Intermediate IBM SPSS

Understanding Your Data (Descriptive Statistics, Graphs and Custom Tables)

Pawel Skuza
Statistical Consultant
eResearch@Flinders / Central Library

Flinders UNIVERSITY
inspiring achievement

- Please note that the workshop is aimed to be a brief introduction to the topic and this PowerPoint is primarily designed to support the flow of the workshop. It cannot be seen as either an exclusive or exhaustive resource on the statistical concepts which are introduced in this course. You are encouraged to refer to peer-reviewed books or papers that are listed throughout the presentation.
- It is acknowledged that a limited number of slides have been adapted from presentations produced by the previous statistical consultant (Kylie Lange) and a colleague with whom I worked with in the past (Dr Kelvin Gregory).

Flinders UNIVERSITY
inspiring achievement

Statistical Consulting Website

http://www.flinders.edu.au/library/research/eresearch/statistics-consulting/

or go to Flinders University Website →A-Z Index →S →Statistical Consultant

Introductory Level
- Introduction to IBM SPSS
- Introduction to Statistical Analysis

IBM SPSS - Intermediate Level
- Understanding Your Data (Descriptive Statistics, Graphs and Custom Tables)
  - Correlation and Multiple Regression
    - Logistic Regression and Survival Analysis
  - Basic Statistical Techniques for Difference Questions
    - Advanced Statistical Techniques for Difference Questions
      - Longitudinal Data Analysis - Repeated Measures ANOVA
  - Categorical Data Analysis

IBM SPSS - Advanced Level
- Structural Equation Modelling using Amos
- Linear Mixed Models
  - Longitudinal Data Analysis - Mixed and Latent Variable Growth Curve Models
- Scale Development
- Complex Sample Survey Design / ABS and FaHCSIA Confidentialised Datasets

Flinders UNIVERSITY inspiring achievement

What you will learn:

- Summarising categorical and continuous data
- Working with 'Custom Tables' Module
- Using graphs to describe and explore the data
- Assessing normality and outliers
- Dealing with missing data
- Use of syntax

Flinders UNIVERSITY inspiring achievement

??? SPSS / PASW / IBM SPSS ???

- In late 2009 SPSS Inc. was taken over by IBM Company and the software changed its official name twice over the period of one year. From SPSS it was relabelled to PASW (Predictive Analytics Software) and later to IBM SPSS. Consequently, there may be books, online resources, etc. that use either of those different names but in fact refer to the same software.

- **SPSS**
  - Statistical Package for the Social Sciences

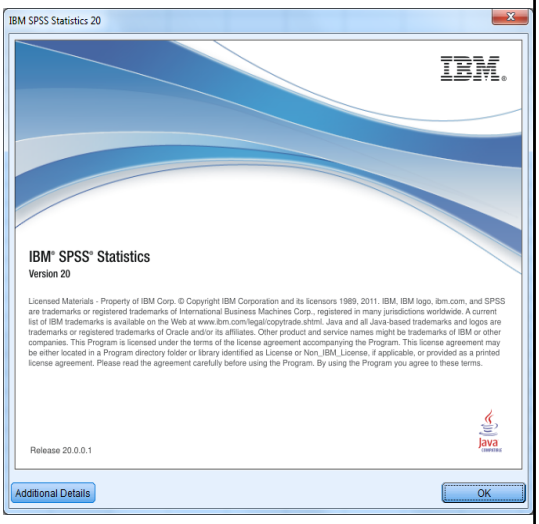
- **PASW**
  - Predictive Analytics Software

- **IBM SPSS Statistics**

SPSS / PASW / IBM SPSS

(1) How to check?

START SOFTWARE → HELP → ABOUT



(2) How to cite?

(Examples with APA Style)

- **SPSS Inc. Released 2007. SPSS for Windows, Version 16.0. Chicago, SPSS Inc.**
- **SPSS Inc. Released 2008. SPSS Statistics for Windows, Version 17.0. Chicago: SPSS Inc.**
- **SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.**
- **IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.**
- **IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.**
- **IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.**

Pawel Skuza 2013

IBM SPSS on Flinders University

- Flinders University has licence for number of IBM SPSS products (versions 19, 20, 21) covering following modules:
  - IBM SPSS Statistics Base
  - IBM SPSS Regression
  - IBM SPSS Advanced Statistics
  - IBM SPSS Complex Samples
  - IBM SPSS Categories
  - IBM SPSS Exact Tests
  - IBM SPSS Missing Values
  - IBM SPSS Forecasting
  - IBM SPSS Custom Tables
  - IBM SPSS Conjoint
  - IBM SPSS Statistics Programmability Extension

  and AMOS
- For details explaining various modes of obtaining access to the software go to

http://www.flinders.edu.au/library/research/eresearch/statistics-consulting/spss-licenses-and-technical-support/licenses-for-university-and-home.cfm

What is 'Statistics' ?

sta·tis·tics (st -t s t ks)

n.

