neither)	Agree	Strongly Agree

7 point scale - Bipolar

How satisfied are you with our service?

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

9 point scale - Bipolar

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**

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

Design: Ordinal Scales – Natural metric anchors

Questions with a natural metric

- Frequency, Amount, Spend, Weight, Probability
- Label with numbers not words since its more accurate e.g. frequency would be x 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.
- NB: LIKERT scales are covered in more detail in their own section

Design: Ordinal Scales – Natural metric anchors

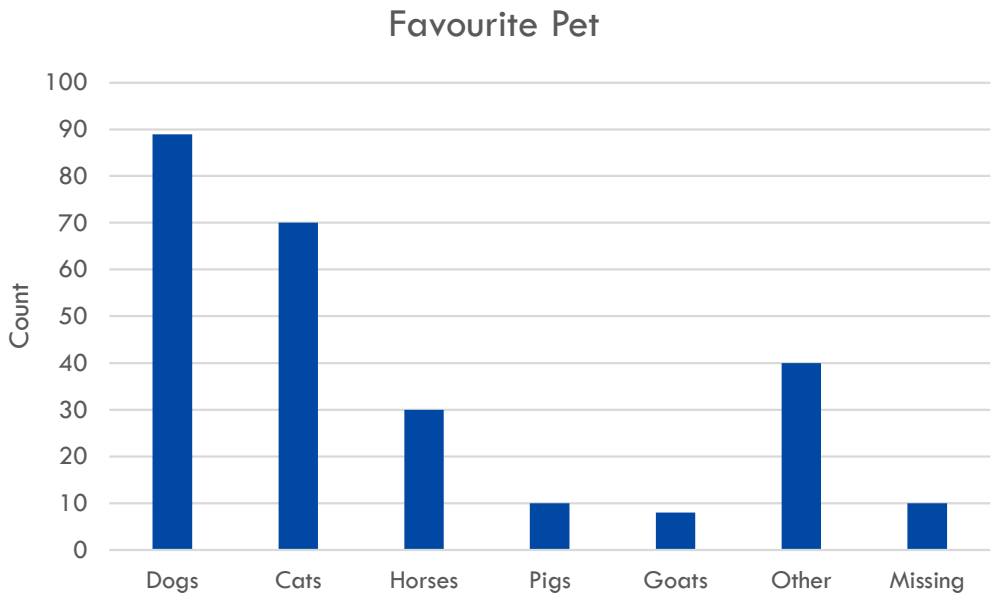
How often do you feel pain when you exercise?

- **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

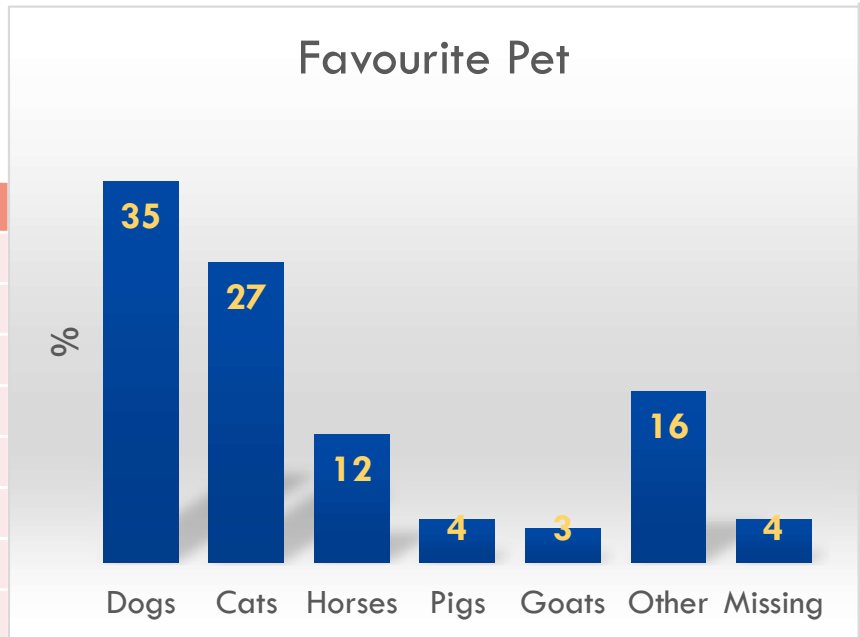


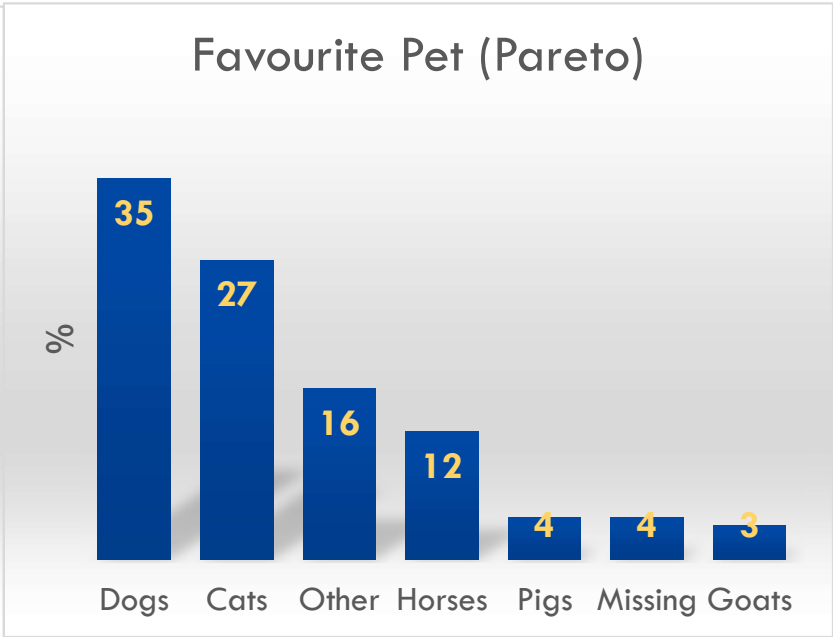
Reporting

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 under stand intuitively.
- Often useful to include the actual scores as labels.
- 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

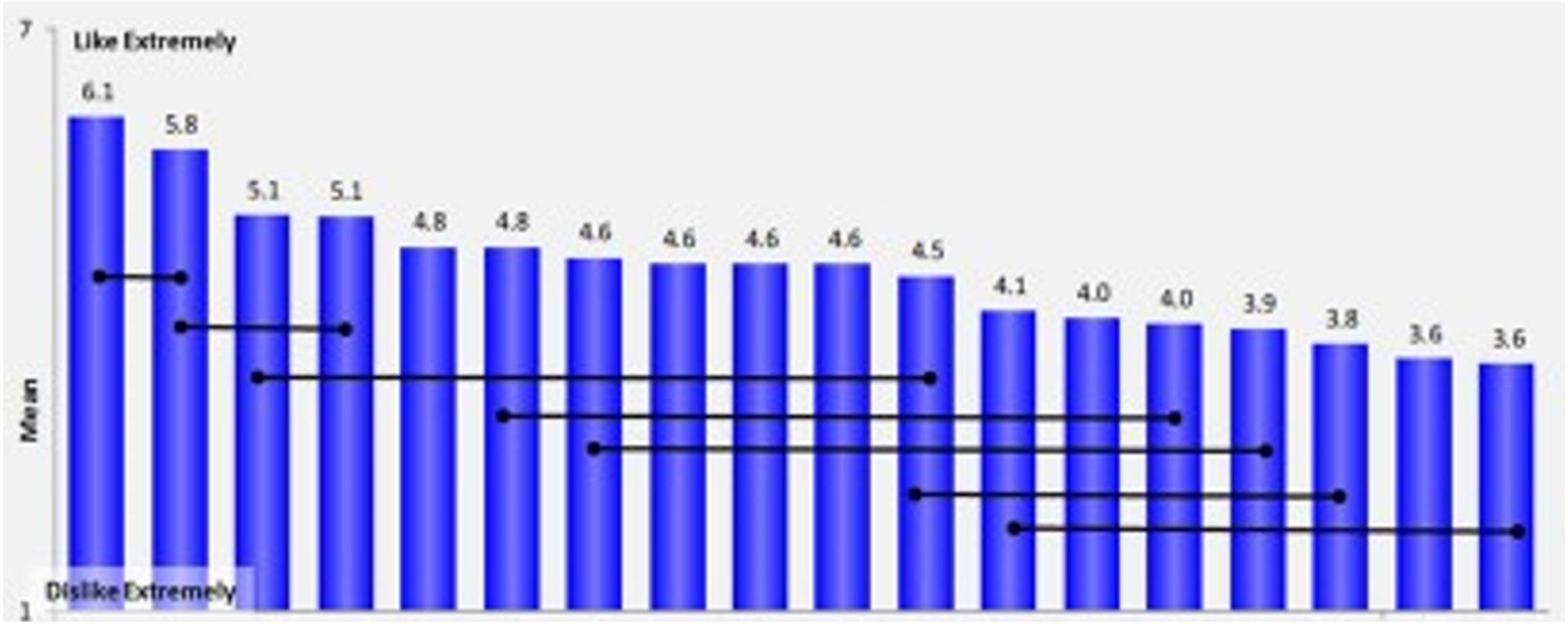
Example 1 - Colour

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 -

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% *

Reporting Splits and comparing to Benchmarks: Homogenous Subsets



Each bar represents liking of a different product. Apologies, I had to remove their labels to debrand it for Commercial in Confidence reasons.

Analysis

Nominal

- Logistic regression
- Poisson Regression
- Multinomial
- Chi Squared Analysis and it's mapping equivalent Correspondence Analysis

Ordinal

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

Free Text / Open Enders

THE UNIVERSITY OF SYDNEY

What are Open Enders/Free Text?

