UNIVERSITY OF CANBERRA

Survey Research and Design in Psychology

Lecture 3 - Descriptives and graphing

Dr Ananthan Ambikairajah

University of Canberra

A quick recap

UNIVERSITY OF CANBERRA

- The research process
- Types of surveys
- Survey design
- Levels of measurement
- Sampling
- Biases

Homework from last week - survey item design

UNIVERSITY OF CANBERRA

- Create some items to measure extroversion
- Try and come up with at least one item using each of the response formats!
  - Dichotomous and multichotomous
  - Multiple response
  - Ranking
  - Verbal frequency
  - Likert
  - Semantic differential
  - Graphical
  - Non-verbal

Homework from last week - survey administration

UNIVERSITY OF CANBERRA

- How would you administer a survey on...
  - Prejudice and discrimination?
  - Children's favourite toys?
  - Experiences of individuals living in remote areas?
- Most importantly, justify your answers!

Homework from last week - sampling methods

UNIVERSITY OF CANBERRA

- How would you recruit participants for a study on...
  - Ecstasy use among students?
  - Comparing cognitive abilities of pregnant and non-pregnant women
  - Polling Australian's voting intentions?
- Most importantly, justify your answers!

Focus/Overview for today

UNIVERSITY OF CANBERRA

- Getting to know a data set
- Levels of measurement and types of statistics
- Descriptive statistics
- Normal and non-normal distribution
- The effect of skew on central tendency
- Principles of graphing
- Univarite graphical techniques

The steps to get to know your data

UNIVERSITY OF CANBERRA

- Have a play around to familiarise yourself with it
- Don't be afraid - you can't break it, just remember to keep back up copies
- Screen and clean your data
- Check your data carefully
- Describe the main features of your data
- Test your hypotheses to answer your research question

Screening data

UNIVERSITY OF CANBERRA

- Check for out-of-range values
- Remove or replace errors
- Reverse code items where necessary
- Change ‘not applicable’ responses to missing data
- Giving variables sensible names and labels
- Dealing with cases with lots of missing values
- Document and report your steps!

Which statistics can I use?

UNIVERSITY OF CANBERRA

- It depends on the level of measurement of your data

Levels of measurement

UNIVERSITY OF CANBERRA

- Nominal
  - Attributes are named
- Ordinal
  - Attributes are named
  - Attributes can be ordered
- Interval
  - Attributes are named
  - Attributes can be ordered
  - Distance is meaningful
- Ratio
  - Attributes are named
  - Attributes can be ordered
  - Distance is meaningful
  - True zero
- Each level (i.e. nominal $>$ ordinal $>$ interval $>$ ratio) can have the properties of the levels that came before, plus something more

Levels of measurement

UNIVERSITY OF CANBERRA

- Pay attention to the level of measurement of your data
- Levels of measurement determines what kind of descriptive statistics, graph and inferential statistics you can use

Levels of measurement and parametric vs non-parametric

UNIVERSITY OF CANBERRA

- Categorical and ordinal dependent variable -> non-parametric
- Interval and ratio dependent variable -> check the distribution
  - If the distribution is normal -> parametric
  - If the distribution is non-normal -> non-parametric

Parametric vs non-parametric

UNIVERSITY OF CANBERRA

- Parametric statistics
  - Estimate the parameters of a population, based on a normal distribution
    - Univariate - mean, standard deviation, skewness, kurtosis
    - Bivariate - correlation, linear regression, t-tests
    - Multivariate - multiple linear regression, analysis of variance (ANOVA)
- Non-parametric statistics
  - Do not assume sampling from a population which is normally distributed
    - Univariate - median, frequencies
    - Bivariate - Spearman's correlation, Chi-square test

Parametric vs non-parametric

UNIVERSITY OF CANBERRA

- Parametric statistics
  - More powerful - more sensitive
  - More assumptions - normal distribution
  - Vulnerable to violations of assumptions - less robust
- Non-parametric statistics
  - Less powerful - less sensitive
  - Fewer assumptions - do not assume a normal distribution
  - Less vulnerable to assumption violations - more robust

How many variables are you working on?

