Data Analysis

Dr Charles Martin

Announcements

- **TODAY:** assignment 1 due **Monday 18 August, 23:59 on GitLab** (template)
- assignment 2 specification will be published soon, you can see the “main idea” already on Canvas
- keep attending labs, if issues, apply for an extension (see course policies on Canvas)
- any questions, problems, **use the forum** - more questions allowed (no limits!), public questions preferred.
- lab marks come out weekly via Canvas

Plan for the class

- Overview of analysis, interpretation, presentation
- Quantitative Analysis (demos!)
- Qualitative Analysis (demos!)
- Analytical Frameworks
- Interpreting and Presenting Findings

Analysis, Interpretation, Presentation

Analysis, Interpretation, Presentation

- last week we talked about getting data
- once we have it, what can we do with it?
- quantitative approaches
- qualitative approaches
- or a combination (common in HCI!)

Processes

- (data gathering)
- data cleaning
  - e.g., `age = 999`
- analysis and interpretation
  - different tasks, but often parallel
- presentation

Bias in Analysis and Interpretation

- **bias:** influence affecting objective judgement and decision making
- integrating new data with past experience (**bias is normal!**)
- **conscious:** we know about them and their effect on us
- **unconscious:** subtle effects that we don’t know about

Specific Biases

- **familiarity bias:** do what we know best
  - ignore unfamiliar methods
  - ignore unfamiliar data sources
- **self-attribution bias:** giving ourselves too much credit
  - falsely attribute improvements to our design and not external factors, user learning, etc
  - overlook alternative explanations
- **confirmation bias:** discarding information that contradicts existing belief
  - more on confirmation bias
  - ignore non-supporting evidence
  - analyse for confirmation, not for discovery
  - overlook alternative analyses and sources

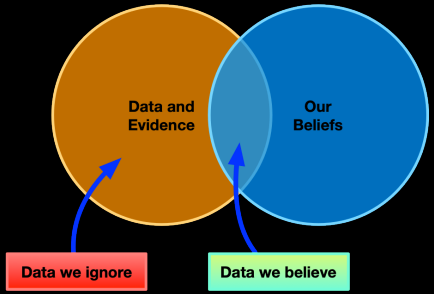

Figure 1: Confirmation bias is problematic

Quantitative and Qualitative

- typical categorisation of data
- **Quantitative data:** form of numbers or easily translated into numbers
  - years experience
  - number of minutes to perform a task
- **Qualitative data:** words, images, sound
  - descriptions
  - interview transcripts
  - photos
- Some data can be represented as both
  - e.g., digital text, sound, images can be represented and analysed numerically (is this useful?)

Quantitative and Qualitative Methods

Quantitative Analysis: find magnitude, amounts or size of *something* and make rigorous comparisons

Qualitative Analysis: find the nature of things, themes, patterns, stories

Use and Misuse of Data

- It’s easy to misuse numerical data and the results of analyses
- This can particularly happen when transforming the type of data (e.g., agreement ratings to a numerical code)
- Qualitative data can also be misused, e.g., content of questions reported as a finding.

Some example problems

- “50% of users took longer than 30 minutes” vs “2 out of 4 users…”
- The mean agreement was 3.67
- (In a study about mobile phone use with many questions about mobile phones) “Participants noted frequent use of mobile phones for many tasks”.

First Steps

- Interviews: transcribe (e.g., using Aiko), expand notes, enter closed questions into spreadsheets (treated as quantitative)
- Questionnaires: enter into spreadsheets, clean up data, filter into datasets
- Observations: expand notes, transcribe recordings, edit videos, synchronise and clean up interaction data

Basic Quantitative Analysis

Basic Quantitative Analysis

What do we do with quantitative data once we have some?

Centre of a set of data

what's the average of a set of values?

- **Mean:** sum of the values divided by the number of entries.
- **Median:** order the set numerically and find the value in the middle (or if even number of entries, halfway between the two middle entries)
- **Mode:** the most common entry

Example: `[2, 2, 3, 4, 873]`

Example shows that outliers mess up *mean*, so *median* is often more useful.

Spread of data

how spread out is a set of values?

- **Range:** `max - min`
- **Standard Deviation:** the *typical distance* of a value from the mean
- **Interquartile range:** range of the middle 50% of the data

Similarly to central measures, **interquartile range** is robust against weird outliers

Normal vs non-normal distributions

The distribution of the data is how it is spread out and where it is bunched up.

- **normal distribution:** a.k.a. bell curve, Gaussian distribution, mean, median, mode are the same, and the data evenly falls either side of the mean
- **skewed distributions:** data with a weirdly long tail in either direction
- **multi-modal distributions:** data that seems to have multiple bumps

This matters because statistical tests often assume data is *normal* so findings might be misleading.

