Published in final edited form as:

J Clin Epidemiol. 2012 February ; 65(2): 179–188. doi:10.1016/j.jclinepi.2011.06.007.

Multivariate Graphical Methods Provide an Insightful Way to Formulate Explanatory Hypotheses from Limited Categorical Data

Peter H. Van Ness, Ph.D., M.P.H.^{*,§}, Terrence E. Murphy, Ph.D.^{*}, Katy L.B. Araujo, M.P.H.^{*}, and Margaret A. Pisani, M.D., M.P.H.[†]

^{*}Department of Internal Medicine, Geriatrics Section, and the Program on Aging, Yale University School of Medicine

[§]Division of Chronic Disease Epidemiology, Yale School of Public Health

[†]Department of Internal Medicine, Pulmonary & Critical Care Section, and the Program on Aging, Yale University School of Medicine

Abstract

Objective—Graphical methods for generating explanatory hypotheses from limited categorical data are described and illustrated.

Study Design and Setting—Univariate, bivariate, multivariate, and multiplicative graphical methods were applied to clinical data regarding very ill older persons. The data to which these methods were applied were limited as to their nature (e.g., nominal categorical data) or quality (e.g., data subject to measurement error and missing values). Such limitations make confirmatory inference problematic but might still allow for meaningful generation of new explanatory hypotheses in some cases.

Results—A striking feature of the graphical results from this study's major illustrative application was that Post-Traumatic Stress Disorder (PTSD) following Intensive Care Unit (ICU) discharge occurred rarely and nearly always co-occurred with two or more other mental health conditions. These results suggest the explanatory hypothesis that PTSD in this context is less attributable to single traumatic causes than to acute illnesses contributing to a cascade of mental health decrements.

Conclusion—Illustrative applications of a sequence of graphical procedures yield more informative and less abstract representations of limited data than do descriptive statistics alone, and by doing so, they aid in the formulation of explanatory hypotheses.

Keywords

graphical methods; multivariate data; older patients; logic; critical care

© 2011 Elsevier Inc. All rights reserved.

Address correspondence to: Peter H. Van Ness, Ph.D., M.P.H., Yale University School of Medicine, Program on Aging, 300 George Street, Suite 775, New Haven, CT 06511-6664, Tel: (203) 737-1958; Fax: (203) 785-4823; peter.vanness@yale.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Clinical researchers often encounter data that are limited in some way that makes confirmatory statistical inference very problematic or relatively uninformative. Datasets with considerable missing data or measurement error—and without relevant auxiliary data for remediation of these shortcomings—are ways in which data can be limited. Another source of limitation is the nature rather than the quality of the data. For instance, fewer statistical techniques are applicable to nominal data than to data that have more ordered structure. Upon encountering such data researchers might decline to analyze them at all or they might apply traditional confirmatory methods with the stipulation that such analyses are exploratory. The first option risks discarding data, that while limited, might still have some research value. The second option makes results vulnerable to misinterpretation, especially when confirmatory methods appear to provide evidence of statistical significance.

Alternatively, researchers might explicitly implement the exploratory intent of their efforts by focusing on hypothesis generation by means of data visualization. Furthermore, attention might be given to multivariate graphical methods that help to overcome the relatively abstract, and so potentially misleading, nature of simple numerical methods of univariate and bivariate analysis. This methodological study recommends the alternative approach to dealing with limited categorical data. It provides a logical rationale for it, a sequence of graphical procedures that instantiates it, and illustrative applications that demonstrate its utility and suggest its own limitations.

Formulating an Explanatory Hypothesis

The methodological strategy of this study aims to identify relevant facts about seriously ill older patients and then present explanatory hypotheses that make these facts comprehensible. The statistical analysis consists of univariate, bivariate, multivariate, and multiplicative components. John Tukey is especially well known for advocating the use of graphical methods for hypothesis generation; he likened the use of such methods to “detective work” that gathers evidence for subsequent confirmatory adjudication [1]. Tukey and colleagues he inspired mostly developed graphical methods for continuous and count data, that is, data following distributions such as Gaussian or Poisson distributions [2]. Here a different logical rationale for graphical exploratory methods will be presented, one that dates back to the American logician Charles Sanders Peirce. It is argued that inference to an explanatory hypothesis has logical structure and rigor, and, when applied to the investigation of sociodemographic characteristics and health conditions of older cohort, also has practical, clinical value. Graphical methods for categorical data will be the focus of interest. This circumscription of the topic will allow recommendations to be more specific; it also directs attention to graphical procedures that have recently become widely accessible in commercial statistical software.