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

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

BUT use with care. If a code frame is more appropriate use them with an other. Not doing so can lead to a lot work post processing the data which may also be 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 them to single topic since this makes them 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.
- A simple code frame for comments which can then be used in "Sentiment analysis":
  - Positive
  - Neutral
  - Negative

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.

Continuous Variables

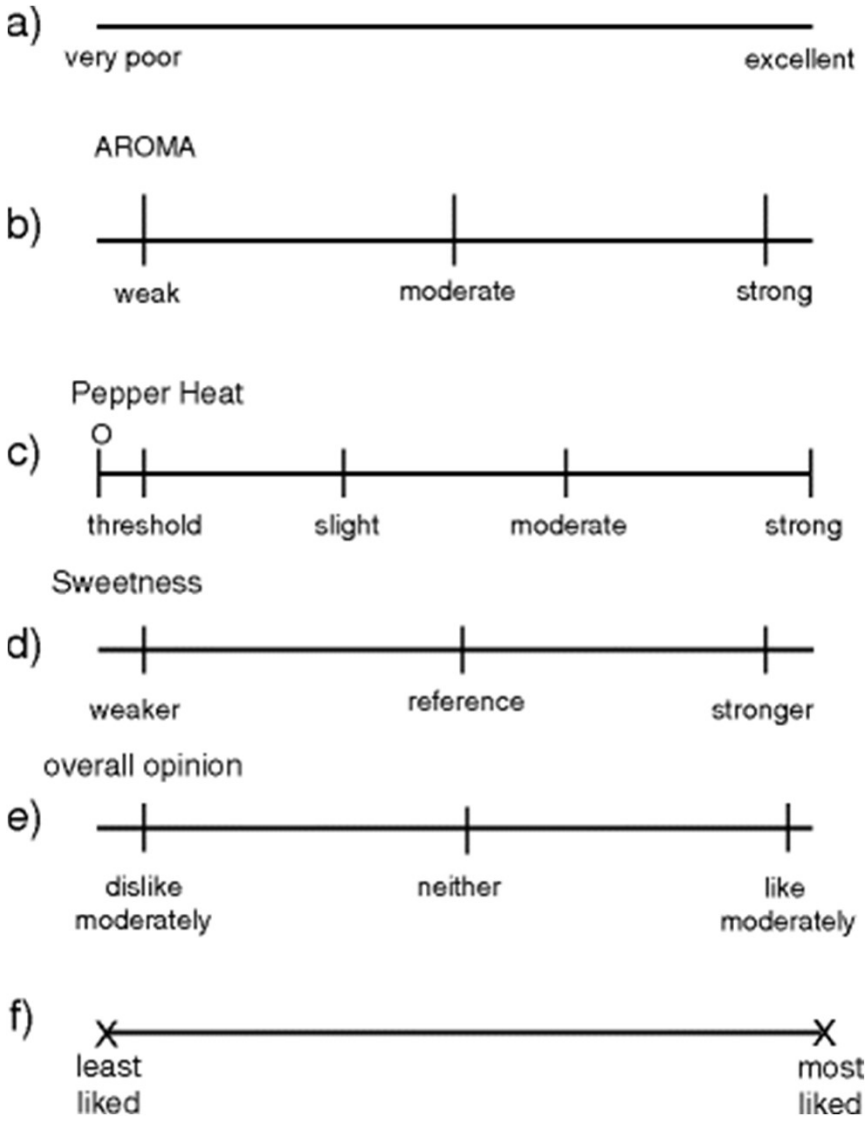
THE UNIVERSITY OF SYDNEY

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)



- **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 at the very least show a half way mark and even a ¼ and ¾ 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 well, but do quite poorly outside of that.

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

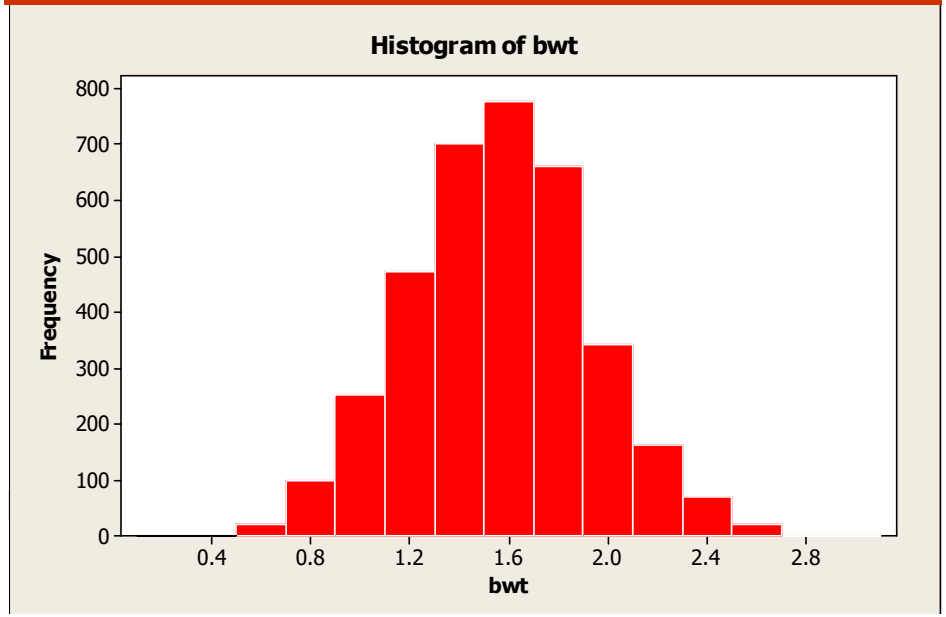
EDA (Exploratory Data Analysis) and data cleaning

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

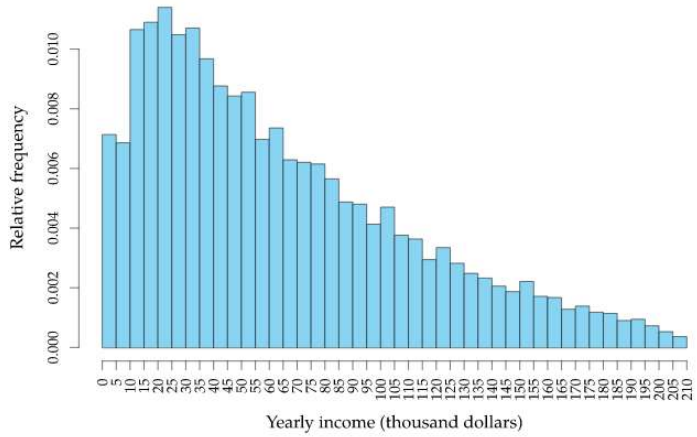
- Outliers (which might be Data Entry mistakes and require removing e.g. someone who says they are 230 years old)
- Unexpected distributions
- If symmetric report using the average, if not and highly skewed decide if the median would be better e.g. house prices, income.

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.

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 Models III workshop). Also refer to the LIKERT section below for some examples.

LIKERT Scales

THE UNIVERSITY OF SYDNEY

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.

They fundamentally 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.

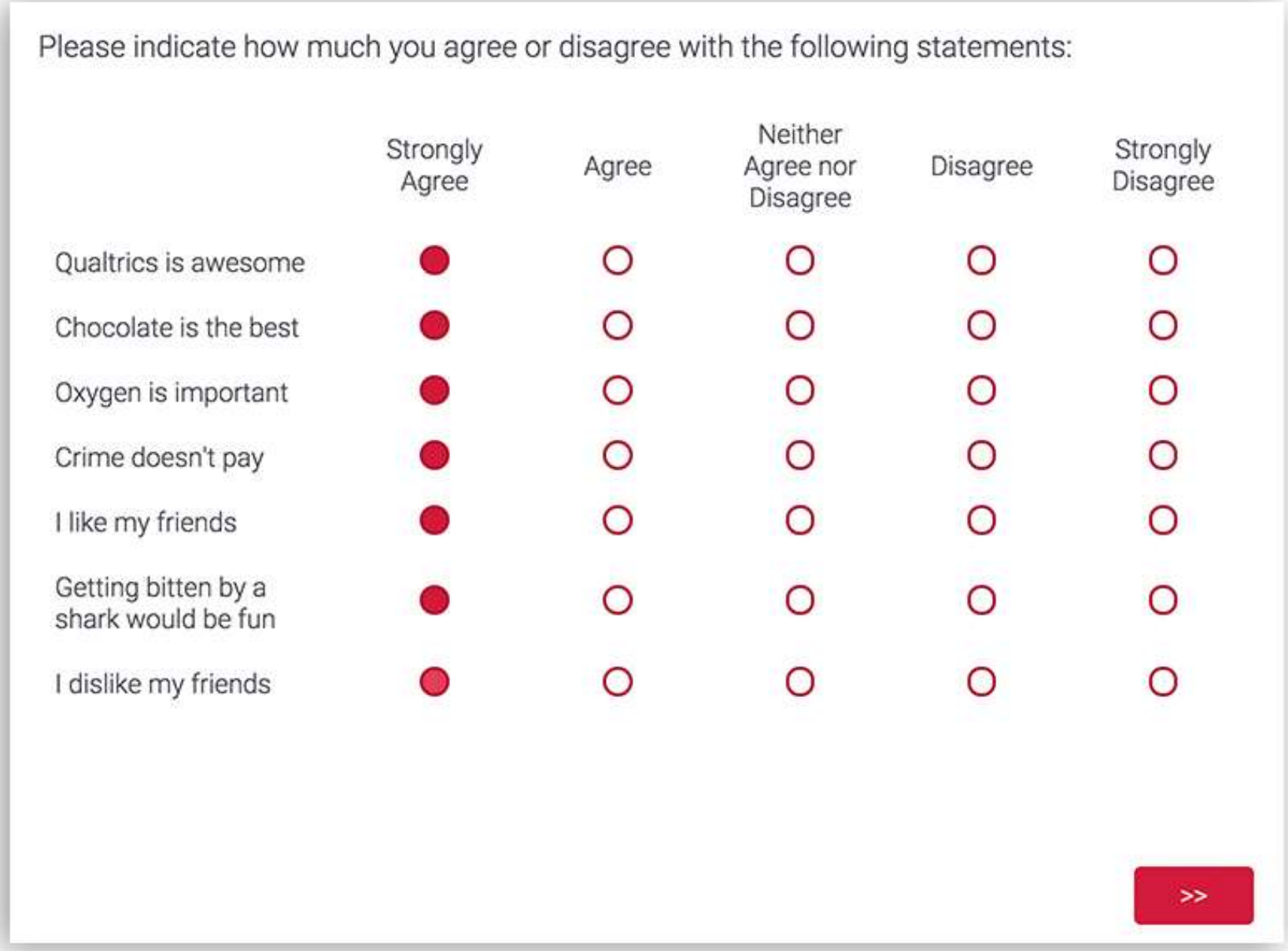
For example: "Please agree or disagree with the following statement: My internet should be faster and more stable".