UNIVERSITY OF CANBERRA

- One -> Univariate (mean, median, mode, histogram, bar chart)
- Two -> Bivariate (correlation, t-test, scatterplot, clustered bar chart)
- More than two -> Multivariate (reliability analyses, factor analysis, multiple linear regression)

Univariate descriptive statistics

UNIVERSITY OF CANBERRA

- What do you want to describe about the data?
  - Central tendency - frequency, mode, median, mean
  - Distribution
    - Spread (dispersion): minimum and maximum value, range, percentiles, variance, standard deviation
    - Shape: skewness, kurtosis.

Central tendency

UNIVERSITY OF CANBERRA

- Statistics which represent the 'center' of a frequency distribution
  - Mode - most frequent
  - Median - 50th percentile (middle score if scores are arranged from smallest to largest)
  - Mean - average
- How do you know which one to use?
  - Depends on your type of data (level of measurement) and the shape of your distribution (especially skewness)
- Reporting more than one might be appropriate

Central tendency

UNIVERSITY OF CANBERRA

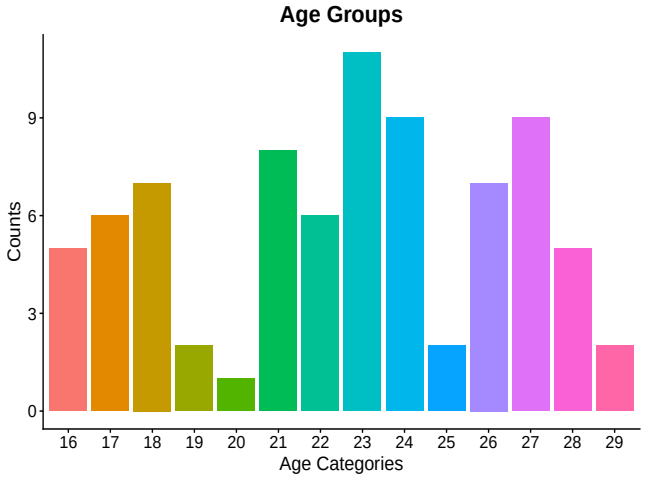
	Mode, frequency & percentages	Median	Mean
Nominal	Yes	No	No
Ordinal	Yes	If it's meaningful	No
Interval	Yes	Yes	Yes
Ratio	If it's meaningful	Yes	Yes

Mode

UNIVERSITY OF CANBERRA

- Mode is the most common score - the highest point in a frequency distribution, the most common response
- Suitable for all levels of data (. . . but it might not be meaningful for continuous/ratio data)
- Is not affected by outliers
- Check frequencies and bar graph to see whether it is useful

What is the mode?



Frequencies (f) and percentages (%)

UNIVERSITY OF CANBERRA

- The number of responses in each category
- The percentage of responses in each category
- Frequency table
- Can also visualise this using a bar or pie chart

Median (Mdn)

UNIVERSITY OF CANBERRA

- The mid-point of the distribution - quartile 2, 50th percentile
- Not badly affected by outliers
- Might not represent the central tendency if the data is skewed
- If the median is useful, other percentiles might be worth reporting

Mean

UNIVERSITY OF CANBERRA

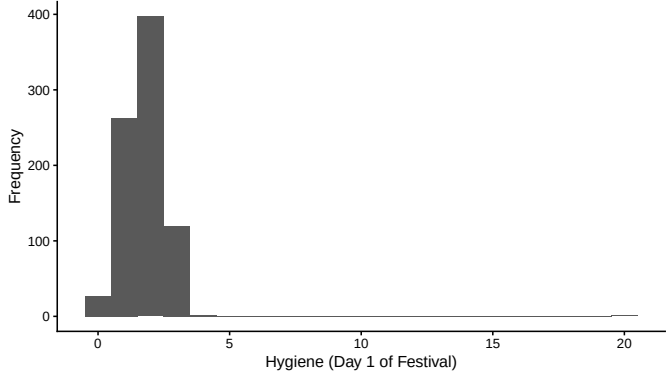
- The mean is the average score
- It is calculated by summing all scores and dividing them by the number of scores
- It is used for normally distributed ratio or interval data
- It is sensitive to extreme scores/outliers
- Sometimes it is inappropriate
  - E.g. If there is a bimodal distribution, the 'average' is describing a value where it is possible we have no scores

Mean - sensitive to outliers (and mistakes!)