1. *(used with a sing. verb)* The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.

2. *(used with a pl. verb)* Numerical data.

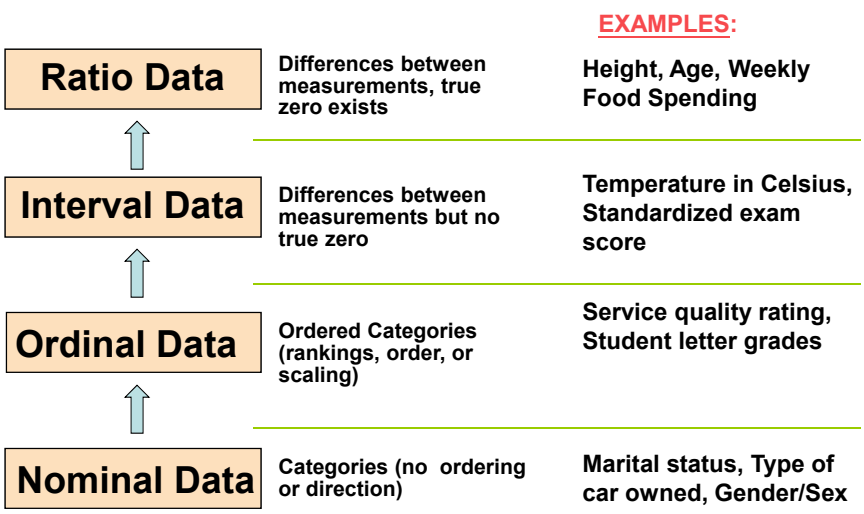
http://www.thefreedictionary.com/Statistics

Harvard President Lawrence Lowell wrote in 1909 that statistics, ***"like veal pies, are good if you know the person that made them, and are sure of the ingredients".***

http://en.wikipedia.org/wiki/Statistics

Flinders UNIVERSITY inspiring achievement

Levels of Measurement and Measurement Scales



Flinders UNIVERSITY inspiring achievement

Real World Data

- Data can be dirty
  - Incomplete data
    - Missing attributes
    - Missing attribute values
    - Only aggregated data
  - Inconsistent data
    - Different coding
    - Different naming conventions
    - Impossible values
    - Out-of-range values
  - Noisy data
    - Errors
    - Outliers
    - Inaccurate values
- Need to pre-process the data before using for analysis

Flinders UNIVERSITY inspiring achievement

Common Data Entry Errors

- Wrong data but within range
  - The marital status of married person is entered as a single
  - Both single and married are legal
    - This type of errors checked by using double entry method
- Wrong data and out of range
  - If 1 stands for male and 2 stands for female, then the value of 3 represents erroneous data
  - Frequency distribution procedures flagged these cases

Flinders UNIVERSITY inspiring achievement

Common Data Entry Errors

- False logic(Consistency)
  - A 25 years old respondent is reported as having experience of government service in 30 years.
- Missing data
  - Missing data codes for items such as “Not applicable” and “refuse to answer” have not been pre coded in the questionnaire, even though they should have been
  - Need to find these cases replace with the appropriate data code
    - 8=refuse to answer
    - 9=missing

Flinders UNIVERSITY inspiring achievement

Screening and cleaning the data

- Exercise 1
  Data set describing the survival status of individual passengers on the Titanic. More information about the data can be found in here:
  http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html

- Use “Save file as” option

Example:

My data_2008_08_19,
My data_2008_08_20

- Keeping diary of undertaken analyses - Syntax

Flinders UNIVERSITY inspiring achievement

- Descriptive statistics
  - Summarising and presenting data
    - Measures of Central Tendency
    - Measures of Dispersion or Variability