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



Statement Batteries – ALWAYS make all Negative or Positive i.e. NO REVERSE CODING *

Often touted as a way to find bots and respondents not paying attention. I do not recommend!

Given that we have this huge list with most of them being 'good' what are we effectively doing?

- Training the respondents that the right part of the scale is 'good' and left 'bad'.

So what do you think happens when we purposely try to confuse them?

- 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 and often therefore don't use it.

A better way to find bots and disinterested respondents

Add some statements that should be answered using either end of the scale while keeping them both positive (or negative) e.g. *Oxygen is Important* vs *Getting bitten by a shark would be fun*.

Note that the *I like my friends* vs *I dislike my friends* statements are breaking the "keep it either positive or negative" rule. But are such easy to answer question it is unlikely to confuse people.

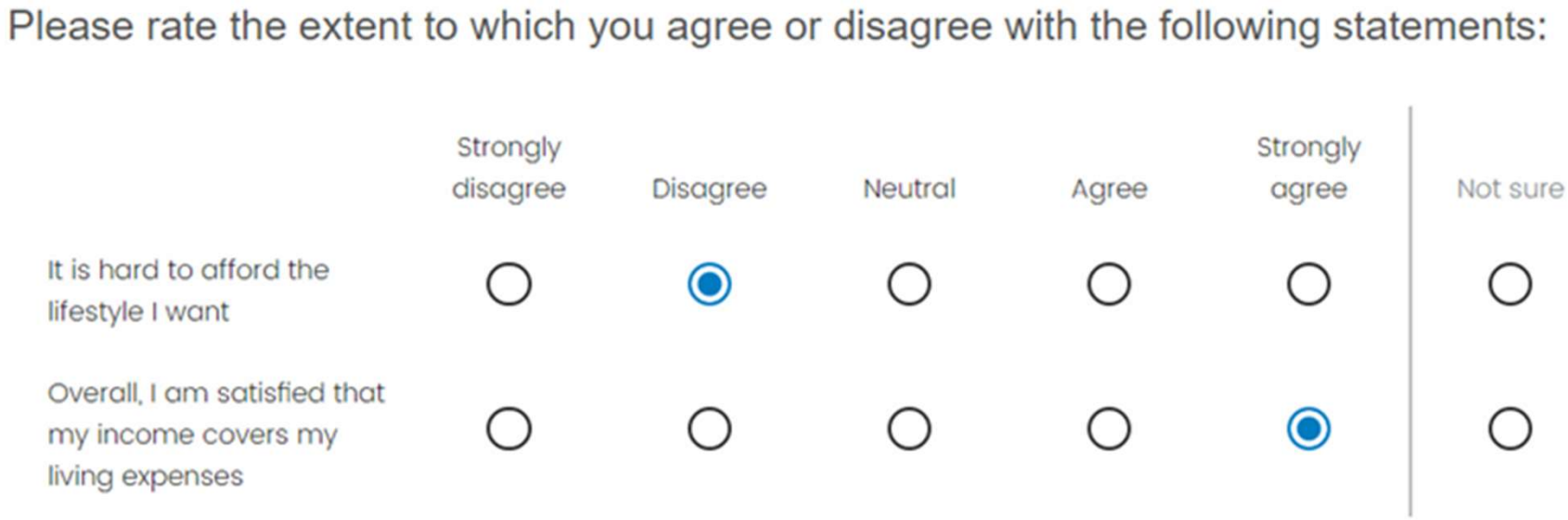
Please indicate how much you agree or disagree with the following statements:

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Qualtrics is awesome	●	○	○	○	○
Chocolate is the best	●	○	○	○	○
Oxygen is important	●	○	○	○	○
Crime doesn't pay	●	○	○	○	○
I like my friends	●	○	○	○	○
Getting bitten by a shark would be fun	●	○	○	○	○
I dislike my friends	●	○	○	○	○

A better way to find bots and disinterested respondents

You can also ask the same question but in a slightly different way. With opposite ends of the scale being the appropriately consistent answer.

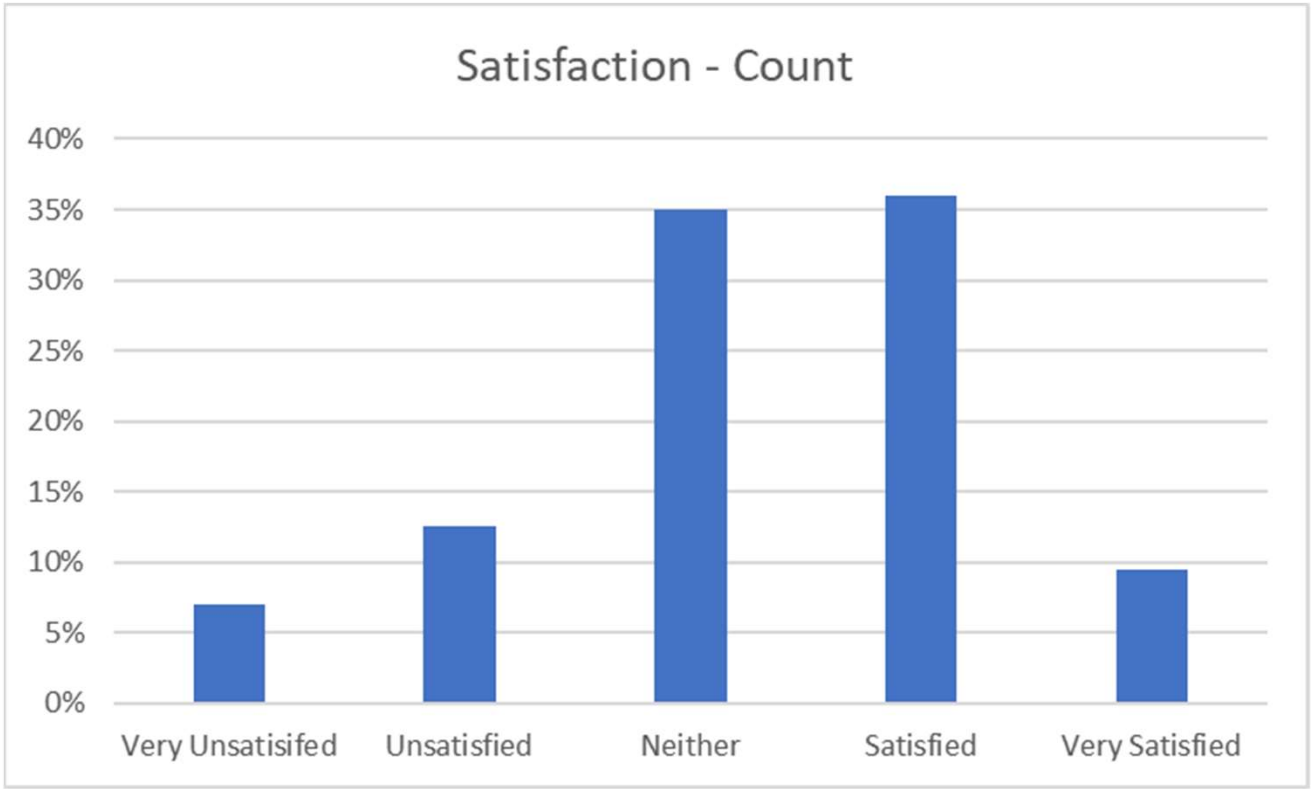
This example is a little less obviously testing respondents than the preceding ones.