UNIVERSITY OF CANBERRA

Data with outlier

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.020   1.312   1.790   1.793   2.230  20.020
```



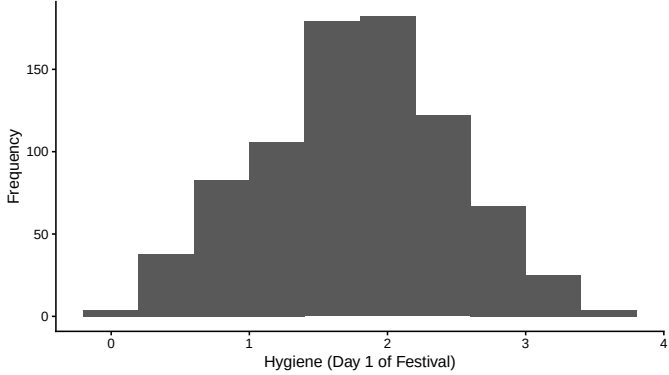
Example and dataset from Field et al. (2012).

Mean - sensitive to outliers (and mistakes!)

UNIVERSITY OF CANBERRA

Data without outlier

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.020   1.312   1.790   1.771   2.230   3.690
```



Example and dataset from Field et al. (2012).

Distribution

UNIVERSITY OF CANBERRA

- Measures the shape, spread and dispersion of your data, as well as the deviation from the central tendency
- How do you decide which statistics to use?
  - Parametric
    - Standard deviation
    - Skewness
    - Kurtosis
  - Non-parametric
    - Minimum and maximum
    - Range
    - Percentiles

Distribution

UNIVERSITY OF CANBERRA

	Min/max and range	Percentile	Variance/SD
Nominal	No	No	No
Ordinal	Yes	If it's meaningful	No
Interval	Yes	Yes	Yes
Ratio	Yes	Yes	Yes

Standard deviation (SD)

UNIVERSITY OF CANBERRA

- The standard deviation is the square root of the variance
- It is calculated with this formula: $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$
- Use the SD for normally distribution or interval data
- Again, the SD is affected by outliers
- You can also derive the standard error from the SD: $\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$

Options for nominal data

UNIVERSITY OF CANBERRA

- Recall that nominal data are labelled categories
- You can describe this data, but it is a little different
  - Which is the most frequent? (similar to the mode, e.g. females)
  - Which is least frequent? (e.g. males)
  - What are the frequencies? (20 females, 10 males)
  - Percentages? (67% females, 33% males)
  - Cumulative percentages
  - Ratios (twice as many females than males)

Descriptives for ordinal data

UNIVERSITY OF CANBERRA

- Recall that ordinal level of measurement means the data tells you about order, but not distance (i.e. options are ranked)
- Descriptives approach is the same as for nominal, but you can use percentiles (including median)

Descriptives for interval data

UNIVERSITY OF CANBERRA

- Interval level of measurement conveys information about order and distance (but the zero is arbitrary/not meaningful)
- You can describe this data in a number of ways:
  - Central tendency - mode, median and mean
  - Shape/spread - minimum, maximum, range, standard deviation, skewness, kurtosis
- Interval data is discrete, but often treated as continuous (especially if there are more than 5 intervals)

Descriptives for ratio data

UNIVERSITY OF CANBERRA

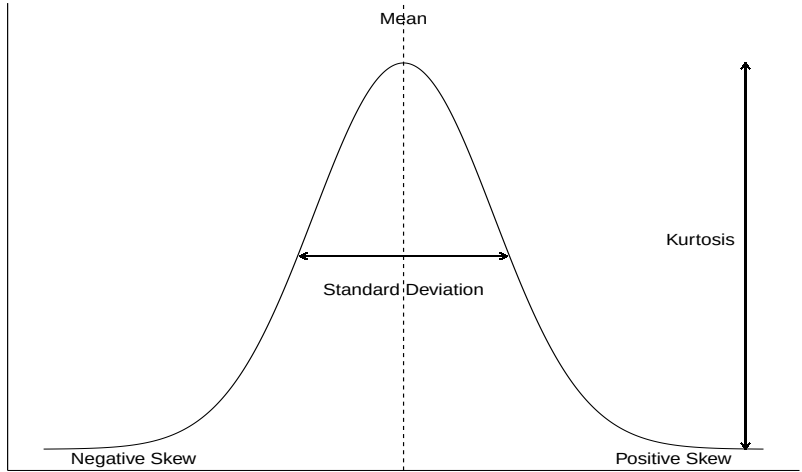
- With ratio data, the numbers convey the order and distance, with a meaningful zero
- Use the same descriptives as for interval data, but you can also use ratios (e.g. Group A is twice as tall as Group B)

