Research Methods

THE ESSENTIAL KNOWLEDGE BASE 2e

William M. Trochim
James P. Donnelly
Kanika Arora

11

Introduction to Data Analysis

11.1 Foundations of Data Analysis

- Data analysis has important implications for conclusion validity
- Data analysis has three main steps
  - Data preparation
  - Descriptive statistics
  - Inferential statistics

11.2 Conclusion Validity

- The extent to which conclusions or inferences regarding relationships between the major variables in your research are warranted
  - Related to internal validity, but is also independent of it
  - Remember, internal validity is concerned with causal relationships
  - Conclusion validity applies to all relationships, not just causal ones

11.2a Threats to Conclusion Validity

- Type I Error
  - You conclude there is a relationship when there is not
  - A false positive
- Type II Error
  - You conclude there is not a relationship when there is
  - A false negative

11.2a Threats to Conclusion Validity – Type I Error

- Finding a relationship when there is not one (or seeing things that aren't there)
  - Level of statistical significance (alpha)
    - Often set to .05, meaning five times out of a 100 you will have a false positive
  - Fishing and error rate problem
    - A problem that occurs as a result of conducting multiple analyses and treating each one as independent

11.2a Threats to Conclusion Validity – Type II Error

- Finding no relationship when there is one (or, missing the needle in the haystack)
  - Small effect size: the ratio of signal to noise
  - Sources of noise
    - Low reliability of measures
    - Poor reliability of treatment implementation
    - Random irrelevancies in the setting
    - Random heterogeneity of respondents
  - Weak signal: due to low strength intervention

11.2a Threats to Conclusion Validity – Other Issues

- Problems that can lead to either conclusion error
  - Violated assumptions of statistical tests
  - Violated assumptions in qualitative research

11.2b Improving Conclusion Validity

- Increase statistical power
  - Increase the sample size
  - Increase the level of significance
  - Increase the effect size
    - Good implementation of the program (e.g. by using trained researchers)

11.3 Data Preparation

- Logging the data in
- Checking the data for accuracy
- Entering data into a computer
- Transform the data
- Develop a database

11.3a Logging the Data

- Use a computer program to log the data as it comes in
  - MS Excel, Access
  - SPSS, SAS, Minitab, Datadesk
- Retain and archive original data records
  - Most researchers keep original data for 5-7 years
  - IRB requires data be stored securely and anonymously

11.3b Checking the Data for Accuracy

- Are the responses legible/readable?
- Are all important questions answered?
- Are the responses complete?
- Is all relevant contextual information included (for example, data, time, place, and researcher)?

11.3c Developing a Database Structure

- The database structure is the system you use to store the data for the study so that it can be accessed in subsequent data analyses
- Develop a codebook
  - Variable name, description, format (number, data, text), and location
  - Instrument/method of collection, data collected, respondent or group, and notes

11.3d Entering the Data into the Computer

- Enter the data into a spreadsheet, document, or statistical program
- Have a unique ID number for each record
  - Write this ID number on the paper the data was collected on
  - Each case (record) will be on its own row
  - Each column represents a variable
- Double-entry

11.3e Data Transformations

- Missing values
- Item reversals
- Scale and subscale totals
- Categories
- Variable transformations

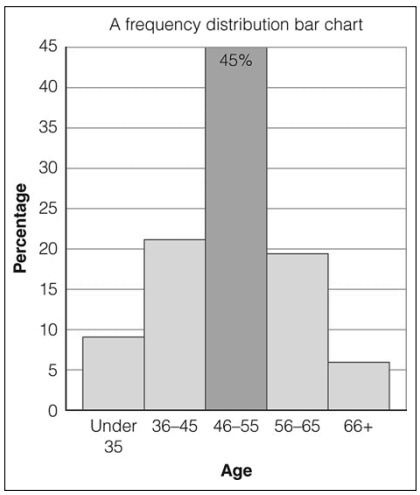
11.4 Descriptive Statistics

- Statistics used to describe the basic features of the data in a study
- Three main measures
  - The distribution
  - The central tendency
  - The dispersion (or variability)

