Valley International Journals
Open Access Journal

The International Journal of Social Sciences and Humanities Invention
Volume 2 issue 05 2015 page no.1327-1349 ISSN: ***2349-2031***
Available Online At: *http://valleyinternational.net/index.php/our-jou/theijsshi*

How Should Very Large Likert Datasets Be Analyzed?

Victor L. Landry, PhD
School of Advanced Studies, The University of Phoenix, Tempe, AZ 85282(USA)
vlandry@email.phoenix.edu

Abstract: Very large Likert databases, consisting of millions of cells, pose a special analytical problem because of the processing limitations of most desk top computers running Microsoft Excel. This paper shows that while there is validity to both the ordinal and the interval approaches to Likert analysis, the Marascuilo Procedure, an ordinal method based upon summations, proportions and Chi Square is superior in its ability to detect statistically significant intergrouprelationships. An illustration of its usage is given. It is the procedure of choice which operates well within desktop and more powerful computer systems. Commercially available add-ins are available.

Background

The Likert scale, developed in 1932 by Rensis Likert as a business management tool, is a psychometric scale that employs questionnaires, typically in survey research[1]. It is widely used in social science research. It is relatively easy to use and lends itself readily to Internet based survey research such as the "Survey Monkey" website[2]. Participants are presented with a questionnaire and are asked to specify their degree of agreement or disagreement by selecting one of several possible choices. Typically there are five but questionnaires can be developed with more. For example, a five item questionnaire would have many questions, each with an answer choiceranging from "strongly disagree" to "strongly agree" with "disagree", "neutral" and "agree" as midpoints. A questionnaire with more choices would include in the midranges with responses such as "moderately agree" between "strongly agree" and "agree" and so forth.

Debate has been ongoing and unresolved concerning how Likert data ought to be analyzed [3]. Proponents fall into two schools, the ordinalists and the intervalists [4]. The former, the conservative group, hold strictly to the ordinal nature of the data and eschew any parametric analysis [5]. The latter, the liberals, while recognizing that Likert is indeed ordinal, believe strongly that valid interval analyses can be made. In terms of the percentage of usage, it seems that the intervalists are in the majority. The ordinalists claim that treating ordinal data as interval is a real but subtle research violation[6].

Interval measurement hides information. If there is a range of numbers, coded from 1 to 5, and it can be determined that our average is 3. But what does that really tell us? If a "3" is "neutral" then we get no information concerning range, skewedness, or distribution. And, because of greater weight of high numbers (the "5s"), the nature of the low range (the "1s" and "2s") is missed. The scale of 1 to 5 is a coding structure.The problem would go away, according to the ordinalists, if Likert scales did not use numerals! This, they say, confuses the issue [6].

From a purist point of view, Likert is clearly ordinal but repeated testing by many over the decades has shown that interval measurement can be valid. It is suspected that Likert is really hybrid data and belongs to a unique class of quasi-interval/ordinal data. This study will show that a hybrid approach to Likert analysis is best, depending upon the researcher's purpose [3].

A Comparison of a Large Dataset By Ordinal and Interval Methods

Aparallel analysis was done on a large publically available dataset to compare and contrast interval and ordinal methods. The dataset that was obtained was called "16pf" [7] which consisted of the results of British psychologist Raymond Cattell's internet-based survey from 2014 [8]. There were 16 personality factor domains or question groups in the set with a total of 163 individual questions. Online questionnaire responses were obtained from 49149 participants from 251 countries. The variable for "country" was electronically obtained from the respondent's IP address.

The download consisted of two files—the main data set in .csv format and a second file had the description of the variables, that is, the personality questions themselves, and 3 demographics. That file was in .html.

The main data file was significantly large with163 question columns and 3 columns of demographics (age, gender and country). There were 15 personality factor domains with 10 questions per domain and one, the second domain,which had 13 questions. There were 49149 rows of Likert data responses, coded from "strongly disagree" (1) to "strongly agree" (5).

Therefore, the main data file consisted of 8,011,287answer cells coded with the numbers "1" through "5" and additional 147,447 cells with three demographic variables. The grand total was 8,158,734 cells for the entire file including null and invalid entries.

Microsoft Excel 2010 (Office Professional 64-bit) with the commercial add-in "XLSTAT" [11]was used throughout. A second add-in was the free downloadable "PowerPivot" from Microsoft [13] which expanded the standard pivot table functionalities. The operating system was 64-bit Windows 7 Professional.