Summary of descriptive statistics

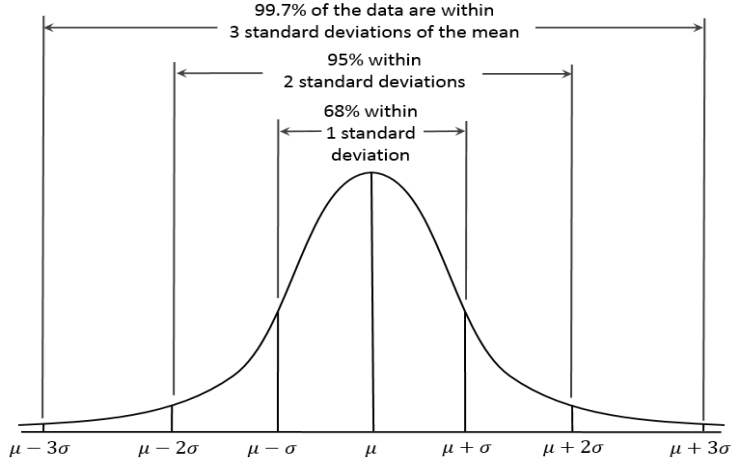
UNIVERSITY OF CANBERRA

- Level of measurement and normality of your data determines whether data can be treated as parametric
- Describe the central tendency using:
  - Frequencies
  - Percentages
  - Mode
  - Median
  - Mean
- Describe the distribution using:
  - Minimum and maximum values
  - Range
  - Quartiles
  - Standard deviation
  - Variance

Four moments of a normal distribution



A normal distribution



Example

UNIVERSITY OF CANBERRA

biobank[uk]

Index | Browse | Search | Catalogues | Downloads | Login | Help

Data-Field 12144

Description: Height

Category: Assessment centre ▸ Physical measures ▸ Anthropometry ▸ Body size measures
Physical measures

Participants	101,798	Value Type	Integer, cm	Sexed	Both sexes	Debut	Sep 2015
Item count	122,854	Item Type	Data	Instances	Defined (2)	Version	Aug 2025
Stability	Accruing	Strata	Primary	Array	No	Cost Tier	o1 s1

Data | 2 Instances | Notes | 1 Related Data-Field | 1 Resource | 1 Application

122,854 items of data are available, covering 101,798 participants.
Defined-instances run from 2 to 3, labelled using Instancing 2.
Units of measurement are cm.

Maximum	204
Decile 9	182
Decile 8	178
Decile 7	175
Decile 6	172
Median	170
Decile 4	167
Decile 3	164
Decile 2	161
Decile 1	157
Minimum	110

- There are 74 distinct values.
- Mean = 169.566
- Std.dev = 9.47034
- 5 items below graph minimum of 132

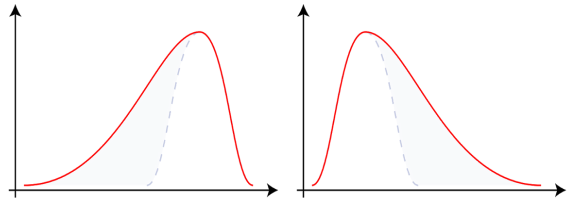
Enabling scientific discoveries that improve human health