EDA (Exploratory Data Analysis)

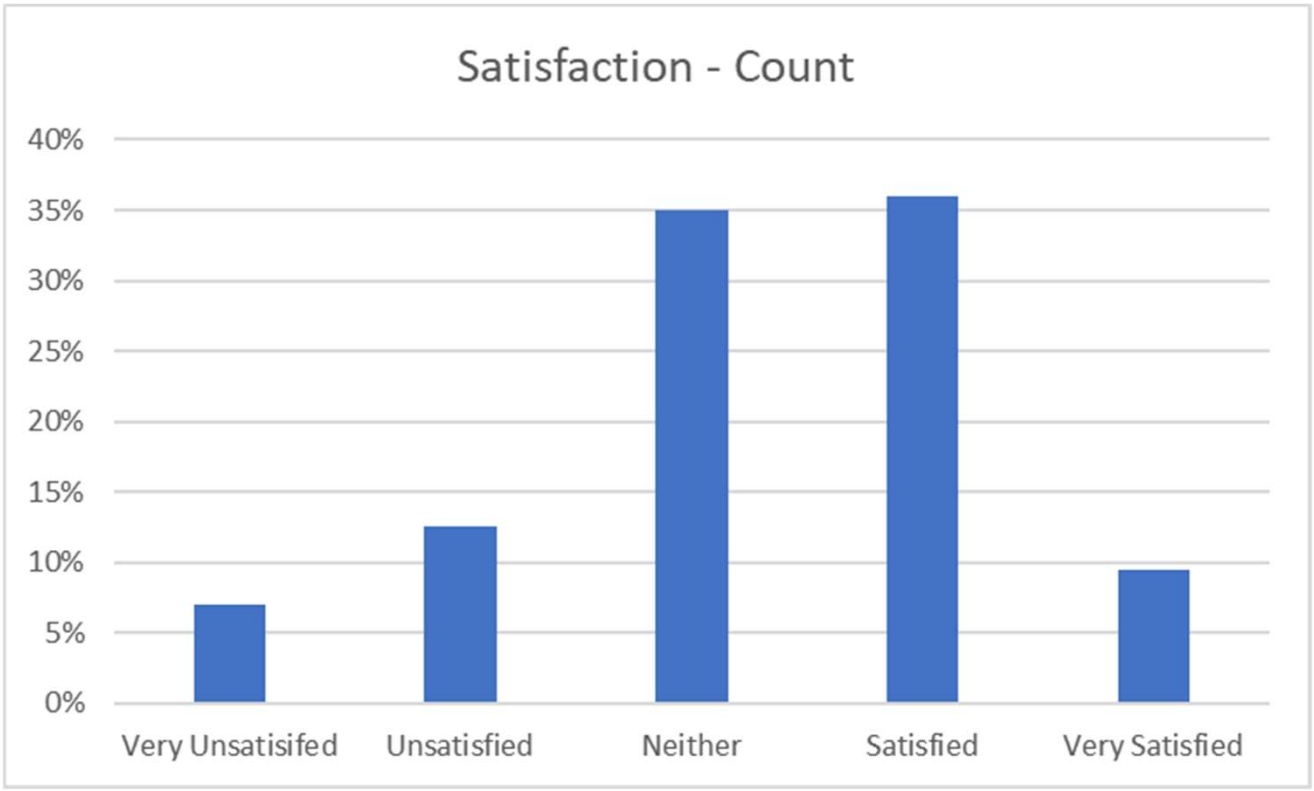
Bar Charts are great for exploratory Data Analysis, and should always be skimmed for problems. Ideally use counts for this, not %, since want to identify low categories with low sample sizes. Common problems are:

- Poor Discrimination i.e. most people in a single category
- Missing Data



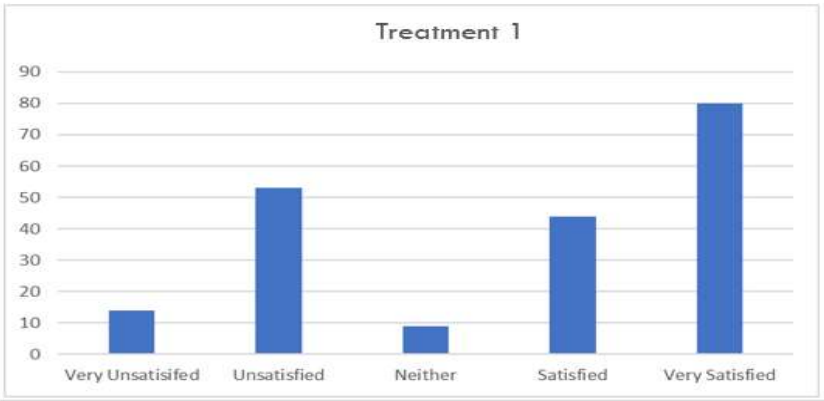
Reporting: Bar Charts of entire scale

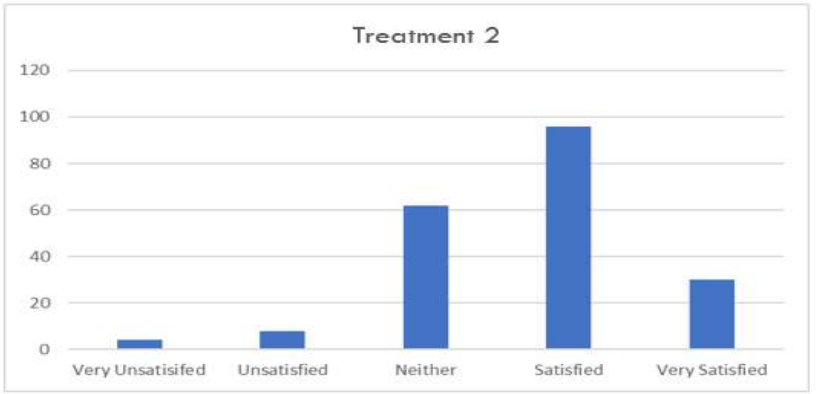
Bar Charts can also be used for reporting.

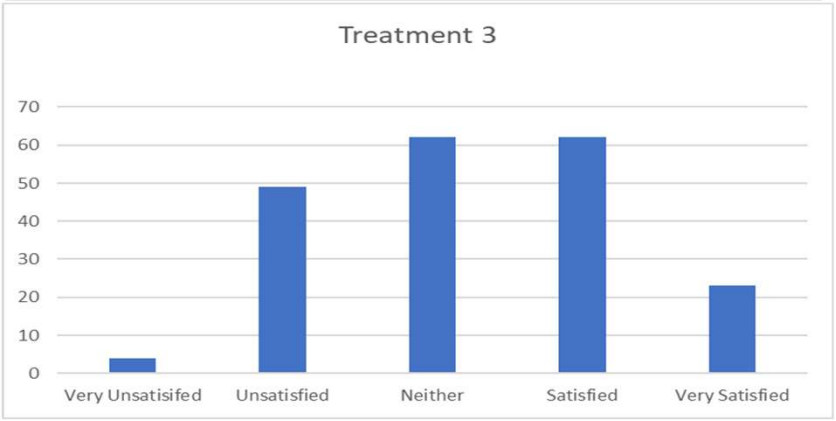


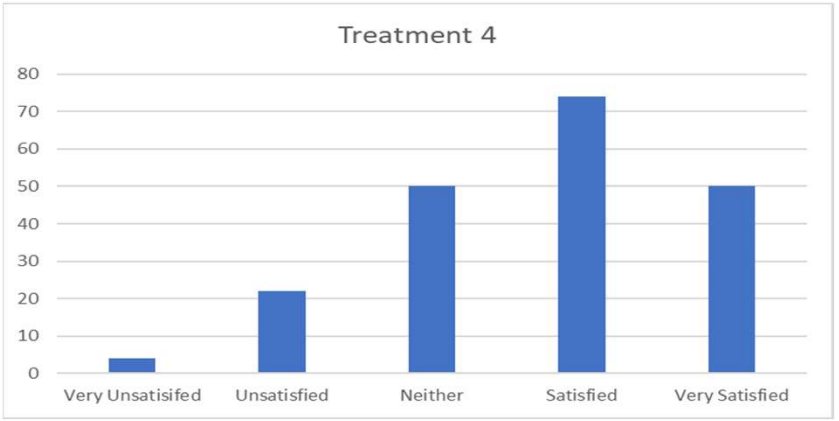
Reporting: Bar Charts of entire scale

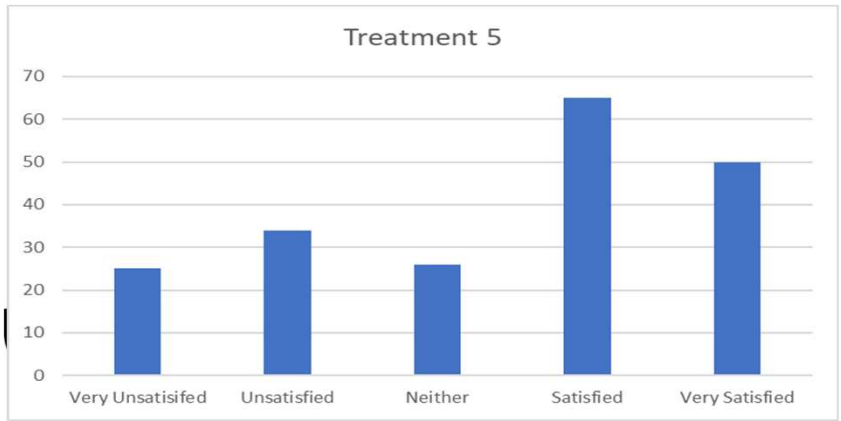
But not so good when reporting say 5+ treatments, statements, products, etc!