To confirm a hypothesis is a task of inductive inference. A statistically significant predictor variable in a longitudinal regression analysis aspires to establish an inductive inference. The analysis seeks to establish an association between some predictive factor (p) and a second outcome factor (q) occurring temporally consequent to it such that their relationships is expressible as a conditional mean or probability (if p then q) unlikely to be attributable to chance. Formulating an explanatory hypothesis identifies an informative conditional relationship of a different sort. First, some surprising factor (q) is identified by perceptual judgment. In this study, such identification will be promoted by presenting major study results in graphical form. Then a conditional relationship (if p then q) is posited that says that this factor (q) follows from a certain explanatory hypothesis (p). Finally, the explanatory hypothesis is formulated on the basis of its ability to explain the surprising factor plausibly and parsimoniously. Of course, the explanatory hypothesis obtained in this

way must then be confirmed with relevant but independent data. Applying graphical methods to limited data does not overcome the data's limitations. Instead, it redirects investigative efforts towards a more modest exploratory goal than confirmatory inference.

Formulating an explanatory hypothesis moves from effect to cause and relies more on judgments of similarity than consequence. Construed deductively it is fallacious ("the fallacy of affirming the consequent"); it is wrong, however, to force such reasoning into deductive or inductive modes. Peirce called this type of reasoning "abduction" [3]; logicians since have argued about its precise status but all agree that it is an integral part of scientific research [4, 5]. A major objective of this study is to show that the process of formulating an explanatory hypothesis has logical structure and rigor, and, that when it is undertaken in the investigation of sociodemographic characteristics and health conditions of older cohorts, it can also have practical, clinical value.

Methods

Statistical and Graphical Methods

The proposed statistical and graphical investigation in its full application consists of univariate, bivariate, multivariate, and multiplicative components: for each statistical component a graphical representation of results is provided. First, the univariate prevalence of each of the variables of interest is given with Bayesian equal-tail credible intervals representing the precision of the estimate. These credible intervals are used instead of more familiar confidence intervals because they have a straightforward visual interpretation: for a 95% credible interval there is a 95% probability that a value drawn from the posterior distribution for some factor's prevalence will fall within the upper and lower bounds of the interval [6]. The credible intervals are not compared to any null value and so no formal inferences are drawn.

A correlation matrix of Kendall tau-b coefficients for various factors is provided as a bivariate analysis. These coefficients are appropriate for binary indicator variables or ordered categorical variables [7]. For nominal categorical variables a Pearson chi-square statistic, or an exact counterpart, would be required for inference about pair-wise associations. When it is appropriate, however, the Kendall tau-b statistic is preferable because a matrix of its values simultaneously conveys information about the direction and magnitude of each pair-wise association on a scale from -1 to 1.

Simple and multiple correspondence analyses are conducted to investigate the multivariate relationships of data points representing the categories—possibly nominal or ordinal—of several variables of interest [8] [9]. Correspondence analysis is analogous to principal component and factor analyses in several ways but differs by being especially applicable to categorical data. Specifically, correspondence analysis involves the decomposition of the total variance of contingency table data by determining mutually independent dimensions that portray prominent departures from independence. The number of dimensions in a correspondence plot is determined by criteria similar to those used in principal components analysis and factor analysis, e.g., by thresholds for percentage of variance accounted for and by marked changes in a scree plot. Since the number of dimension displayed in correspondence plots is characteristically fewer than the maximum possible, it is often used as a data reduction technique.

Unlike a correlation matrix, the relationships identified in a correspondence plot are multivariate, taking into account more than two categories or variables at once. A plot of correspondences can provide a visual display of how categories of variables cluster together in specific ways. A correspondence analysis is multiple rather than simple when associations

are among three or more variables instead of between only the multiple categories of two variables. Both types will be illustrated in following sections.

Finally, log-linear models employing counts from a multi-dimensional contingency table as its outcome are fit. Some log-linear models can investigate the multiplicative relationship of effect modification and the occurrence of conditional independence among the variables of interest [10]. In some applications this model can answer questions such as: “Does the association between two variables differ across the levels of a third?” and “Given the occurrence of a certain factor specified by an indicator variable, are two other factors similarly indicated independent of one another?” Fitting log-linear models will utilize overall model goodness-of-fit statistics whenever possible: dividing the model deviance by its degrees of freedom provides a measure of goodness of fit. Mosaic plots are used to represent visually the results of the log-linear analysis [11–13].