The first procedure was to test the entire dataset for normalcy using the Kolmogorov-Smirnov goodness-of-fit routine. This tests whether one distribution (all our data) differs substantially from theoretical expectations.It was found that the null hypothesis of no difference between theoretical and actual distributions could not be rejected (p=0.995). The total data set was judged therefore to be normally distributed (Table 1).

Kolmogorov-Smirnov Test of Normalcy for Entire Dataset of 163 Columns and 49149 Rows With Descriptors	
N (cells)	8011287
D	0.168
p	0.995
α	0.05
Mean	3.115
Std Dev	1.2
Mode	4
Kurtosis	-1.744
Skewness	0.295

Table 1. Normalcy test and descriptors for entire dataset

In order to test if the normal distribution of treating data as intervals was a function of a very large sample size, Kolmogorov-Smirnov tests for normality were done on various sample sizes. In all cases, even with the size as low as n=25, the distributions were normal, albeit with distribution curves that flattened with lower sample sizes. It was concluded that Likert ordinal data retained a parametric form which is not a function of sample size.

The 163 columns of data were then compared and contrasted by the parametric ANOVA and non-parametric Kruskal-Wallis tests. The output showed a global (all columns, all rows)interval mean of 3.115 (sd=1.2). The mode was 4. Comparisons showed both with p-values of essentially zero with infinitesimal small differences. The null hypothesis ofno statistically significant difference within the dataset was rejected by both methods. However, although both the parametric and non-parametric routines demonstrated strong differences within the data, neither could show where the differences laid.

The "16pf" data file also included demographic variables including country (two letter code),

gender and age. The 251 country codes were translated into the "First World", "Second World" and "Third World", according to criteria given in "One World Nations Online" [9]. This allowed for further analysis to see if normal distributions were affected by socio-economic factors inherent in the concept of world position.The Kolmogorov-Smirnov test showed high p-values for all data, regardless of global economic position indicating that each subsection was normally distributed.

Following the Kruskal-Wallis test, a column chart was made for the entire non-subsected dataset. Visually, it can be seen that the data skews towards "agree" and "strongly "agree" (total = 56.21%) and that "neutral" is chosen less frequently (19.92%). The results in actuality are bimodal (Figure 1).

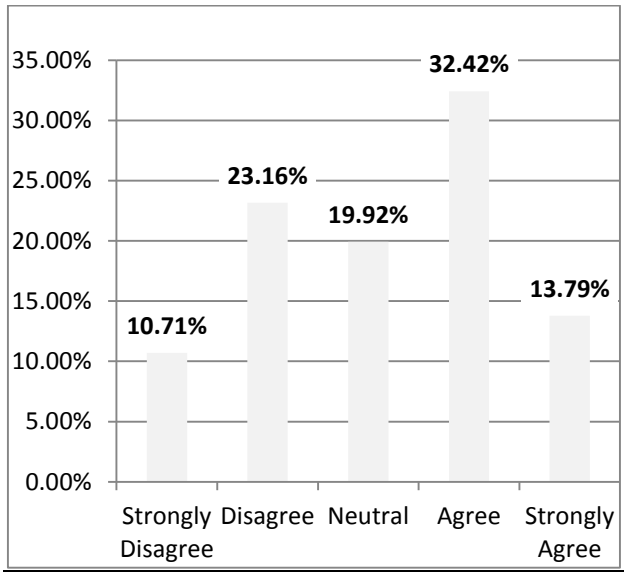

Figure 1. Proportion of choices for the entire dataset of 49149

This, along with the superimposed normal distribution that was obtained by the Kolmogrov-Smirnov goodness-of-fit method and the results of the ANOVA testing, indicated that parametric treatment would not be inappropriate, especially if the goal was group comparisons.

Using the Marascuilo Method

It was determined that while both parametric and non-parametric testing of the data were in agreement, showing very significant differences between and among the data ($p<0.00001$), neither was able to pinpoint the location of difference.

The most significant problem, moreover, with either the parametric or non-parametric analyses for very large databases was that the computing power of Excel was severely strained, resulting in long processing times or total processing failure.

The Marascuilo method was then employed because it allows for the simultaneous testing of differences between pairs of proportions [10]. The Marascuilo routine is included in the XLSTAT add-in for Excel [11].