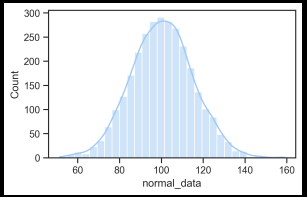

Figure 2: Normal distribution

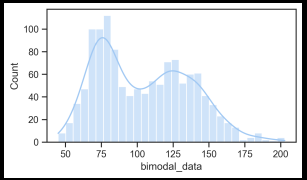

Figure 3: Bimodel distribution

Looking at the data

First think to do after loading it in. May not be the most helpful approach... but still important to check it's not garbled and the columns make sense.

interactive activities	attend in person	watch online	degree	time in CBR
5	2	2	undergraduate	1-3 years
3	5	1	postgraduate	3+ years
4	5	1	postgraduate	<1 year
5	2	4	undergraduate	<1 year
4	3	1	undergradu-	<1 year

Descriptive Statistics

Second thing to do when loading up data for analysis, calculate:

- minimum, maximum
- lower and upper quartile
- median and mean
- number of data points (count)

Ask: are these values what you expected? do they suggest any *interesting* points about your data?

stat	interactive activities	atter
count	75	75
mean	3.36	3.15
std	1.30	1.24
min	1	1
25%	2	2
50%	3	3
75%	4	4
max	5	5

Plotting Data

Third thing to do when loading data

- Plot the data to see the distribution and compare data points
- **Scatter plot:** see all the data! good for checking outliers and comparing aspects of data
- **Histogram:** useful to check distribution
- **Boxplot:** useful to compare distributions clearly (more abstract) **charles approved plot!**
- **Bar plots:** really just show one value (mean), may not be useful (too abstract!)
- **Line plot:** useful for showing data over time, not distributions

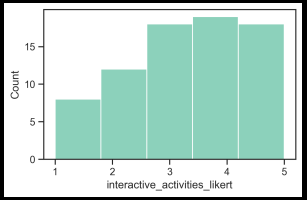

Figure 4: A histogram of some data

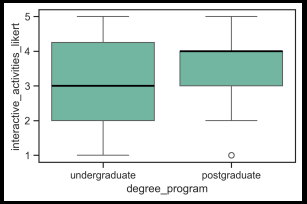

Figure 5: Box plots of the same data

More complex plots

You can get *more plots* into one plot. Good for surfacing contrasts or telling a story about the data graphically.

```
sns.set_theme(style="ticks", palette="Set2")
plt.figure(figsize=(10, 6))
sns.boxplot(data=survey_data, x='degree_progr
            y='interactive_activities_likert'
            hue='time_in_canberra',
            medianprops={'linewidth': 2, 'col
plt.savefig('plots/fake_data_complex_boxplot.png',
            bbox_inches='tight', dpi=300)
plt.show()
```

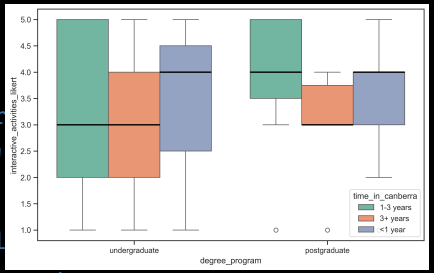

Figure 6: A more complex box plot

Comparing data and tests

- You can compare two sets of data by finding the difference between their centres or other descriptive statistics but **is the difference meaningful?**
- We use statistical significance to help ascertain meaningful differences that might be a research finding.
- A classic test is the *t*-test which compares the *means* of two sets of data, the output of a *t*-test can tell us how likely differences are to be random or **significant** (meaningful)

Signficance test notes

- we will come back to this later...
- *t*-tests assume normality and can only compare simple situations
- other tests can be used on any non-normal data 📊 and complex datasets (e.g., multiple values from each participants, multiple experimental conditions)

Quantitative Analysis with Python

Lots of ways to do data analysis:

- Excel/spreadsheets can do this a bit, but inferior to coding approaches
- Special programming languages exist: SPSS (1968–), S (1976–), R (1993–), Python (1991) + pandas (2008), julia (2012)
- R is where the statisticians, social scientists and psychologists live, Python and Julia are more where computer science folks hangout.

In this class we'll use Python, numpy, pandas, scipy, seaborn, and matplotlib as a default stack for data analysis (yes, libraries are a problem in python...)