UK Biobank, https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=12144

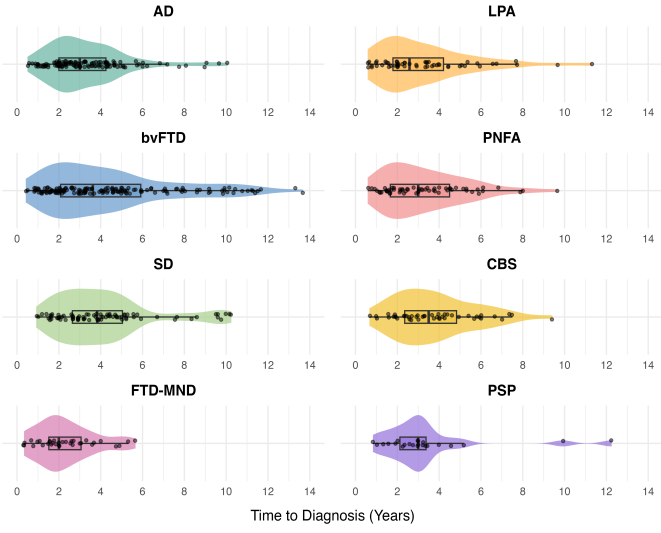
Skewness

UNIVERSITY OF CANBERRA

- Skewness is a measure of the lean of a distribution
- Look for the tail (where there are fewer values)
  - Tail to the right = positive skew
  - Tail to the left = negative skew
- What causes skew?
  - An outlier
  - Floor effects
  - Ceiling effects
  - Check the chart to see what's going
- Skewed data is not always a mistake



Example



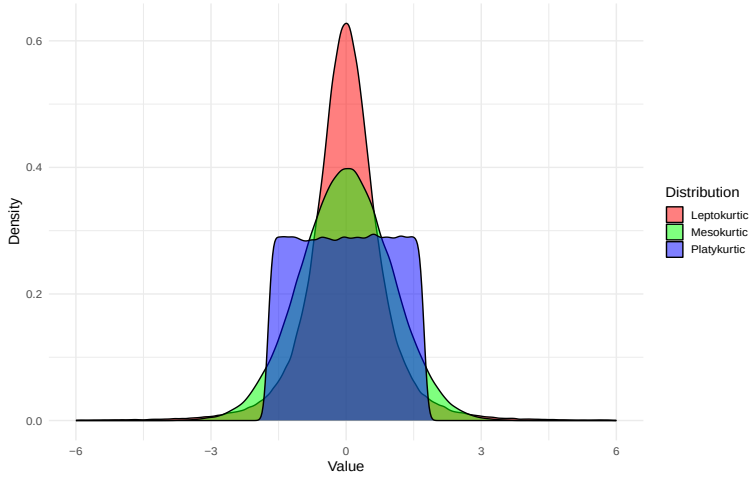
Ambikairajah et al. (2025)

Kurtosis

UNIVERSITY OF CANBERRA

- Kurtosis is how flat vs how peaked the distribution is
  - Peaked data = positive kurtosis
  - Flat data = negative kurtosis
- A distribution can look more peaked or flat depending on how the graph is set up (the X and Y axes) so add a normal curve to judge kurtosis visually

Kurtosis example



How do you judge how severe the skewness and kurtosis is?

UNIVERSITY OF CANBERRA

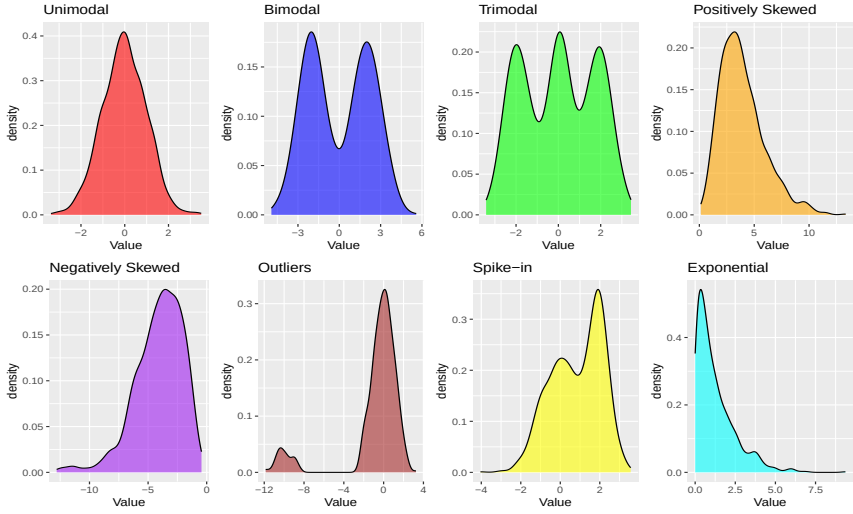
- Check the histogram
- Are there outliers? Deal with them
- Run the skewness/kurtosis analyses (will give you a value and significance for the test)
- A rule of thumb: Skewness and kurtosis values between -1 and $+1$ is generally 'normal enough' to meet the assumptions for parametric inferential statistics, but many use $\pm$ 2.5 as the cut off
- The significance test for skewness tends to be overly sensitive