The procedure begins by counting the ordinal Likert cells. This summation, and the proportions that result, subtly changes the nature of the dataset from being decidedly ordinal data into interval data. This happens when a count is made which, of course, has a true zero and clear distance between values. A conversion of sorts has been made from ordinal to interval data.

The first task was to generate the frequency distribution and histogram. At the bottom of each column of data, that is, in the 49150th through 49154th place, five summary bins, were created for each of the 163 columns. Excel's frequency function was attempted but, due to the size of the spreadsheet, this built-in frequency function exceeded the computer's memory. Instead, as recommended by Kyd [12], Microsoft Excel's powerful "countif" was used. The syntax of the countif function asked for a cell reference range and a condition. Each cell would be counted only if it met a specific criterion. In the case of "strongly disagree", the condition was having a "1".The countif function is extremely fast and does not strain Excel's processing power. This alone is ample reason to use the Marascuilo Procedure.

Likewise, empirical counts were made for the other categories of answers, counting them and inserting them into their respective bins based upon the number that was in their particular cell. This was done for all cells, resulting in 815 (163 columns times 5 categories) summaries. Finally,

an overall grand summary of the five choices was produced and tabled. A sixth holding row was made at the 49155th place for a grand total of the five sub-counts. Because missing or inappropriate values were not counted there was some variance in grand totals across the 163 columns resulting in 99,058 exclusions (Table 2).

Survey Selection	Total Counts	Proportion
Strongly Disagree	847365	10.71%
Disagree	1832849	23.16%
Neutral	1576087	19.92%
Agree	2565016	32.42%
Strongly Agree	1090912	13.79%
Grand Total	*7912229*	*100.00%*
Mean	*1582445.8*	
Std Dev	*672706.4096*	

Table 2: Bin frequencies for the data set.

Using the column totals from the 49155th row, five proportions of the grand frequencies were calculated. This showed a bimodal display with high points for "agree" and "disagree" and with "neutral" as a low point. This gives more information that the interval measurement test which, while showing centrality at 3.115 (within the neutral zone),did not reveal that the majority of the respondents (bins #4 and #5, 56.21%) reported that they "strongly agreed" and "agreed" with the survey questions (Figure 2).

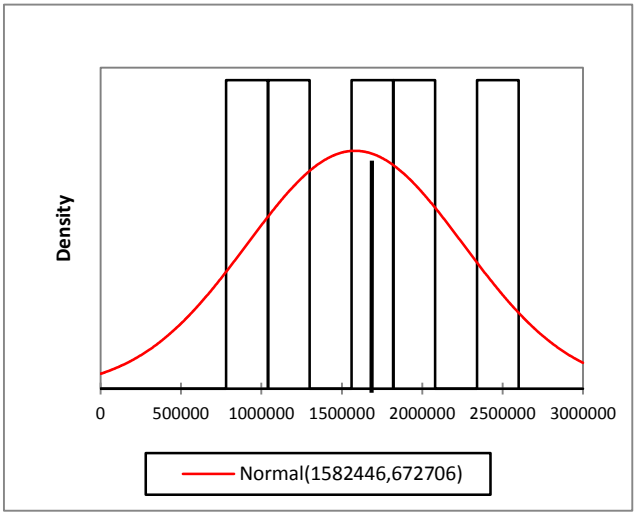

Figure 2: This histogram showed 5 bins corresponding to the 5 choices, from "strongly disagree" to "strongly agree", with the interval mean of 3.115 indicated.

The Marascuilo routine makes it possible to do intergroup comparisons. It tests the null hypothesis and allows for grouping. This is highly useful when dissecting the dataset by various demographics including global regions, gender and age groups.There are three main steps in this procedure.

The first step is to determine how many pairs of proportions there will be. The number of pairs from k samples is determined equaling k(k-1)/2 sets of proportions. The absolute differences between proportions within each pair will be the test statistics. Therefore, if there are three groups ("First World", "Second World", and "Third World") there would be 3(3-1)/2 or three sets of pairings. If there were 6 groups then there would be 6(6-1)/2 or 15 pairings and so forth. The second step was to pick a significance level, typically α=0.05, and compute the corresponding critical values for the procedure using Chi Square and multiplying by square root of the proportion ratios [10]

$$r_{ij} = \sqrt{X^2_{1-a,k-1}}\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}$$

where k=sample size, p_i and n_i= first proportion and size, p_j and n_j=proportion and size of the second part of the pair, and α=0.05.

