Statistics essay: interpreting social data

Interpreting Social Data

The British Household Panel Survey of 1991 measured many opinions, among otherthings, of the UK population. One of the questions asked was whether thehusband should be the primary breadwinner in the household, while the wifestayed at home. Answers to the questions were provided on an ordinal scale, progressing in five ordinances from Strongly disagree to Strongly agree. Results for each ordinance were recorded from male respondents and femalerespondents. Of survey respondents, 96. 75, or N = 5325. 162 answered thisquestion of a total survey population of N = 5500. 829. 3. 2%, or N = 175. 667 ofsurvey respondents did not answer the question. In lay terms, this meansapproximately 97% of the survey respondents answered the question, while 3% didnot.

The study presents ordinal ranking, or ranking in a qualitative manner, of fivesets of concordant pairs of variables: the male and female count for those whostrongly agree the husband be the primary earner while the wife stays at home, the male and female count for those who agree, the male and female count forthose who are neutral, the male and female count for those who disagree, andthe male and female count for those who strongly disagree. The sexcross-tabulation presents numeric data for responses for each of the tenvariables, arranged in five variable pairs with male and female responses foreach variable pair. Data is presented in terms of number of responses for eachof the ten variables.

The counts or number of responses for each variable aredependent variables in the data analysis. We know they are dependentvariables because first,

they are presented on the y-axis in the chartgraphically representing the data. Dependent variables are graphicallyrepresented on the y-axis, with independent variables presented on the x-axis. Causally it becomes more difficult to distinguish between dependent andindependent variables at first glance. Dependent variables usually change as aresult of independent variables. For example, if one were studying the effectof a certain medication on blood sugar in diabetics, the independent variablewould be the amount of medication given to the patient. In a test group orcohort of patients, each would be given a set dosage and their blood sugarresponses recorded. One patient may respond with a blood sugar reading of 110when given 20mg of medicine. Another day the patient, again given 20mg ofmedicine, may respond with a blood sugar reading of 240. The amount ofmedicine provided to the patient is fixed, or the independent variable. Theresponse of the patient is variable, and believed to be influenced by, ordependent on, the amount of medicine provided. The dependent variable wouldtherefore be the responding blood sugar reading in each patient.

In this survey, independent variables are the fivechoices of answers available to the survey takers. These five possibleresponses are presented to each survey respondent, just as the medicine isprovided to the patient in the example above. The respondent then chooses hisor her reply to the five possible answers, or chooses not to answer thequestion at all. The amount of those choosing not to answer at all, 3. 2%, isconsidered statistically irrelevant in the analysis of this data. Data relatedto non-response is not considered from either an independent variable ordependent variable standpoint.

The amount of responses or response count for a givenindependent variable in the survey is a dependent variable. The response countwill change, at least slightly, from survey to survey. This could be a due tochange in survey size, response rate or number of those choosing to respond tothe statement, or possible minor fluctuation in percentage response for thefive answer possibilities. Although the statistical results of the responsesshould be similar, given a large enough and representative sample for eachsurvey attempt, some variance is likely to occur. The independent – dependentvariable relationship in the Husband should earn, wife should stay at homeanalysis is trickier to get one's mind around than the medical example givenabove. In the medical example, it is easy to grasp how a medicine could affectblood sugar, and the resulting cause-effect relationship. In this survey, thecreation of five answer groups causes the respondents to categorise theiropinion into one of the groups, a much more difficult mental construction thanmore straightforward cause-result examples.

Fourexamples of dependent variables in these statistics are the number of men whoagreed with the statement (525), the number of women who agreed with thestatement (520), the number of men who disagreed with the statement (688), andthe number of women who disagreed with the statement (997). As describedabove, we know these are dependent variables because they are caused by theindependent variables, the five ordinal answer groups, in the survey.

Overall, empirical data for the results is skewed towards the Disagree / Stronglydisagree end of the survey. Three of the independent variables are ofparticular note. Strongly agree is the lowest response for both men and

women, with Disagree being the highest response for both men and women althoughaccording to Gaussian predictions the Not agree/disagree variable should have thehighest distribution.

Inlay terms, the graphical representation of each of the five possible answersshould have looked like a bell-shaped curve. The two independent variables oneach end of the chart, Strongly agree and Strongly disagree, should have had alow but approximately equal response. The middle independent variable on thechart, Not agree / disagree, should have been the largest response. Thisshould have produced dependent variables of approximately 935 each for both menand women for the Not agree / disagree variable. Instead, the response for menwas 586, or 63% of typical distribution of answers. The response for women was702, or 75% of the typically distributed answers. The mean, or average, of allresponses in this survey is 1065. 2, with the mean or average of male responsesbeing 464. 6 and the mean or average of female responses being 600. 6. Were theresponses distributed evenly amongst all five possible answers, these would bethe anticipated response counts.

Inexamining this data, a hypothesis can be put forth that the correlation betweenthe counts on two of the answer possibilities (two of the dependent variables)will be some value other than zero, at least in the population represented bythe survey respondents. This hypothesis can be tested using the ordinalsymmetric measures produced in the data analysis. As Pilcher describes, whendata on two ordinal variables are grouped and given in categorical order, wewant to determine whether or not the relative positions of categories on twoscales go together' (1990, 98). Three ordinal symmetric

measures, Kendall’stau-b, Kendall’s tau-c, and Gamma, were therefore calculated to determine ifthe order of categories on the amount of agreement to the question would helpto predict the order of categories on the count or amount of those selectingeach ordinal category. The most appropriate measures of association toevaluate this hypothesis are the two Kendall’s tau measures. The Kendall tau-cmeasure allows for tie correction not considered in the Kendall tau-b measure. The results of these measures, value . 083 and . 102 with approximate T ^{b} of 6. 75 indicate there is neither a perfect positive or perfect negativecorrelation between variables. Results do indicate a low level of predictionand approximation of sampling distribution. The correlation between two of thedependent variables is indeed a value other than zero, proving the hypothesiscorrect.

Three nominal symmetric measures were also calculated. These showed weak relationship between category and count variables, with avalue of only . 096 for Phi, Cramer’s V, and Contingency Coefficient. Thesewere not used in testing the above hypothesis.

Atheory of distribution, Chebyshev’s theorem states that the standard of deviationwill be increased when data is spread out, and smaller when data is compacted. While the data may or may not present according to the empirical rule(bell-shaped), Chebyshev’s theorem contends that defined percentages of thedata will always be within a certain number of standard deviations from themean (Pilcher 1990).

Inthis example, data is compressed into five possible answer variables. The datadoes not present according to the empirical rule, but is skewed towards

thedisagreement end of the variable scale. However, Chebyshev’s theorem doesapply relating to the distribution of data according to standard deviation fromthe mean for nine of the ten dependent variables. The response count of womenwho Disagree with the statement the Husband should earn, the wife stay at home, was proportionately larger than would be indicated along normal distribution. While the response count for men is also statistically high, it is not beyondthe predictions of Chebyshev’s theorem. If the survey had been conducted withfewer independent variables, say three ordinances instead of five, theresulting data would be more tightly compacted. If the survey had beenconducted with ten ordinances, the data would have been more spread out.

REFERENCES

Pilcher, D., 1990. *Data Analysis forthe Helping Professions* . Sage Publications, London.