Flinders UNIVERSITY inspiring achievement

Central Tendency

- To summarise the “location” of a distribution
- Mode
- Median
- Mean

Flinders UNIVERSITY inspiring achievement

Measures of Central Tendency

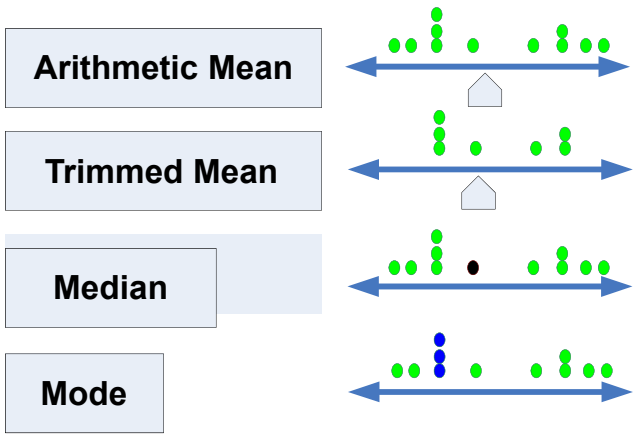
- Three common measures
  - Mode
    - The mode of a data set is the value that occurs with the most frequency
  - Median
    - The median is the central of an ordered distribution
      - Order the data from smallest to largest
      - For an odd number of data values in the distribution
        - » Median=middle value of the data
      - For an even number of data values in the distribution
        - » Median=(sum of the middle two values)/2
  - (Arithmetic) mean or average
    - Mean is the sum of all the entries divided by the number of entries

Trimmed Mean

- The trimmed mean is produced by discarding the most extreme values
  - In SPSS, the trimmed mean is calculated by discarding the top and bottom 5% of the cases
- The Trimmed Mean procedure is available through the Explore procedure
- If the Trimmed Mean and Mean are similar it suggests that there are few, if any, influential outliers

Flinders UNIVERSITY inspiring achievement

Measures of Central Tendency



Flinders UNIVERSITY inspiring achievement

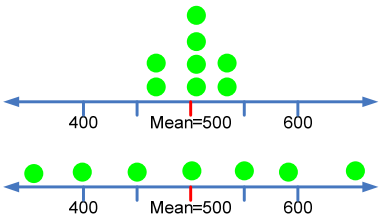
Variability

- To summarise the “spread” or “dispersion” of a distribution
- Low variability => scores are similar
- High variability => scores differ
- Range
- Interquartile range (IQR)
- Standard deviation / Variance

Flinders UNIVERSITY inspiring achievement

The Limitation of Point Estimates

- The sample median and sample mean estimate the corresponding center points of a population
- Such estimates are called **point estimates**



Flinders UNIVERSITY inspiring achievement

Pawel Skuza 2013

Range

- Two definitions used
  - Exclusive range

    $X_{max} - X_{min}$

    - **This is the most commonly used way of calculating the range**

Flinders UNIVERSITY inspiring achievement

Deviation Score

- Remember
  - The arithmetic mean uses information about every observation
- A good measure of variation should also summarize how much each observation deviates from the measure of central tendency
- The deviation score is the distance a score is from the arithmetic mean

$$d_i = X_i - \overline{X}$$

Flinders UNIVERSITY inspiring achievement

Variance and Standard Deviation

- The variance is the mean squared deviation from the average
- There are two formulas
  - One for populations
  - One for samples

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \qquad s^2 = \frac{\sum (X-\bar{X})^2}{n-1}$$

$$\sigma^2 = \frac{\sum (d)^2}{N} \qquad s^2 = \frac{\sum (d)^2}{n-1}$$

Flinders UNIVERSITY inspiring achievement

Variance and Standard Deviation

- The sample variance formula has the (N-1) divisor
  - This produces an unbiased estimate of the population variance
- The standard deviation is the positive square root of the variance

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}} \qquad s = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}$$

$$\sigma = \sqrt{\frac{\sum (d)^2}{N}} \qquad s = \sqrt{\frac{\sum (d)^2}{n-1}}$$

