Intermediate IBM SPSS –

Categorical Data Analysis

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 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).

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**

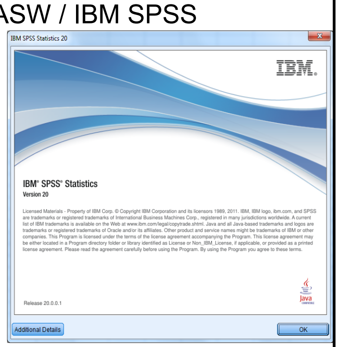
??? SPSS / PASW / IBM SPSS ???

- 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.**

Levels of Measurement and Measurement Scales

		EXAMPLES:
Ratio Data	Differences between measurements, true zero exists	Height, Age, Weekly Food Spending
Interval Data	Differences between measurements but no true zero	Temperature in Celsius, Standardized exam score
Ordinal Data	Ordered Categories (rankings, order, or scaling)	Service quality rating, Student letter grades
Nominal Data	Categories (no ordering or direction)	Marital status, Type of car owned, Gender/Sex

MEASUREMENT

Selection of statistical methods

Example 1

Figure 4.11 from Dancey, C. P., & Reidy, J. (2004). Statistics without maths for psychology : using SPSS for Windows (3rd ed.). New York: Prentice Hall.

Example 2

Table from Pallant, J. (2007). SPSS Survival Manual : A step by step guide to data analysis using SPSS for Windows (SPSS Version 15) (3rd ed.). Maidenhead, Berkshire. U.K. ; New York, NY: Open University Press.

Example 3

Flowchart from http://gjyp.nl/marta/Flowchart%20(English).pdf

Similar ones in other resources …

Pawel Skuza 2013

Table 5.5. *Interpretation of the Strength of a Relationship (Effect Sizes)*

	The *d* Family[a]	The *r* Family[b]			Risk Potency
General Interpretation of the Strength of a Relationship	*d*	*r* and ϕ	*R*	η (eta)	*RD* (%)
Much larger than typical	$\geq \|1.00\|$[c, d]	$\geq \|.70\|$	$\|.70\|+$	$\|.45\|+$	≥ 52
Large or larger than typical	$\|.80\|$	$\|.50\|$	$\|.51\|$	$\|.37\|$	43
Medium or typical	$\|.50\|$	$\|.30\|$	$\|.36\|$	$\|.24\|$	28
Small or smaller than typical	$\|.20\|$	$\|.10\|$	$\|.14\|$	$\|.10\|$	11

[a] *d* values can vary from 0.0 to + or −infinity, but *d* greater than one is relatively uncommon.
[b] *r* family values can vary from 0.0 to + or −1.0, but except for reliability (i.e., same concept measured twice), *r* is rarely above .70. In fact, some of these statistics (e.g., phi) have a restricted range in certain cases; that is, the maximum phi may be less then 1.0.
[c] We interpret the numbers in this table as a range of values. For example, a *d* greater than .90 (or less than −.90) would be described as "much larger than typical," a *d* between say .70 and .90 would be called "larger than typical," and *d* between say .60 and .70 would be "typical to larger than typical." We interpret the other three columns similarly.
[d] Note that | | indicates absolute value of the coefficient. The absolute magnitude of the coefficient, rather than its sign, is the information that is relevant to effect size. *R* and eta usually are calculated by taking the square root of a squared value, so that the sign usually is positive.

Reproduced from (Leech, Barrett, & Morgan, 2008, p. 81)

Pawel Skuza 2013

Exercise 1

Comparisons of Column Proportions (z-test)

- Please open - PISA_2000_Part1b.sav

  Simplified data from PISA 2000 Study – Few countries selected
  (The **P**rogramme for **I**nternational **S**tudents **A**ssessment)

  http://www.pisa.oecd.org

Flinders University inspiring achievement

Pawel Skuza 2013

Comparisons of Column Proportions (z-test)

- Page 27 in Kanji, G. K. (2006). *100 statistical tests* (3rd ed.). London ; Thousand Oaks, Calif.: Sage Publications.
- Pages 637 in Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton: Chapman & Hall/CRC.
- Page 25 in Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

13

Contingency Tables

- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.