Things to look at on non-normal distributions

UNIVERSITY OF CANBERRA

- How many peaks are there?
  - One peak = unimodal
  - Two peaks = bimodal
  - More than two peaks = multi-modal
- Is there a tail? (skewness)
  - To the right = positive skew
  - To the left = negative skew
- How peaked or flat is it? (kurtosis)
  - Flat = platykurtic
  - Peaked = leptokurtic

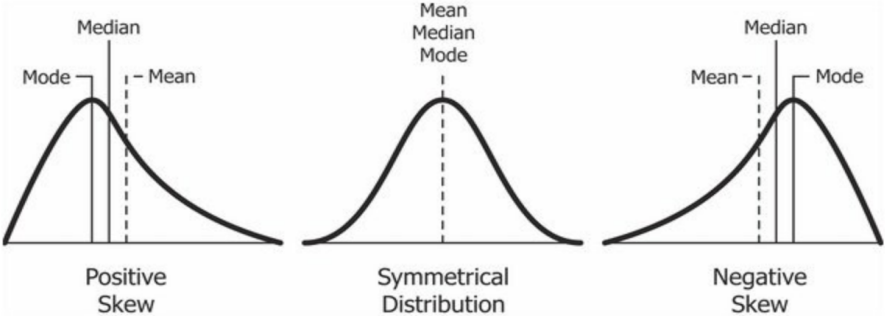
Types of distributions



How does skew effect measures of central tendency?

UNIVERSITY OF CANBERRA

- In a normal distribution (symmetrical), the mean = median = mode
- If there is a positive skew, then mode < median < mean
- If there is a negative skew then mean < median < mode



Statistics to describe a non-normal distribution

UNIVERSITY OF CANBERRA

- Use non-parametric descriptive statistics
- A reminder of these:
  - Minimum and maximum value
  - The range of values (maximum - minimum)
  - Percentiles
  - Quartiles
    - Q1
    - Q2
    - Q3
    - Interquartile ratio (Q3 - Q1)

Ways to ‘fix’ a non-normal distribution

UNIVERSITY OF CANBERRA

- You can use transformations to convert your data into a normal distribution
- This will allow you to do more powerful tests (i.e. parametric ones)
- However, you lose the original metric, which complicates interpretation

Things to keep in mind when you create a graph

- Have a clear purpose
- Make it as clear as you can
- Try to avoid clutter
- Allow for visual comparison

Steps to take

- Ask yourself, what is the purpose of the graph?
  - To make large amounts of data coherent?
  - To present many numbers in a small space?
  - To encourage the eye to make comparisons?
- Select the type of graph to use
- Draw and modify graph to be clear, non-distorting and well-labelled
  - This helps you maximise clarity and minimise clutter
  - You want to show the data - avoid distortion

Graphing one variable (univariate graphs)

UNIVERSITY OF CANBERRA

- Non-parametric (nominal or ordinal)
  - Bar graph
  - Pie chart
- Parametric (normally distributed interval or ratio)
  - Histogram
  - Stem and leaf plot
  - Box plot

Bar chart/graphs

- With a bar chart, you can compare the height of bars e.g. to see which is most common
- For discrete (i.e. data that can only take certain values) data (histogram for continuous) on x-axis
- Y-axis can show frequencies, percentages, or means
- To better show your data:
  - Collapse the x-axis if there are too many categories
  - Truncating the y-axis to exaggerate differences
  - Can add data labels (values for each bar)