The third and final step is to compare each of the k(k-1)/2 test statistics against its corresponding critical r_{ij} value. If that test statistic that exceeds its corresponding critical value at theα level, it is then considered significant.
There can bemultiple k(k-1)/2 sets of iterations of the Marascuilo Procedure as we compare and contrast the questionnaire by global, gender and age variables. Comparisons can also be made between the various domains.

An Illustration

This was an attempt to determine if the proportions of those who strongly agreed (throughout the data) differed significantly by global world designation. In other words, do the peoples of various countries express the same personality traits as measured by the 16 Personality Factor Test[5]. This was a summation analysis for the purpose of illustration for only those who expressed a "strongly agree" opinion. Other opinions were not analyzed.

It is noted that the three 'worlds' each differed from each other and that there was a proportional value differences between each. There was a significant difference between the three groups; no one was like the others. They sort into three separate 'groups', "A", "B", and "C". While those in the 'second world' have a strongly agree opinion 14% of the time, those in the 'first world' feel that way 17.4%, a small yet statistically significant difference (Table 3).

The Marascuilo Procedure shows us that not only are there is statistically significant difference between the groupings but exactly how of a difference that is.

Sample	Proportion	Groups		
Second	0.140	A		
Third	0.157		B	
First	0.174			C

Table 3.Proportional differences are great enough to separateeach sample group from the other.

The Marascuilo procedure generated the test value (difference between proportions) and compared it to the Chi Square critical value. If the critical value was exceeded, then the null hypothesis was rejected (Table 4).

Contrast	Value	Critical value	Signif-icant
|p(First) - p(Second)|	0.034	0.005	Yes
|p(First) - p(Third)|	0.017	0.004	Yes
|p(Second) - p(Third)|	0.017	0.006	Yes

Table4.Value, critical value and inter-group significance between and among the three groupings for 'strongly agree'.

Discussion

The ordinal-interval controversy of how to best analyze Likert data has been going on for almost 70 years. There is no assumption in this study that either the ordinalist or the intervalist positions will change.

However, there is a middle ground that recognizes, counter-intuitively, that Likert data can be validly treated as interval. This works well with group comparisons but fails to yield as much information as the ordinal approach or Chi Square or the related Marascuilo Procedure. This procedure tests significance both between and among various groupings.

Ultimately the choice of interval vs ordinal testing depends upon the researcher's goal. It is believed, however, that the Marascuilo Procedure is the better way to proceed. Much is gained and little is lost.

References

[1] Likert, R. (1932). A technique for the measurement of attitude scales. *Archives of Psychology*, *22*(140), 5-53.

[2] Survey monkey. (nd). Retrieved fromhttp://www.surveymonkey.com

[3] Knapp, T. (1989). *Treating ordinal scales as interval scales: an attempt to resolve the controversy*. Retrieved from http://www.mat.ufrgs.br/~viali/estatistica/mat2282/material/textos/treating_ordinal_scales[1].pdf

[4] Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38*(12), 1217-1218.

[5] Norman, G. (2010). Likert scales, levels of measurement and the "laws" of

statistics. *Advances in health science education*, *15*(5), 625-632.

[6] Uebersax, J. S. (2006). *Likert scales: dispelling the confusion*. Retrieved from http://john-uebersax.com/stat/likert.html

[7] Raw data from online personality tests. (2014). Retrieved from http://www.personality-testing.info/rawdata

[8] Cattelll's 16 personality factors. (nd). Retrieved from http://personality-testing.info/tests/16PF.php

[9] One World Nations Online. (2015). Retrieved from http://www.nationsonline.org/oneworld/third_world_countries.htm

[10] Comparing multiple proportions: the Marascuilo procedure. (2013). Retrieved from http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm

[11] XLSTAT (Version 2015) [Computer Software]. Retrieved from http://www.xlstat.com

[12] Kyd, C. (2014). *Use countifs, not frequency, to calculate frequency distribution for charting histograms*. Retrieved from http:// http://exceluser.com/formulas/countifs-frequency-distributions.htm

[13] Power Pivot Add-in. (2015). Retrieved from https://support.office.com/en-nz/article/Power-Pivot-Add-in-a9c2c6e2-cc49-4976-a7d7-40896795d045 (Microsoft.com)