Flinders University inspiring achievement

Pawel Skuza 2013

Contingency Table Example

Left-Handed vs. Gender

Dominant Hand: Left vs. Right

Gender: Male vs. Female

- 2 categories for each variable, so called a 2 x 2 table
- Suppose we examine a sample of size 300

Flinders University inspiring achievement

Pawel Skuza 2013

Contingency Table Example

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

Gender	Hand Preference: Left	Hand Preference: Right	
Female	12	108	120
Male	24	156	180
	36	264	300

χ^2 Test for the Difference Between Two Proportions

H_0: $\pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)

H_1: $\pi_1 \neq \pi_2$ (The two proportions are not the same – Hand preference is not independent of gender)

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

- where:
  f_o = observed frequency in a particular cell
  f_e = expected frequency in a particular cell if H_0 is true

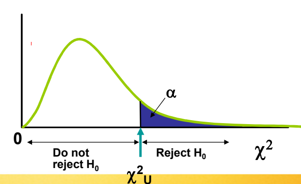
  χ^2 for the 2 x 2 case has 1 degree of freedom

(Assumed: each cell in the contingency table has expected frequency of at least 5)

Decision Rule

The χ^2 test statistic approximately follows a chi-squared distribution with one degree of freedom

Decision Rule:
If $\chi^2 > \chi^2_U$, reject H_0, otherwise, do not reject H_0



Computing the Average Proportion

The average proportion is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{X}{n}$$

120 Females, 12 were left handed

180 Males, 24 were left handed

Here:

$$\bar{p} = \frac{12 + 24}{120 + 180} = \frac{36}{300} = 0.12$$

i.e., the proportion of left handers overall is 0.12, that is, 12%

Finding Expected Frequencies

- To obtain the expected frequency for left handed females, multiply the average proportion left handed ($\bar{p}$) by the total number of females
- To obtain the expected frequency for left handed males, multiply the average proportion left handed ($\bar{p}$) by the total number of males

If the two proportions are equal, then

P(Left Handed | Female) = P(Left Handed | Male) = .12

i.e., we would expect (.12)(120) = 14.4 females to be left handed
(.12)(180) = 21.6 males to be left handed

Observed vs. Expected Frequencies

Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

The Chi-Square Test Statistic

Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

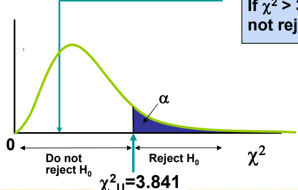
The test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

Decision Rule

The test statistic is $\chi^2 = 0.7576$, χ^2_U with 1 d.f. $= 3.841$



Decision Rule:
If $\chi^2 > 3.841$, reject H_0, otherwise, do not reject H_0

Here,
$\chi^2 = 0.7576 < \chi^2_U = 3.841$, so we do not reject H_0 and conclude that there is not sufficient evidence that the two proportions are different at $\alpha = 0.05$

Exercise 2

- Please open - PISA_2000_Part1b_AUSTRALIA.sav

Simplified data from PISA 2000 Study – Few countries selected
(The **P**rogramme for **I**nternational **S**tudents **A**ssessment)

http://www.pisa.oecd.org

Exercise 3

- Please start a new data file.

		Sex - Q3: Female	Sex - Q3: Male
		Count	Count
Internet - Q21d	Yes	92	97
	No	51	44

χ^2 Test of Independence

- **Similar to the χ^2 test for equality of more than two proportions, but extends the concept to contingency tables with r rows and c columns**

H_0: The two categorical variables are independent
(i.e., there is no relationship between them)
H_1: The two categorical variables are dependent
(i.e., there is a relationship between them)

χ^2 Test of Independence

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

- **where:**
  f_o = observed frequency in a particular cell of the r x c table
  f_e = expected frequency in a particular cell if H_0 is true

χ^2 for the r x c case has (r-1)(c-1) degrees of freedom

(Assumed: each cell in the contingency table has expected frequency of at least 1)

Expected Cell Frequencies

- **Expected cell frequencies:**

$$f_e = \frac{\text{row total} \times \text{column total}}{n}$$

Where:
row total = sum of all frequencies in the row
column total = sum of all frequencies in the column
n = overall sample size

Decision Rule

- **The decision rule is**

If $\chi^2 > \chi^2_U$, reject H_0, otherwise, do not reject H_0

Where χ^2_U is from the chi-squared distribution with (r – 1)(c – 1) degrees of freedom

Example

- **The meal plan selected by 200 students is shown below:**

Class Standing	Number of meals per week			Total
	20/week	10/week	none	
Fresh.	24	32	14	70
Soph.	22	26	12	60
Junior	10	14	6	30
Senior	14	16	10	40
Total	70	88	42	200

Example

- **The hypothesis to be tested is:**

H_0: Meal plan and class standing are independent
(i.e., there is no relationship between them)
H_1: Meal plan and class standing are dependent
(i.e., there is a relationship between them)

Example: Expected Cell Frequencies

Observed:

Class Standing	Number of meals per week			Total
	20/wk	10/wk	none	
Fresh.	24	32	14	70
Soph.	22	26	12	60
Junior	10	14	6	30
Senior	14	16	10	40
Total	70	88	42	200

Expected cell frequencies if H_0 is true:

Class Standing	Number of meals per week			Total
	20/wk	10/wk	none	
Fresh.	24.5	30.8	14.7	70
Soph.	21.0	26.4	12.6	60
Junior	10.5	13.2	6.3	30
Senior	14.0	17.6	8.4	40
Total	70	88	42	200

Example for one cell:

$$f_e = \frac{\text{row total} \times \text{column total}}{n} = \frac{30 \times 70}{200} = 10.5$$

Example: The Test Statistic

- **The test statistic value is:**

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

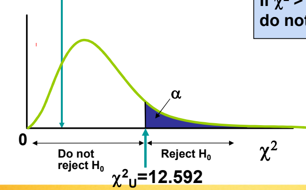
$$= \frac{(24-24.5)^2}{24.5} + \frac{(32-30.8)^2}{30.8} + \cdots + \frac{(10-8.4)^2}{8.4} = 0.709$$

χ^2_U = 12.592 for α = 0.05 from the chi-squared distribution with (4 – 1)(3 – 1) = 6 degrees of freedom

Example: Decision and Interpretation

The test statistic is $\chi^2 = 0.709$, χ^2_U with 6 d.f. = 12.592

Decision Rule:
If χ^2 > 12.592, reject H_0, otherwise, do not reject H_0



Here,
χ^2 = 0.709 < χ^2_U = 12.592, so do not reject H_0
Conclusion: there is not sufficient evidence that meal plan and class standing are related at α = 0.05

Fisher Exact Test of Significance

- "Fisher's exact test directly computes p, the probability of getting a table as strong as the observed table or stronger. This requires computing Fisher's for the given table and all stronger tables, then summing the separate p's to get the total probability of a table that strong or stronger." Garson 2011

For more details see
http://faculty.chass.ncsu.edu/garson/PA765/fisher.htm

Exercise 4

- Please open - PISA_2000_Part1b_AUSTRALIA_SMALL.sav

Simplified data from PISA 2000 Study – Few countries selected
(The **P**rogramme for **I**nternational **S**tudents **A**ssessment)

http://www.pisa.oecd.org

McNemar Test (Related Samples)

- Used to determine if there is a difference between proportions of two related samples
- Uses a test statistic the follows the normal distribution

McNemar Test (Related Samples)

- **Consider a 2 X 2 contingency table:**

	Condition 2		
Condition 1	Yes	No	Totals
Yes	A	B	A+B
No	C	D	C+D
Totals	A+C	B+D	n

McNemar Test (Related Samples)

- The sample proportions of interest are

$$p_1 = \frac{A+B}{n} = \text{proportion of respondents who answer yes to condition 1}$$

$$p_2 = \frac{A+C}{n} = \text{proportion of respondents who answer yes to condition 2}$$

- Test H_0: $\pi_1 = \pi_2$
  (the two population proportions are equal)
  H_1: $\pi_1 \neq \pi_2$
  (the two population proportions are not equal)

McNemar Test (Related Samples)

- **The test statistic for the McNemar test:**

$$Z = \frac{B-C}{\sqrt{B+C}}$$

where the test statistic Z is approximately normally distributed

Flinders University inspiring achievement Pawel Skuza 2013

Cochran's Q test

- Cochran's Q tests whether the percentages (proportions) of a given variable are the same across multiple dependent samples. It extends the McNemar test beyond two related samples.

Flinders University inspiring achievement Pawel Skuza 2013

Exercise 5

- Please open
  - Related_Samples.sav

Data from

Janzen, D. H. 1967. Interaction of the bull's-horn acacia (Acacia cornigera L.) with an ant inhabitant (Pseudomyrmex ferruginea F. Smith) in eastern Mexico. University of Kansas Science Bulletin 47:315-558.'

Flinders University inspiring achievement Pawel Skuza 2013

Chi-Square Goodness-of-Fit Test

- Does sample data conform to a hypothesized distribution?
  - Examples:
    - Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)
    - Do measurements from a production process follow a normal distribution?