Flinders UNIVERSITY inspiring achievement

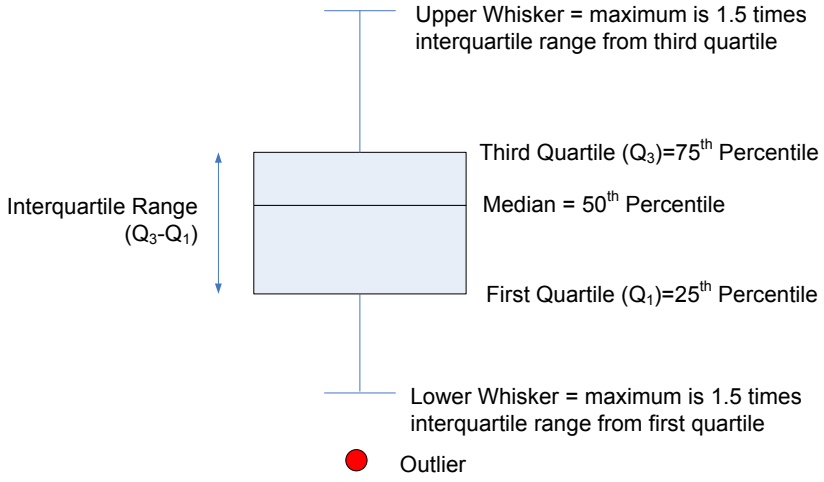
Pawel Skuza 2013

Quartiles and Percentiles

- The distribution can be divided into four equal regions
  - Based upon the cases
- So
  - First quartile (Q_1) is the 25th percentile
    - 25% or one quarter of the cases fall below this score
  - Second quartile is the 50th percentile or median
    - 50% or one quarter of the cases fall below this score
  - Third quartile is the 75th percentile
    - 75% or one quarter of the cases fall below this score

Flinders UNIVERSITY inspiring achievement

Revision: Box and Whisker Plot



Flinders UNIVERSITY inspiring achievement

Outliers

- Outliers are observations that deviate significantly from the majority of observations
- Outliers can lead to
  - Model misspecification
  - Biased parameter estimation
  - Incorrect analysis results

Flinders UNIVERSITY
inspiring achievement

How to report?

- Altman, D. G. (1980). Statistics and ethics in medical research. VI - Presentation of results. *British Medical Journal, 281*(6254), 1542-1544.
- Lang, T. A., & Secic, M. (2006). *How to report statistics in medicine : annotated guidelines for authors, editors, and reviewers* (2nd ed.). New York: American College of Physicians.
- Thabane, L., & Akhtar-Danesh, N. (2008). Guidelines for reporting descriptive statistics in health research. *Nurse researcher, 15*(2), 72-81.
- Whitley, E., & Ball, J. (2002). Statistics review 1: Presenting and summarising data. *Critical Care, 6*(1), 66-71.

inspiring achievement

	Nominal	*Dichotomous*	*Ordinal*	*Normal*
Frequency Distribution	Yes[a]	Yes	Yes	OK[b]
Bar Chart	Yes	Yes	Yes	OK
Histogram	No[c]	No	OK	Yes
Frequency Polygon	No	No	OK	Yes
Box and Whiskers Plot	No	No	Yes	Yes
Central Tendency				
Mean	No	OK	Of ranks, OK	Yes
Median	No	OK = Mode	Yes	OK
Mode	Yes	Yes	OK	OK
Variability				
Range	No	Always 1	Yes	Yes
Standard Deviation	No	No	Of ranks, OK	Yes
Interquartile range	No	No	OK	OK
How many categories	Yes	Always 2	OK	Not if truly continuous
Shape				
Skewness	No	No	Yes	Yes

[a]Yes means a good choice with this level of measurement.
[b]OK means OK to use, but not the best choice at this level of measurement.
[c]No means not appropriate at this level of measurement.

Reproduced from Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2007). *SPSS for introductory statistics : use and interpretation* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

inspiring achievement

Summary Statistics

Categorical variables in SPSS

- Analyse > Descriptive Statistics
  - Frequencies (Statistics: quartiles, percentiles)
  - Explore (median, percentiles)
- Graphs
  - Bar chart

Flinders UNIVERSITY
inspiring achievement

Summary Statistics

Continuous variables in SPSS

- Analyse > Descriptive Statistics
  - Descriptives
  - Explore
- Graphs in Explore
  - Histogram
  - Boxplot

Flinders UNIVERSITY inspiring achievement

Exercise 2

Describing categorical data
Summarising continuous data

- Data – Exercise_2.sav

Simplified data from PISA 2003 Study - Australia
(The **P**rogramme for **I**nternational **S**tudents **A**ssessment)

http://www.pisa.oecd.org

Flinders UNIVERSITY inspiring achievement

Exercise 3

Checking Normality

- Data – Exercise_3.sav --- Sample of Chicago high schools
- Research question – Are variables “Average class size”, Reading at national norm %, Limited English % normally distributed?