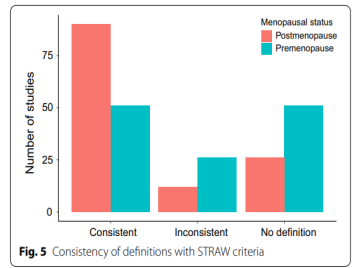


Fig. 5 Consistency of definitions with STRAW criteria

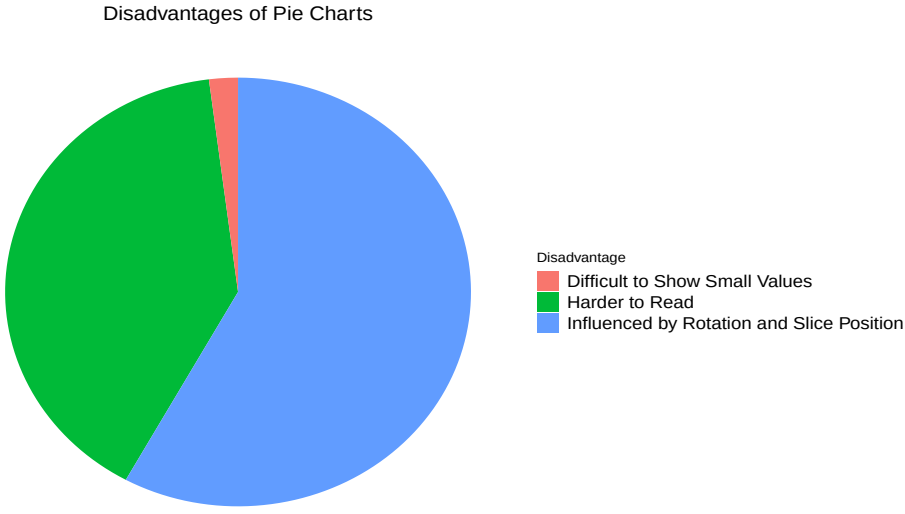
Ambikairajah et al. (2022)

Pie chart

UNIVERSITY OF CANBERRA

- A pie chart can display the same information
- The disadvantages
  - It is harder to read
  - Difficult to show small values or small differences
  - Rotating the chart and position of slices influences perception

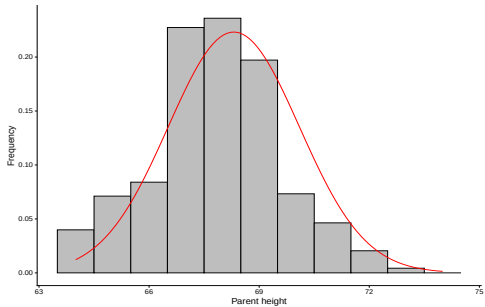
Pie chart



Histogram

UNIVERSITY OF CANBERRA

- For continuous data (likert with more than 5 categories, ratio data)
- The x-axis needs a happy medium number of categories
- The y-axis really matters - can be used to exaggerate findings



Stem and leaf plots

UNIVERSITY OF CANBERRA

- Use it to graph ordinal, interval and ratio data (if it is rounded to whole numbers)
- Contains all the data and presents it in a visual way similar to bar graph

Stem and leaf plots

UNIVERSITY OF CANBERRA

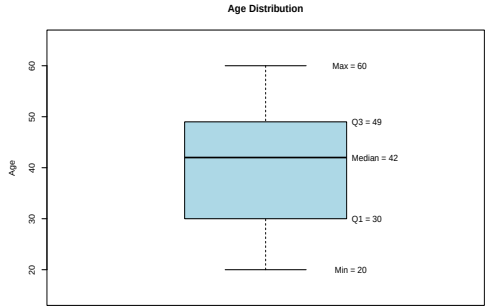
```
##  [1] 48 32 31 20 59 60 54 31 42 43 44 22 44 45 26 46 52 25 43 24 59 26 36 53 31
## [26] 34 60 56 29 32 49 59 24 26 58 27 40 44 24 44 49 55 42 51 46 22 25 29 30 35

##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 | 022444
##   2 | 55666799
##   3 | 0111224
##   3 | 56
##   4 | 022334444
##   4 | 566899
##   5 | 1234
##   5 | 568999
##   6 | 00
```

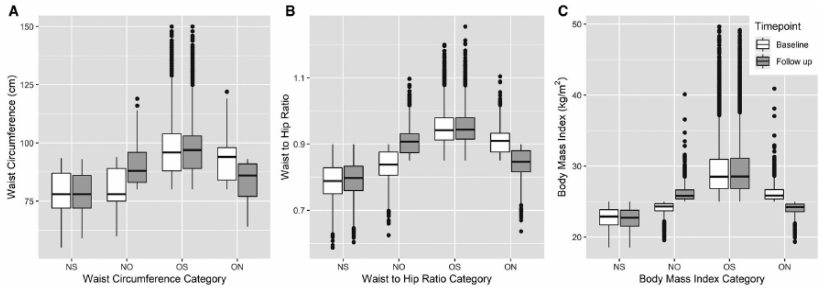
Box plot

UNIVERSITY OF CANBERRA

- Use this for interval and ratio data
- The box plot shows minimum and maximum values, as well as media, quartiles and outliers
- It is an alternative to the histogram
- Good for screening data, comparing variables
- Can get messy (information overload)



