PSYU1105:

Introduction to Psychology II

Statistics

Semester 2, 2020

TOPICS

Data input in stata
Summarising data
Fundamental concepts
T tests
Correlations
Chi square tests

Introduction

- Population - wider group you're interested in
- Sample - selection from the population
- Parameter - numeric summary of the population
- Statistic - numeric summary of the sample
- Unit of observation - level of which you’re interested in sampling
- Data - collection of information that has been recorded from your sample
- Variable - information you have collected that varies among participants
- Quantitative - numeric
- Qualitative - descriptive
- Discrete - categories on a scale
- Continuous - any point on a scale
- Nominal - unordered, categorical
    - Binary
- Ordinal - ordered categorical
- Interval - numeric scale with consistent differences between points
- Ratio - numeric scale with consistent differences between points AND absolute zero
- Measurement error - difference between the actual value of a phenomenon and the value of the data we collect about that phenomenon
- Independent variable - predict or explain a change in outcome
- Dependent variable - outcome
- Extraneous variable - another variable that's not IV or DV
- Confounding variable - extraneous variable that may explain the relationship between IV and DV
- Experimental designs
    - IV can cause a change in DV
- Observational designs
    - IV can be associated with or predict a change in DV
- Descriptive stats - describe the sample only
- Inferential stats - gather data from a sample and make generalisations back to the population
- Research hypothesis - developed from research question
- Statistical hypotheses
    - Null hypothesis - no differences between groups, no relationship
    - Alternate hypothesis - difference between groups with relationship

Stata and Data Input

- Storing data
    - Numeric - data entered is in the form of numbers only
    - String - data entered can be anything
- Need numeric-type data to perform statistical analyses
    - Even qualitative data will be entered as numeric
    - Categorical variables - coded and entered into the data as coded values
        - Tell program what coding scheme is
- Variable names rules
    - Can be uppercase or lowercase or mix (but case-sensitive!)
    - Max 32 characters
    - No spaces or symbols except letters, numbers or underscore
    - First character must be a letter (or underscore)
- Variable labels - longer descriptions of variable
- Value labels - coding scheme for categorical variables

Summarising Data

- Categorical data - discrete categories or groups
    - Frequency tables and bar chart / pie chart
- Numeric data - a score on a scale
    - Numeric summary statistics and a histogram
- Typicality
    - Most typical score - mean, mode, median
    - Mean Mean is represented by $\bar{x}\ (sample)$ or $\bar{\mu}\ (population)$
        - Advantages
            - most common
            - easy to calculate and understand
            - represents all of the data
        - Disadvantages
            - very affected by extreme scores
            - not always an actual score in the dataset
    - Median
        - Advantages
            - easy to find
            - not affected by extreme values
        - Disadvantages
            - may not best represent data if distribution is unbalanced
            - not always an actual score in the data
    - Mode
        - Advantages
            - always an actual value in the data
            - easy to find
        - Disadvantages
            - can be multiple
            - doesn't take into account all the data
- Variability
    - how spread out or varying or dispersed the scores are
    - Range - difference between biggest and smallest score
        - Advantages
            - Easy to calculate
        - Disadvantages
            - affected by extreme scores
            - doesn't take all the data into account
    - Interquartile range - difference between first and third quartile scores
        - Advantages
            - easy to calculate
            - not affected by extreme scores
        - Disadvantages

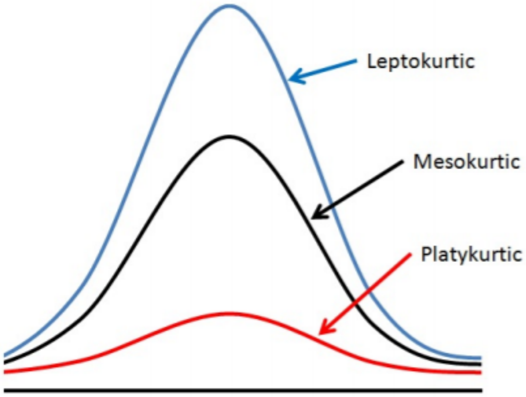
-
  -
    -
      - Not always an actual score in the data
  - Variance - average squared deviation of scores from the mean

    $$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

    - Advantages
      - easy(ish) to calculate
      - widely understood
      - takes all the data into account
    - Disadvantages
      - Can be affected by extreme scores
  - Standard deviation - average deviation of scores from the mean score

    $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

    - Advantages
      - easy(ish) to calculate
      - widely understood and used
      - takes all the data into account
    - Disadvantages
      - Can be affected by extreme scores
- Shape
  - Shape or pattern of distribution
    - Skew - symmetric or skewed
      - Left - negatively skewed
      - Right - positively skewed
    - Kurtosis - peaked/pointy or flat



  -
    - Lepto - leaping
    - Meso - middle
    - Platy - flat
  - Quantitative/numeric variables are often ‘normally distributed’

- Variability
- Unimodality
- Central tendency
- Symmetrical
- Mesokurtic