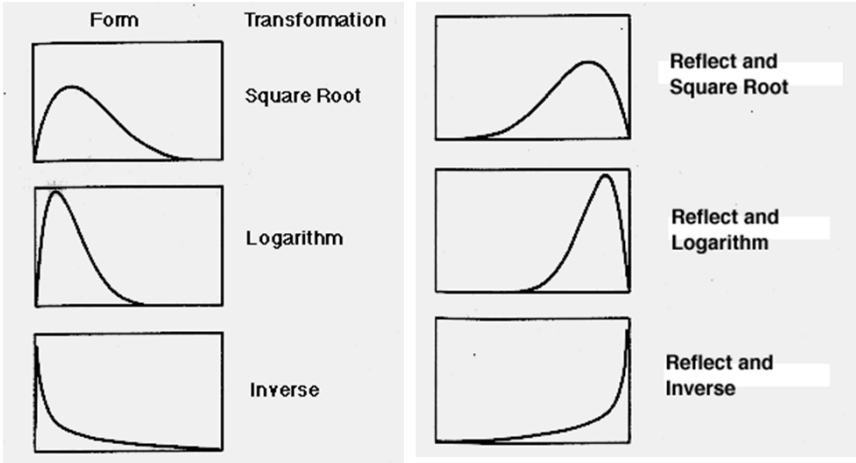
Flinders UNIVERSITY inspiring achievement

Transformations

- Why?
  - A lot of statistical tests and methods are based around the normal distribution assumption
  - Often skewness and heterogeneity of variances is a problem
- Advantages
  - Allows the use of standard methods
  - Allows the use of more powerful methods
- Disadvantages
  - Converts measurements into a foreign unit
- Using statistical nonparametric tests may be an alternative

Flinders UNIVERSITY inspiring achievement

Transformations



Flinders UNIVERSITY inspiring achievement

Exercise 4

Missing Values Analysis

- Data – Exercise_4.sav --- Sample of passengers from the Titanic

Flinders UNIVERSITY inspiring achievement

Dealing with Missing Data

- Handling missing data
  - Ignore record (not advisable)
  - Fill in with attribute mean or median (not advisable)
  - Fill in with most likely value based upon imputation process (various approaches available – see below references for more information)

Flinders UNIVERSITY inspiring achievement

Missing Data - References

- Abraham, W. T., & Russell, D. W. (2004). Missing data: A review of current methods and applications in epidemiological research. *Current Opinion in Psychiatry, 17*(4), 315-321.
- Allison, P. D. (2003). Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37.
- Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92.
- Enders, Craig K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Everitt, Brian. (2003). Missing Values, Drop-outs, Compliance and Intention-to-Treat. In B. Everitt (Ed.), *Modern medical statistics : A practical guide* (pp. 46-66). London: Arnold
- Fitzmaurice, Garrett. (2008). Missing data: implications for analysis. *Nutrition, 24*(2), 200-202. doi: DOI: 10.1016/j.nut.2007.10.014
- McKnight, Patrick E. (2007). *Missing data : a gentle introduction*. New York: Guilford Press.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556.
- Streiner, D. L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry, 47*(1), 68-75.

Flinders UNIVERSITY inspiring achievement

Time permitting
Additional Exercises 5 & 6

Multiple Response Set

- Data – Exercise_5.sav

Merging files

- Merge ITEMS_1.sav
- Merge ITEMS_2.sav

Flinders UNIVERSITY
inspiring achievement

SPSS – BOOKS (**Hard copies)**

- **Chapter 3 in Allen, Peter James, & Bennett, Kellie. (2012). SPSS statistics : a practical guide : version 20. South Melbourne, Vic.: Cengage Learning Australia.**
- **Chapters 3, 4, 9,10,11 in Argyrous, George. (2011). Statistics for research : with a guide to SPSS (3rd ed.). Los Angeles: Sage.**
- **Chapter 2 in Landau, Sabine, & Everitt, Brian. (2004). A handbook of statistical analyses using SPSS. Boca Raton: Chapman & Hall/CRC.**
- **Chapters 4 & 5 in Kinnear, Paul R., & Gray, Colin D. (2009). PASW statistics 17 made simple (replaces SPSS statistics 17). London ; New York: Psychology Press.**
- **Chapters 4 & 5 in in Field, Andy P. (2009). Discovering statistics using SPSS : (and sex, drugs and rock 'n' roll) (3rd ed.). Los Angeles: SAGE Publications.**
- **Chapters 4, 5, 7, 9 & Appendix A - Norušis, M. J. (2008). SPSS 16.0 [or later versions] Guide to Data Analysis. Upper Saddle River, NJ: Prentice Hall.**