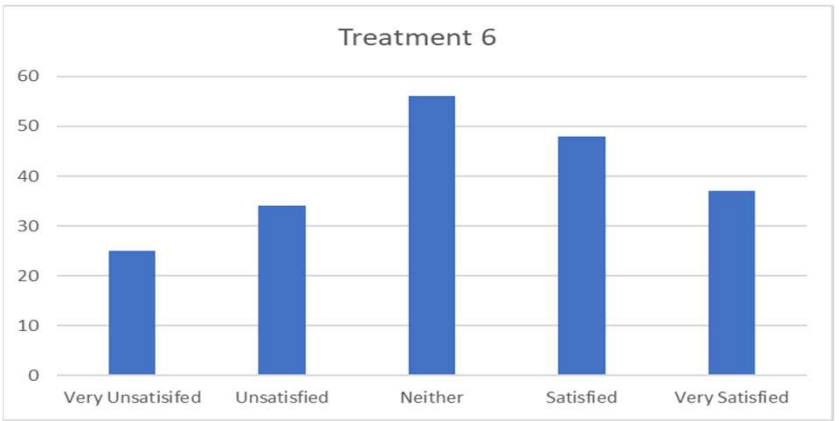






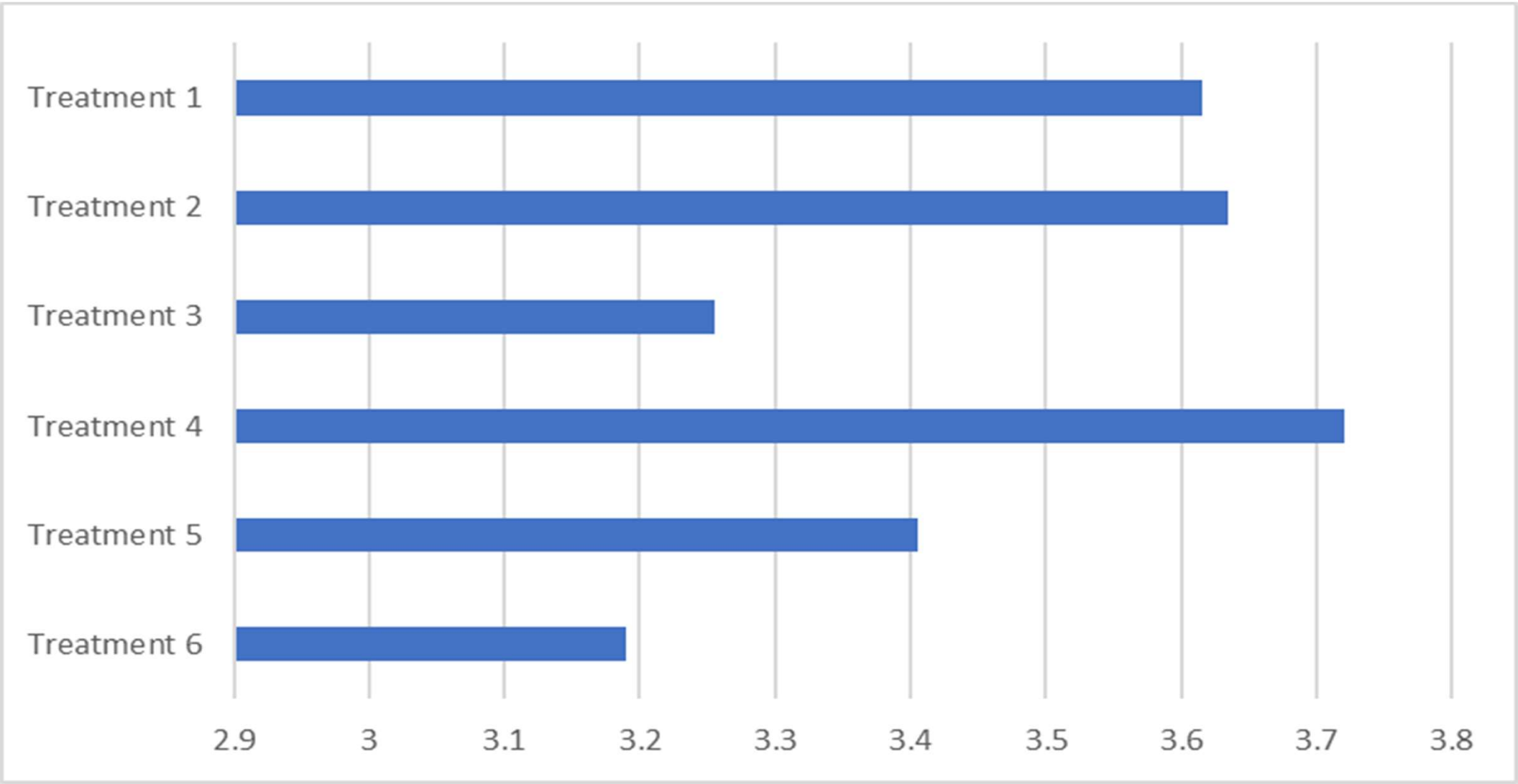






Reporting: using Averages

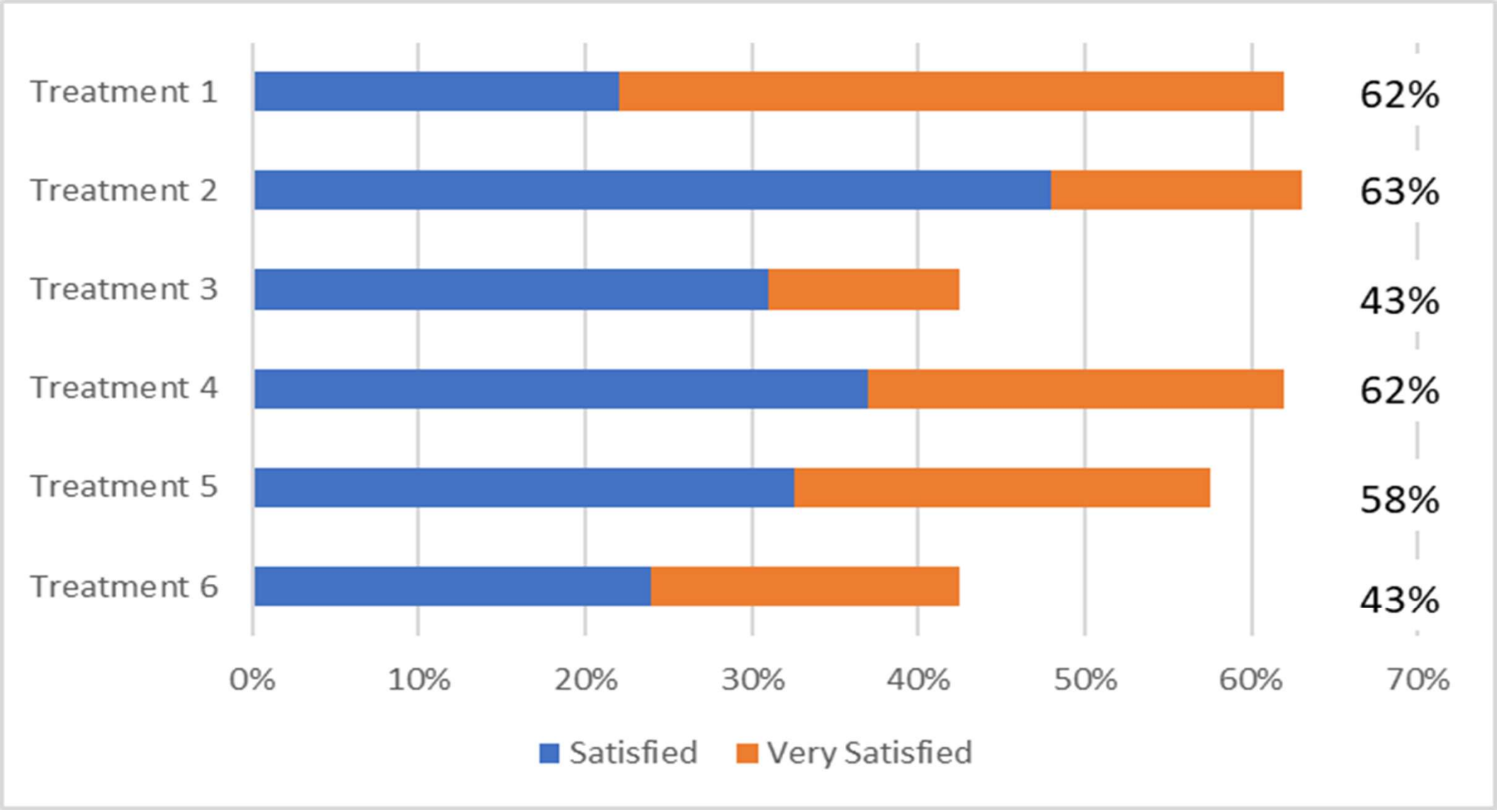
Averages allow you to report lots of items in a succinct way