Demo time: analysing and plotting data in Python

Let's do some *data analysis*

1. Deploy a short questionnaire on PollEverywhere
2. Download the data from pollev and load it into Google Colab
3. View descriptive statistics
4. Plot the data in a few ways

Figure 7: PollEverywhere link: https://pollev.com/charlesmarti205

Basic Qualitative Analysis

Basic Qualitative Analysis

Figure 8: Qualitative analysis, fewer numbers, still a lot of work. (Photo by Jessica Lewis 🦋 thepaintedsquare on Unsplash)

Coding in Qualitative Analysis

- in **qualitative research** the word *coding* has nothing to do with programming
- **coding:** annotating data with key words or phrases that provide a means for analysis over a large corpus
- codes can be inductive (bottom up, from the data) or deductive (top-down start with a framework of codes)
- challenges: creating meaningful, non-overlapping codes that are clearly defined and determining granularity
- inter-rater reliability measures the clarity and reliability of the coding scheme rather than correctness of analysis

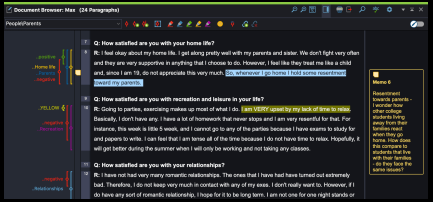

Figure 9: Coding text in MAXQDA (Image by MAXQDA)

Analysing Video Material

- initial viewing involves watching entire recording while writing high-level narrative and noting timestamps of interesting events
- chronological and video times used to index events
- data augmentation: video plus screen captures, interaction logs, transcriptions
- coding schemes applied systematically to ensure reliability and consistency
- team-based coding can address subjectivity, ambiguity, and potential errors
- expert knowledge can help with unclear

Figure 10: Analysing some performance videos in 2012 (Charles)

Identifying Themes

- Many qualitative research approaches distill *themes* from collected data.
- code the data (apply labels) and then create higher level themes from codes.
- (Reflexive) Thematic Analysis (RTA) (Braun & Clarke, 2022) is a well-known and accessible methodology.

Typical process

1. Familiarise with the data
2. Coding (short labels, multiple rounds)
3. Generating initial themes
4. Developing, reviewing, and refining themes

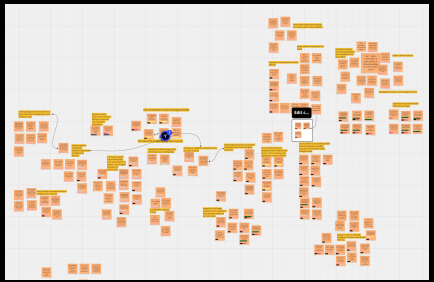

Figure 11: A Miro board from Yichen Wang's thematic analysis (2025)

What is a theme, and why does it matter?

A theme is a high level finding from qualitative analysis, but what that means can differ.

- **Two kinds of theme:** patterns of meaning (uncovering implicit meaning behind words) versus data summaries (summarising responses across participants) (Braun & Clarke, 2019)
- themes can be created organically from the data through interpretation, or follow predetermined frameworks and categories
- no one way is correct, but need to be deliberate in methodology in particular to be careful about:
  - whether themes are a summary or reflect hidden or implicit meaning
  - whether themes arise from data or come from a pre-determined framework
  - whether we want “correct” codes, or to ensure that codes are clearly defined and interpreted consistently (can use multiple researchers to help)

Affinity Diagram

- **affinity diagrams** used for organising large amounts of data and identifying themes and overall narratives
- both digital (e.g., Miro) and physical (e.g., sticky notes) diagramming approaches
- not necessarily a lot to this: summarise the data on notes, and arrange to find relationships between them.
- my PhD/master/Honours students tend to do this with Miro
- See this resource for a guide.

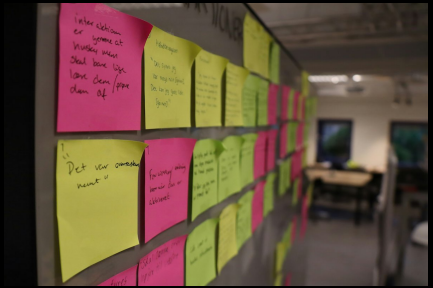

Figure 12: Analogue affinity diagrams often use sticky notes. (Photo by Christian Brok on Unsplash)

Categorising Data

- deductive analysis applies pre-existing theoretical frameworks or categories to analyse data
- data is systematically coded to segment and categorise specific elements, allowing for pattern identification and quantitative analysis
- e.g.: take a specific categorisation scheme from previous research (interface problems, or design recommendations), apply to new situation
- taking a quantitative approach: counting occurrences of categories per participant, identifying specific situations or issues

Critical Incident Analysis

- focus on significant behavioral incidents rather than general impressions
- identifying specific moments that are pivotal in either positive or negative ways
- makes data analysis more manageable and focused
- can be identified by
  - participants during retrospective discussions,
  - observers through real-time monitoring
  - through video analysis of recorded sessions
- more info: nngroup, usability bok

Questions: Who has a question?

Who has a question?

- I can take *cathchbox* question up until 2:55
- For after class questions: meet me outside the classroom at the bar (for 30 minutes)
- Feel free to ask about **any aspect of the course**
- Also feel free to ask about **any aspect of computing at ANU**! I may not be able to help, but I can listen.

Figure 13: Meet you *at the bar* for questions. 🍸🥤🫖☕ Unfortunately no drinks served! 🙃

References

References i

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage Publications.