11.4a The Distribution

A frequency distribution in table form

Category	Percent
Under 35	9%
36–45	21%
46–55	**45%**
56–65	19%
66+	6%



11.4b Central Tendency

- Mean
  - The average
- Median
  - The centermost score (50th percentile)
- Mode
  - The most frequently occurring score
- If the distribution is normal, the mean, median, and mode are all equal

Valori caratteristici di tendenza centrale e di dispersione per tipi di variabili

TAB. A

	TIPO DI VARIABILE	VALORI CARATTERISTICI DI TENDENZA CENTRALE	NOTE	VALORI CARATTERISTICI DI DISPERSIONE	NOTE	RAPPRESENTAZIONI GRAFICHE (della distribuzione di frequenza)
	CATEGORIALE NON ORDINATA (Es. Genere)	MODA (Categoria più frequente)	M F MODA	Sq= $\sum_{j=1}^{k} p_j^2$	$\frac{1}{k}$ Max dispersione 1 (min. dispersione)	Diag. barre
	CATEGORIALE ORDINATA (Es. Livello di istruzione)	MEDIANA (Valore che divide la distribuzione in due parti uguali)	Non subisce l'influenza dei valori estremi	d= $\frac{4\sum_{h-1}^{k-1} p_h(1\text{-}ph)}{K\text{-}1}$	0 (MINIMA DISPERSIONE) 1 (MAX DISPERSIONE)	Diag. barre
CARDINALI	INTERVALLI (Es. Data di nascita)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	SCARTO TIPO $s = \frac{\sqrt{\sum (xi - \bar{x})^2}}{n}$	0 (MINIMA DISPERSIONE + HIGHEST (MAX DISPERSIONE	Istog
	Metriche/enumerate (Es. Peso/N. libri posseduti)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	COEFFICIENTE DI VARIAZIONE cv= s/ $\bar{x}$	0 (MINIMA DISPERSIONE) + HIGHEST (MAX DISPERSIONE)	Istog
	QUANTITA' (Es. Reddito)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	COEFFICIENTE DI CONCENTRAZIONE	0 (max dispersione) 1 (minima dispersione)	Istog

11.4c Dispersion or Variability

- Dispersion: spread of values around the central tendency
  - Range: highest score minus the lowest score
  - Standard deviation: variability of the scores around their average in a single sample
    - 68% of scores will fall within 1 SD of the mean
    - 95% of scores will fall within 2 SD of the mean
    - 98% of scores will fall within 3 SD of the mean
  - Variance: spread of scores around the mean, in squared units

11.4d Correlation

- A single number that describes the degree of relationship between two variables
- Positive correlations
  - Always fall between 0 and 1
  - As one value increases, so does the other
- Negative correlations
  - Always fall between -1 and 0
  - As one value increases, the other decreases

11.4d Correlation

- You can test to see if a correlation happened by chance—this is the correlation's significance
- You can also construct a correlation matrix

Table 11.5 Hypothetical correlation matrix for ten variables.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	
C1	1.000									
C2	.274	1.000								
C3	−.134	−.269	1.000							
C4	.201	−.153	.075	1.000						
C5	−.129	−.166	.278	−.011	1.000					
C6	−.095	.280	−.348	−.378	−.009	1.000				
C7	.171	−.122	.288	.086	.193	.002	1.000			
C8	.219	.242	−.380	−.227	−.551	.324	−.082	1.000		
C9	.518	.238	.002	.082	−.015	.304	.347	−.013	1.000	
C10	.299	.568	.165	−.122	−.106	−.169	.243	.014	.352	1.000

11.4d Types of Correlations

- Pearson Product Moment Correlation: used when both variables are **interval level**
- Spearman Rank Order Correlation (rho): used when both variables are **ordinal level**
- Kendall Rank Order Correlation (tau): used when both variables are **ordinal level**
- Point-Biserial Correlation: used when one variable is continuous, and one is dichotomous

Discuss and Do

- Why should you be cautious when interpreting a correlation?
- Discuss the relationship between a frequency distribution, measures of central tendency, and variability
- Find wild codes (if any) in the d2 dataset
- Choose 1 variable for each *level of measurement* and calculate the appropriates central tendency measures