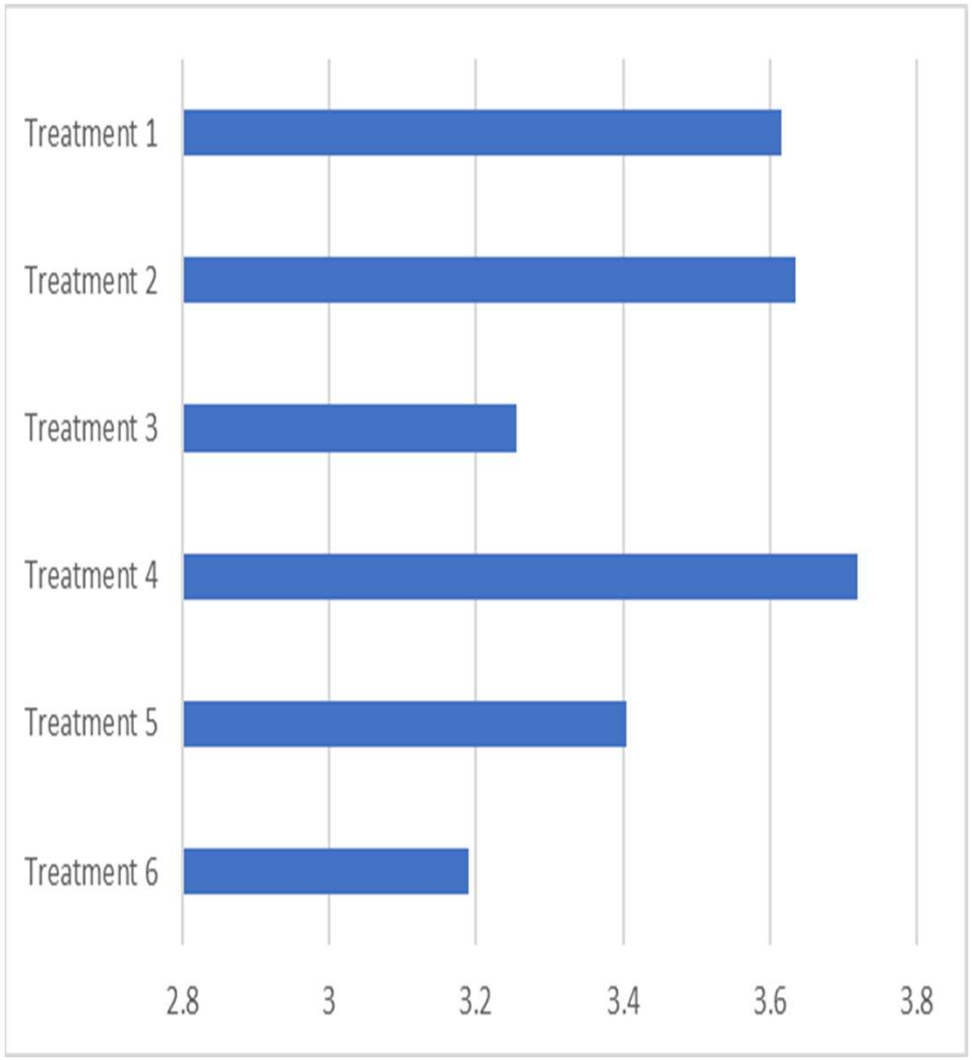
Reporting: using Top Box

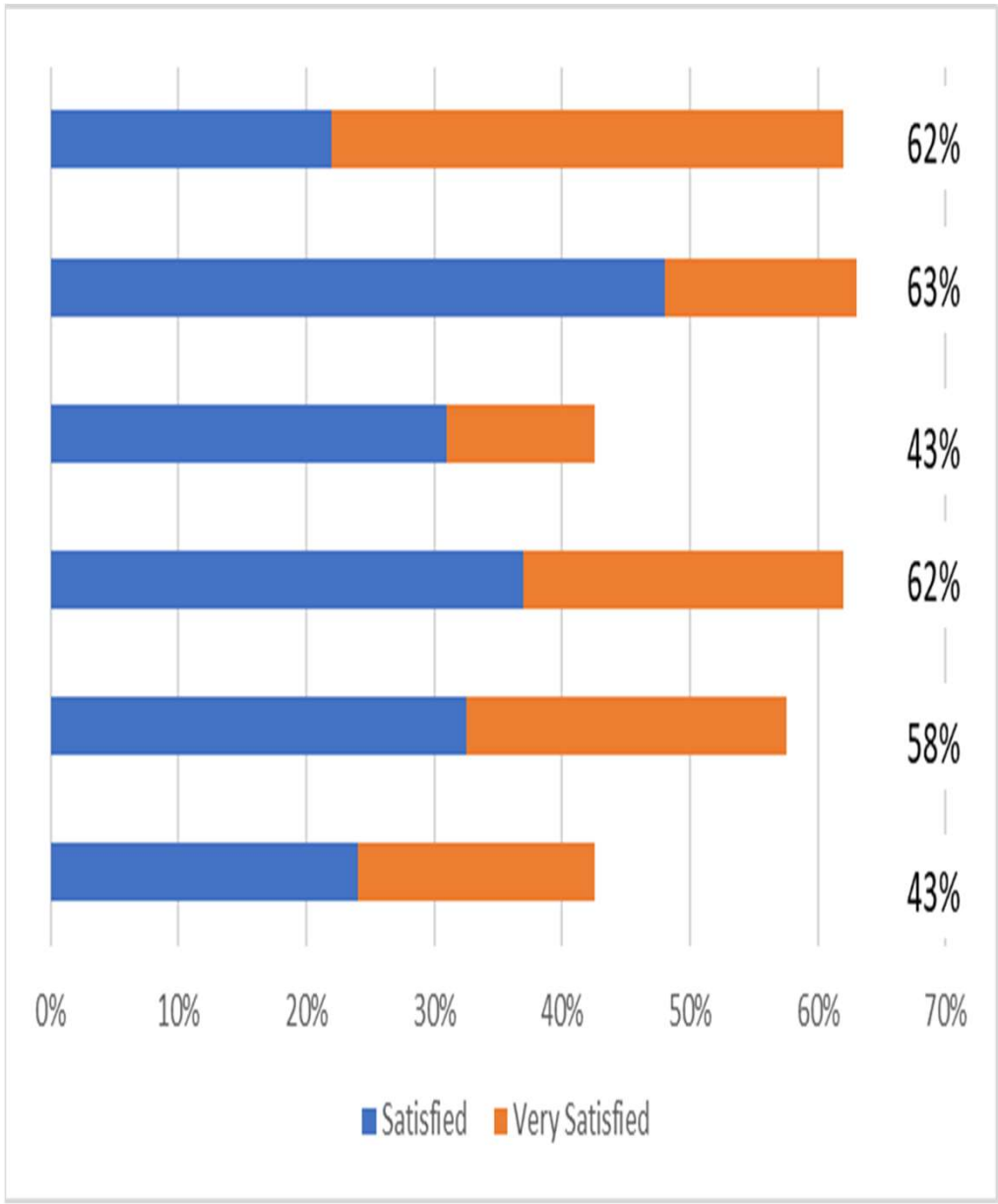
Top Box also allows you to report lots of items in a succinct way

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 **continuous** we use linear regression
  - If **ordinal** we try to use ordinal regression or if the proportional assumption fails we use logistic or multinomial
- Reporting as a mean (continuous) or as counts of the Categories (or Top Box)
  - Again, some say it's OK and others not.

Top Box

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

Normal linear regression vs ordinal regression

The case for Normal Linear Regression (and means)

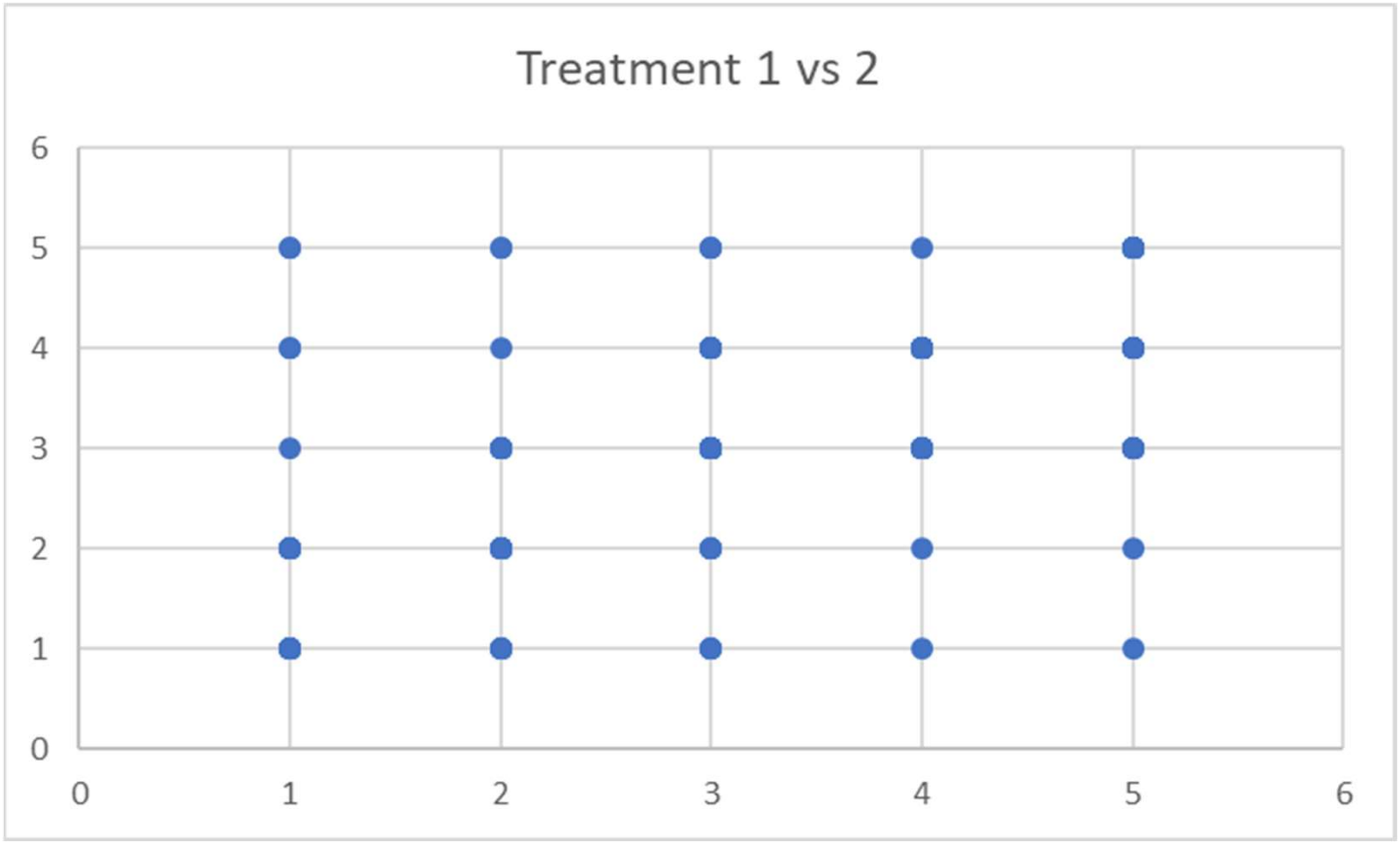
- I've been working in Market Research since 2004, across the USA, Europe and Asia. Market Research works and everyone I have ever worked with does it like this!
- Likely works since most people treat the labels as if they were equally spaced out.
- **BUT** If you do choose to do it this way consider 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?