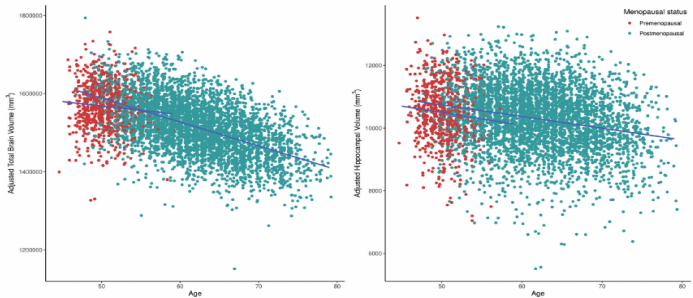
Box plot



Ambikairajah et al. (2020)

Line graph

- Alternative to histogram
- Implies continuity (e.g. time)
- Can show multiple lines for different information



Ambikairajah et al. (2021)

Graphical integrity

UNIVERSITY OF CANBERRA

Tufte & Graves-Morris (1983) on graphical integrity

- Lie factor (should $= 1$) $= \frac{\text{Size of effect in graph}}{\text{Size of effect in data}}$
- Lying in graphs is sometimes intentional and sometimes not
- Misleading graphs are those that:
  - Use area or perspective in misleading ways
  - Leave out important context
  - Lack taste/aesthetics

Summary

Things you should know

- The properties of a normal distribution
- Descriptive and graphing techniques for each level of measurement
- Difference between parametric/non-parametric and what kind of statistics you can use
- Skewness and kurtosis
- Interpreting graphs and understanding graphical integrity

Summary

UNIVERSITY OF CANBERRA

Types of questions

- Based on examination of a histogram
  - Which type of statistics should you use?
  - How would you describe the shape of the distribution
- Check descriptive statistics of variables in the dataset to identify
  - Standard deviations, mean, median, mode, minimum, maximum, sample size, skewness, kurtosis
- Which graphs to use, how to construct them
- Relationship between mean/median/mode under conditions of skew

Summary

- To make sure you understand the content we covered today, fill in the cells in this table for homework

Level	Properties	Examples	Descriptive statistics	Graphing
Nominal				
Ordinal				
Interval				
Ratio				

Next week - correlations

UNIVERSITY OF CANBERRA

- Covariation
- Purposes of correlation
- Linear correlation
- Types of correlation
- Interpreting correlation
- Assumptions and limitations

Contributions to this course

UNIVERSITY OF CANBERRA

Dr James Neill

Dr Samantha Stanley

Dr Jeroen van Boxtel

References I

Ambikairajah, A., Foxe, D., de Lange, A.-M. G., Carrick, J., Cheung, S. C., Srikanth, V. K., Hwang, Y. T., Ahmed, R. M., Burrell, J. R., & Piguet, O. (2025). A Bayesian analysis of diagnostic timelines across Alzheimer's disease, frontotemporal dementia, and other neurodegenerative conditions. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *17*(3), e70184. https://doi.org/10.1002/dad2.70184

Ambikairajah, A., Tabatabaei-Jafari, H., Hornberger, M., & Cherbuin, N. (2021). Age, menstruation history, and the brain. *Menopause*, *28*(2), 167–174. https://doi.org/10/ghtfz7

Ambikairajah, A., Tabatabaei-Jafari, H., Walsh, E., Hornberger, M., & Cherbuin, N. (2020). Longitudinal Changes in Fat Mass and the Hippocampus. *Obesity*, *28*(7), 1263–1269. https://doi.org/10/ggwqg5

References II

Ambikairajah, A., Walsh, E., & Cherbuin, N. (2022). A review of menopause nomenclature. *Reproductive Health*, *19*(1), 29. https://doi.org/10.1186/s12978-022-01336-7

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2). Graphics press Cheshire, CT.