Flinders University inspiring achievement Pawel Skuza 2013

Chi-Square Goodness-of-Fit Test

- **Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)**
  - **Sample data for 10 days per day of week:**

	Sum of calls for this day:
Monday	**290**
Tuesday	**250**
Wednesday	**238**
Thursday	**257**
Friday	**265**
Saturday	**230**
Sunday	**192**
	Σ = 1722

Flinders University inspiring achievement Pawel Skuza 2013

Logic of Goodness-of-Fit Test

- **If calls are uniformly distributed, the 1722 calls would be expected to be equally divided across the 7 days:**

$$\frac{1722}{7} = 246 \quad \text{expected calls per day if uniform}$$

- Chi-Square Goodness-of-Fit Test: test to see if the sample results are consistent with the expected results

Observed vs. Expected Frequencies

	Observed f_o	Expected f_e
Monday	290	246
Tuesday	250	246
Wednesday	238	246
Thursday	257	246
Friday	265	246
Saturday	230	246
Sunday	192	246
TOTAL	1722	1722

Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

- **The test statistic is**

$$\chi^2 = \sum_k \frac{(f_o - f_e)^2}{f_e} \quad (\text{where } df = k - p - 1)$$

where:

k = number of categories
f_o = observed frequency
f_e = expected frequency
p = number of parameters estimated from the data

The Rejection Region

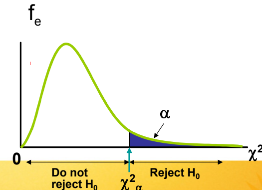
H_0: The distribution of calls is uniform over days of the week

H_1: The distribution of calls is not uniform

$$\chi^2 = \sum_k \frac{(f_o - f_e)^2}{f_e}$$

- **Reject H_0 if** $\chi^2 > \chi^2_\alpha$

k – 2 degrees of freedom, since p = 1 here (the mean was estimated)



Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week

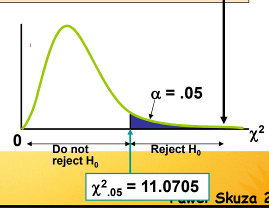
H_1: The distribution of calls is not uniform

$$\chi^2 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.05$$

k – 2 = 5 (k = 7 days of the week) so use 5 degrees of freedom:

$\chi^2_{.05}$ = 11.0705

Conclusion:
χ^2 = 23.05 > χ^2_α = 11.0705 so reject H_0 and conclude that the distribution is not uniform



Exercise 6

- Please open
  - PISA_2000_Part1b_AUSTRALIA_SMALL.sav

SPSS – BOOKS (**Hard copies)**

- **Chapter 16 in Allen, Peter James, & Bennett, Kellie. (2012). SPSS statistics : a practical guide : version 20. South Melbourne, Vic.: Cengage Learning Australia.**
- **Chapters 21-24 in Argyrous, George. (2011). Statistics for research : with a guide to SPSS (3rd ed.). Los Angeles: Sage.**
- **Chapter 3 in Landau, Sabine, & Everitt, Brian. (2004). A handbook of statistical analyses using SPSS. Boca Raton: Chapman & Hall/CRC.**
- **Chapter 11.5, 13 in Kinnear, Paul R., & Gray, Colin D. (2009). PASW statistics 17 made simple (replaces SPSS statistics 17). London ; New York: Psychology Press.**
- **Chapter 18 in Field, Andy P. (2009). Discovering statistics using SPSS : (and sex, drugs and rock 'n' roll) (3rd ed.). Los Angeles: SAGE Publications.**
- **Chapter 18 in Norušis, M. J. (2008). SPSS 16.0 [or later versions] Guide to Data Analysis. Upper Saddle River, NJ: Prentice Hall.**
- **Chapters 10 in Norušis, Marija J. (2008). *SPSS 16.0* [or later versions] *Statistical Procedures Companion*. Upper Saddle River, NJ: Prentice Hall.**

SPSS – BOOKS (**Online copies)**

Hard copies and online versions

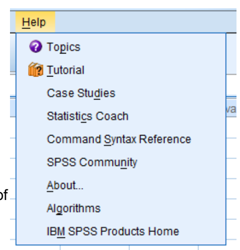
- **Chapter 16 in Pallant, Julie. (2010). SPSS survival manual a step by step guide to data analysis using SPSS (4th ed.). Maidenhead: Open University Press/McGraw-Hill.**
- **Chapter 7 in Morgan, George A. (2011). IBM SPSS for introductory statistics : use and interpretation (4th ed.). New York: Routledge.**

Online versions

- **Chapters 7 & 8 in Bryman, Alan, & Cramer, Duncan. (2011). Quantitative data analysis with IBM SPSS 17, 18 & 19 : a guide for social scientists. Hove ; New York: Routledge.**
- **Chapter 8 in Marston, Louise. (2010). Introductory statistics for health and nursing using SPSS. Los Angeles: SAGE.**
- **Chapters 8 & 14 in Larson-Hall, Jenifer. (2010). A guide to doing statistics in second language research using SPSS**

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

Pawel Skuza 2013

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

Pawel Skuza 2013