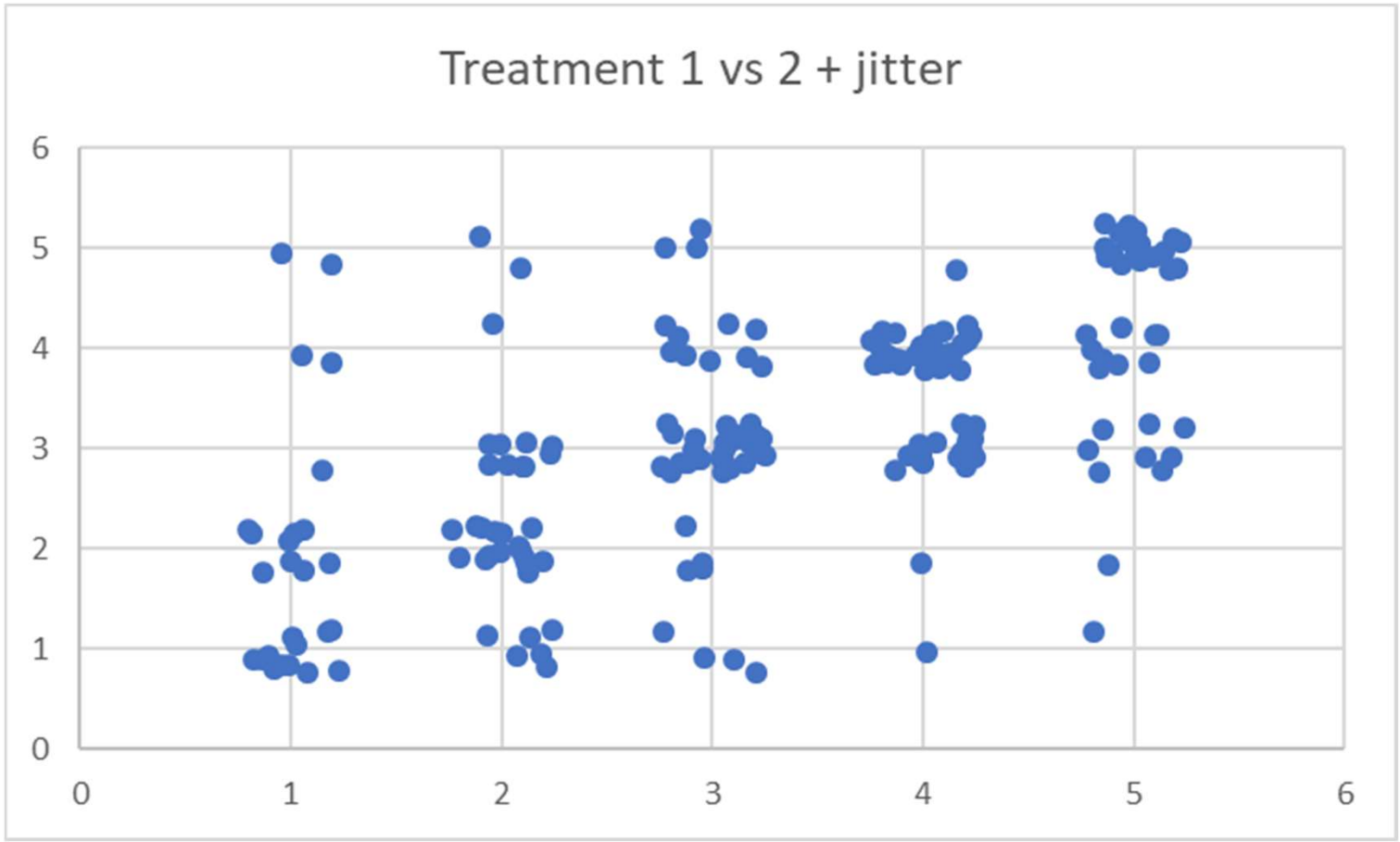
As the number of scores are limited it often comes out as a grid!!! Which doesn't help us much since we don't know how many times each combination actually occurs!



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).

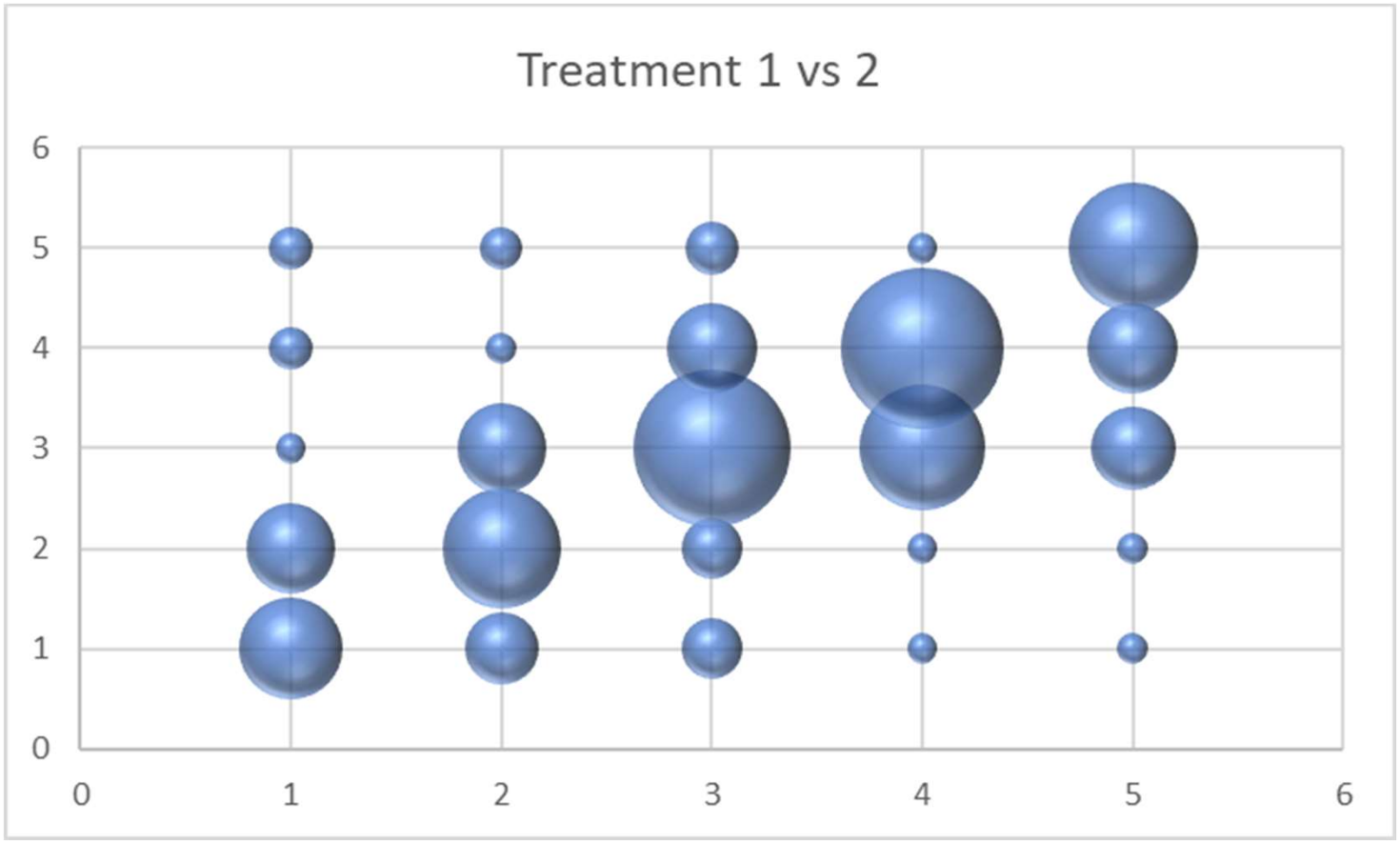
As roughly linear we can use linear regression.



Always look at the scatterplot to test linearity assumption

SOLUTION: use a bubble plot

As roughly linear we can use linear regression.



Normal linear regression vs ordinal regression

The case for Ordinal Regression

- Some researchers are uncomfortable making the assumption that the 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).

Tricks of the Trade

THE UNIVERSITY OF SYDNEY

Should I always include an "opt out" option?

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 its 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.

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 = % done
  - Ask same list + the dodgy activity you want to measure. Y= % done
  - Y-X = % who are doing the dodgy thing
    - Since they never actually select the dodgy option they are more likely to answer truthfully.