SPSS – BOOKS (**Online copies)**

Hard copies and online versions

- **Chapters 5, 6, 7 in Pallant, Julie. (2010). SPSS survival manual a step by step guide to data analysis using SPSS (4th ed.). Maidenhead: Open University Press/McGraw-Hill.**
- **Chapters 3, 4 in Morgan, George A. (2011). IBM SPSS for introductory statistics : use and interpretation (4th ed.). New York: Routledge.**

Online versions

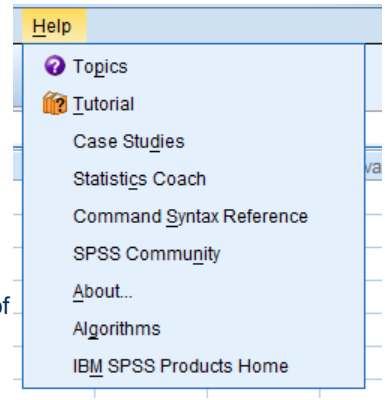
- **Chapter 5 in Bryman, Alan, & Cramer, Duncan. (2011). Quantitative data analysis with IBM SPSS 17, 18 & 19 : a guide for social scientists. Hove ; New York: Routledge.**
- **Chapters 5 & 6 in Marston, Louise. (2010). Introductory statistics for health and nursing using SPSS. Los Angeles: SAGE.**
- **Chapter 3 in Larson-Hall, Jenifer. (2010). A guide to doing statistics in second language research using SPSS**

THEORY - Marsh, Catherine, & Elliott, Jane. (2008). *Exploring data : an introduction to data analysis for social scientists* (2nd ed.). Cambridge ; Malden, Mass.: Polity.

SPSS – Help and Resources

- SPSS has a range of help options available
  - Topics
    - Used to find specific information
  - Tutorial
    - Find illustrated, step-by-step instructions for the basic features
  - Case studies
    - Hands-on examples of various types of statistical procedures
  - Statistics coach
    - To help you find the procedure you want to use

And manuals available online - http://www-01.ibm.com/support/docview.wss?uid=swg27021213



SPSS – **Online tutorials and resources**

(!!! Please keep in mind that usually online resources are not academically peer reviewed. Despite many of them being of high quality as well as being very useful from educational point of view, they shouldn't be treated as a completely reliable and academically sound references)

- **Statnotes: Topics in Multivariate Analysis, by G. David Garson**
  http://www.statisticalassociates.com/
- **UCLA Institute for Digital Research and Education - SPSS Starter Kit**
  http://www.ats.ucla.edu/stat/spss/sk/default.htm
- **Getting Started with SPSS for Windows by John Samuel, Indiana University**
  http://www.indiana.edu/~statmath/stat/spss/win/index.html
- **Companion Website for the 3rd edition of Discovering Statistics Using SPSS by Andy Field**
  http://www.uk.sagepub.com/field3e/SPSSFlashmovieslect.htm
- **SPSS for Windows and Amos tutorials by Information Technology Services, University of Texas**
  http://ssc.utexas.edu/software/software-tutorials#SPSS
- **Journey in Survey Research by John Hall**
  http://surveyresearch.weebly.com/index.html

SPSS – Help and Resources

- **Online SPSS FORUMS**

(!!! Please keep in mind that usually online resources are not academically peer reviewed. Despite many of them being of high quality as well as being very useful from educational point of view, they shouldn't be treated as a completely reliable and academically sound references.

!!! Suggestions / Guidance found on forums should be especially treated very doubtfully, yet they may point to more reliable academic resources and be somewhat of help.

Archives of SPSSX-L@LISTSERV.UGA.EDU – List Serve that is endorsed by IBM SPSS
http://www.listserv.uga.edu/archives/spssx-l.html

Other forums

http://groups.google.com/group/comp.soft-sys.stat.spss/topics?gvc=2
http://www.spssforum.com/

THANK YOU

Please provide us with your feedback by completing the short survey.

Flinders UNIVERSITY
inspiring achievement