Categorical: Ancestry

Australia

The Australian census questions are difficult to find, but they ABS has written this: *"A person's ancestry, when used in conjunction with the person's birthplace, language and religion, and whether the person's parents were born in Australia or overseas, provides a good indication of the ethnic background of first and second generation Australians. To analyse ancestry, both variables (Ancestry 1st Response (ANC1P) and Ancestry 2nd Response (ANC2P)) must be used. Further information about using ancestry data is provided below."*
https://www.abs.gov.au/websitedbs/censushome.nsf/home/factsheetsa?opendocument&navpos=450

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

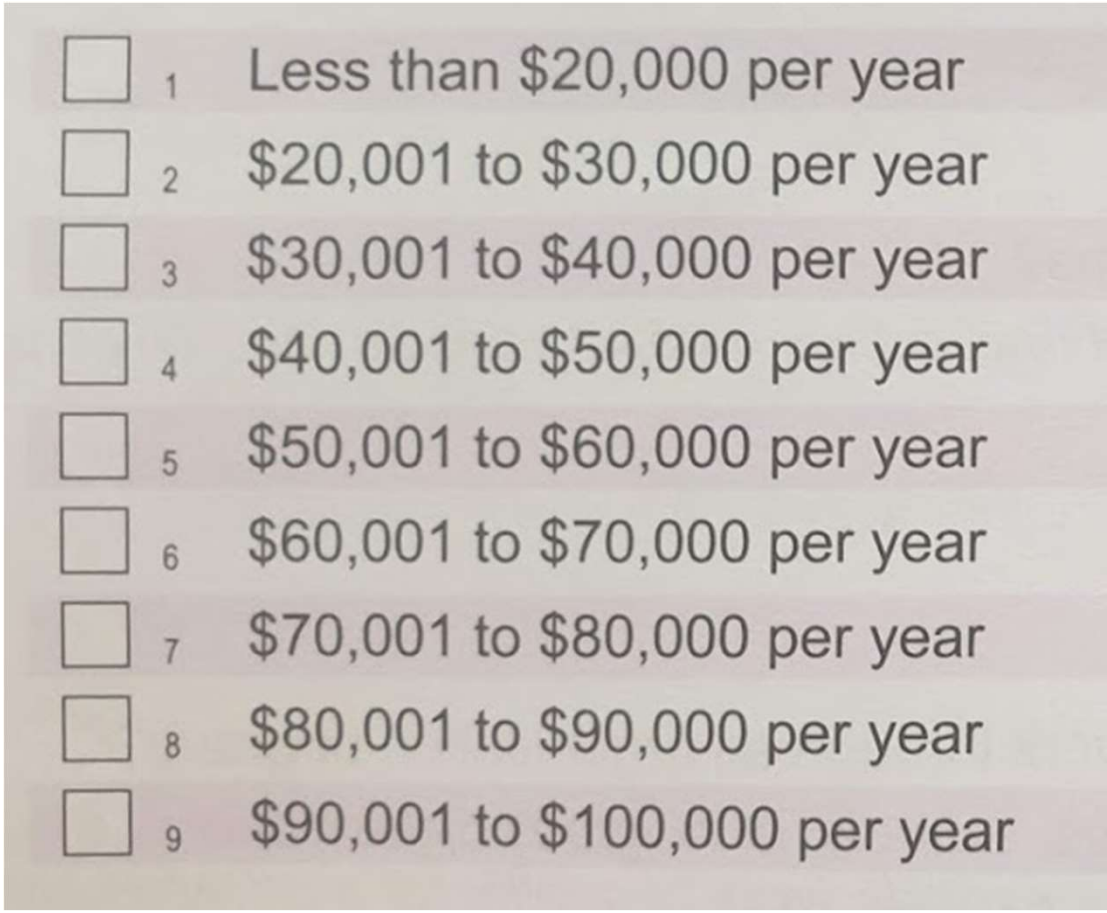
Exceptions

- 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 accidently 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 with very fine



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.

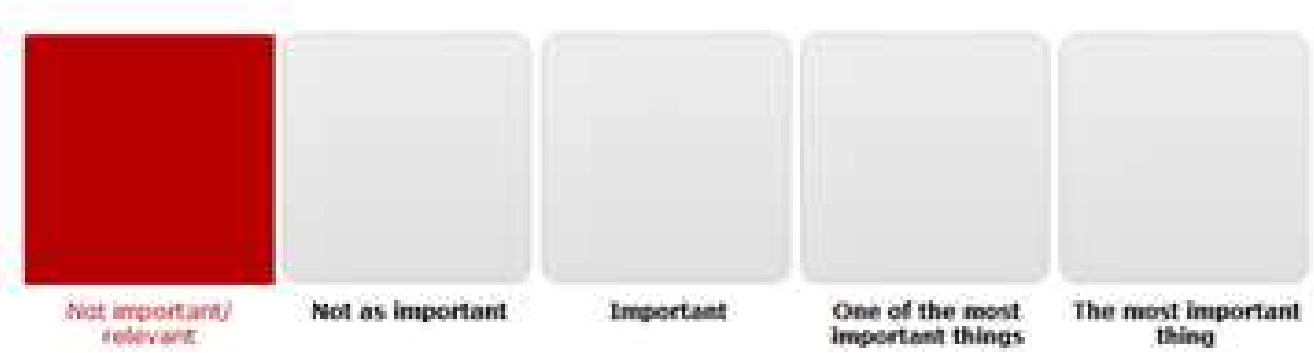
How to bias the answers

Q) Who do you think should have more influence in important policy decisions.

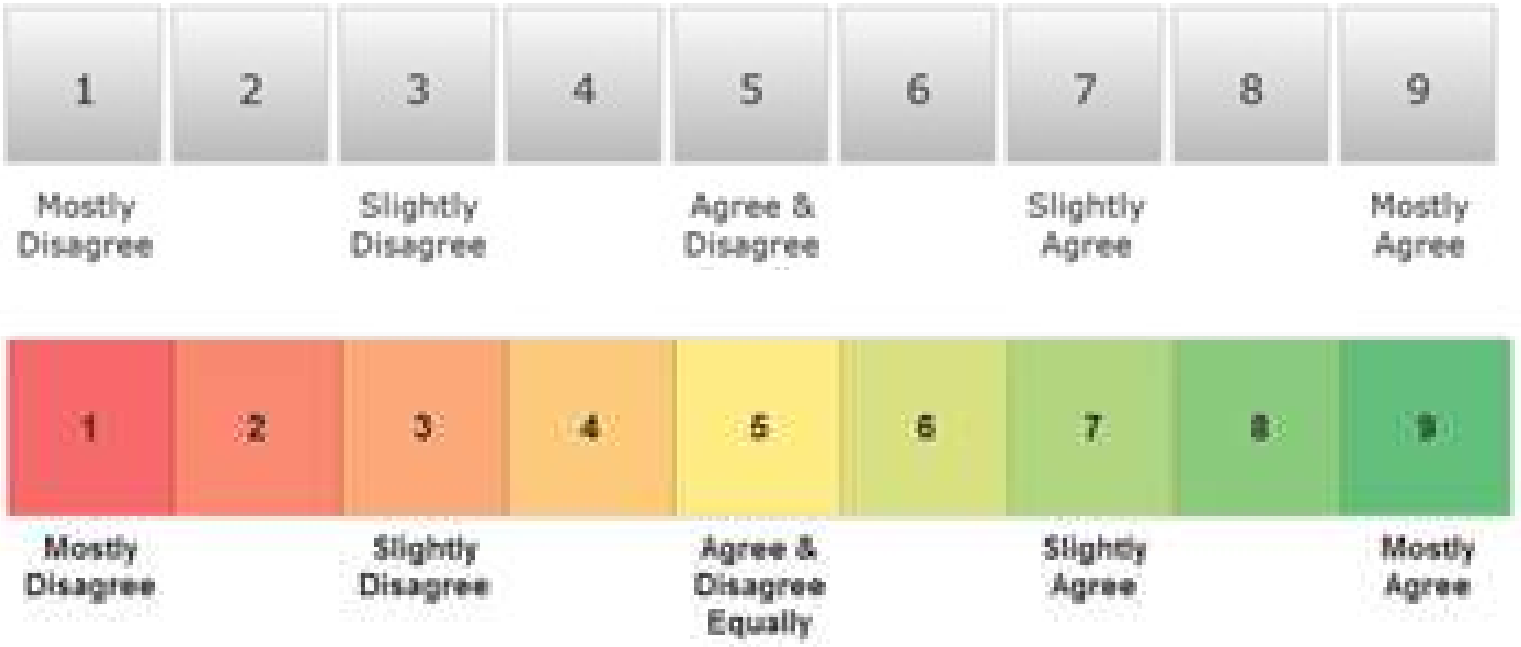
1) The people instead of politicians should make our most important policy decisions.
2) Politicians are corrupt and can't be trusted to make important decisions.
3) Politicians make decisions that harm the interests of ordinary people and can't be trusted.
4) The ordinary people should have more influence than big companies that only want to make profits.

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.**

The End! Any Questions?

Further Assistance at Sydney University

SIH

- **1on1 Consults** can be requested on our website: www.sydney.edu.au/research/facilities/sydney-informatics-hub.html OR Google "Sydney Informatics Hub" with the "I'm feeling lucky" button
- **Training** Sign up to our mailing list to be notified of upcoming training: https://signup.e2ma.net/signup/1945889/1928048/
  - Research Essentials
  - Experimental Design
  - Power Analysis
- **Online library.** Useful links and the most recent version of all our workshops.
  - https://sydney-informatics-hub.github.io/stats-resources/
- **Hacky Hour** www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html OR Google "Sydney Hacky Hour"

OTHER

- **Open Learning Environment (OLE) courses**
- **Research Data Consulting** for help with **RedCap.** To ask a question
- **LinkedIn Learning**: https://linkedin.com/learning/
  - **SPSS** https://www.linkedin.com/learning/machine-learning-ai-foundations-linear-regression/welcome?u=2196204

A reminder about Acknowledging SIH

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

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

Suggested wording:

General acknowledgement:

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

Acknowledging specific staff:

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

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

We would like your feedback on this workshop

- We will email you a link to the survey shortly
- It only takes a few minutes to complete (*really!*)
- Completing this survey is another way to help us keep providing these workshop resources free of charge