The software used to conduct the statistical analyses was SAS[®] version 9.22 [14]; the R statistical package (version 2.13.1) was used to generate correspondence and mosaic plots [15].

Illustrative Applications

Sociodemographic Data from the LEAP Study—The Longitudinal Examination of Attitudes and Preferences study (LEAP) investigated a seriously ill cohort at the end of life with regard to their treatment preferences [16]. Dr. Terri Fried gathered data (n=226) for the LEAP study in Connecticut during the years 1999–2002. Her research objective for the LEAP study was to describe changes in treatment preferences among this cohort of older men and women with advanced illness over time and to determine the associations between participants’ sociodemographic, health, and psychosocial profiles and their preferences as measured by several different outcomes.

The seriously ill character of the cohort was reflected in stratified sampling that resulted in approximately one third of the study participants having cancer (35%), chronic obstructive pulmonary disease (36%) or congestive heart failure (29%). The age criterion for enrollment (60 years or older) yielded a cohort with an average age of 73 (± 7); slightly more men (57%) than women were enrolled. Univariate descriptions of the marital and housing statuses of the LEAP study participants are provided by cross-tabulations of these two factors in Table 1; unadjusted bivariate results are also provided in this table. Two of the graphical procedures described above will be applied to the data in this table as a first illustration.

Health Conditions from the EPIC Study—A second illustration uses data from Dr. Margaret Pisani’s Evaluation of Psychoactive Medications in the Intensive Care Unit (EPIC) study. Study participants were drawn from a cohort study of 309 consecutive patients aged 60 years and older admitted to the medical (Intensive Care Unit) ICU at Yale-New Haven Hospital during the years 2002–2004 [17]. Yale-New Haven is a 900-bed teaching hospital with a 14-bed ICU. This study includes the 152 patients who were alive and completed an interview at 1-month post-ICU discharge. See Table 2.

The 1-month follow-up interview included information about five health conditions measured by Katz’s Basic Activities of Daily Living (ADL) [18], a two-question screen for depression [19], the Mini-Mental State Examination (MMSE) [20], the Confusion Assessment Method (CAM) [21], and a Post-Traumatic Symptom Scale (PTSS-10) [22]. For this analysis functional disability was assessed using Katz’s ADL disability measure, comparing any versus no functional disability. Depression was considered present if there were positive responses to both depression screening questions. Cognitive impairment was identified by an MMSE less than 24. Delirium was dichotomized as present or absent based

on the Confusion Assessment Method (CAM) criteria [21]. Persons scoring 35 or higher on Part B of the Post-Traumatic Symptom Scale (PTSS-10) were identified as having Post-Traumatic Stress Disorder (PTSD).

Informed consent for participation in the EPIC and LEAP studies was obtained (separately) from surrogates and patients according to procedures approved by the Institutional Review Board of the Yale University School of Medicine.

Results

The LEAP Study

Table 1 provides a cross-tabulation of the marital and housing status variables from the LEAP study. Almost 60% of study participants own their own homes and almost the same percentage (58%) are married; some 44% are both married and own their own homes. A Pearson's chi-square statistic and a Fisher's exact test statistic give evidence that there is a dependency between table rows and columns. Relationships between the other marriage and housing nominal categories are not as immediately evident. Often researchers confronted with nominal categorical data discontinue their analysis at this point. Further investigation, however, is possible and can be summarized visually with the help of correspondence and mosaic plots.

Figure 1 provides an asymmetric correspondence plot for the marriage and housing variables. In an asymmetric plot the row and column points are scaled with different types of coordinates. Here the marital status or row points are of most interest and their scaling is such that the distances between them are most readily interpretable. Symmetric plots, e.g., Figure 5, use the same type of coordinates for both row and column points. In general, and especially for symmetric plots, it is not possible to infer from the closeness of two row and column points that the relevant row and column factors are associated. For asymmetric plots, however, as Greenacre notes, "such an interpretation is justified to a certain extent [9:72]." Asymmetric plots are thus amenable to fairly intuitive visual interpretations. For instance, the relative closeness of the married point to the home-owning point suggests a relatively strong association. So too does the relative closeness of the separated/divorced point to senior housing/assisted living point suggest some association. The former association is evident in Table 1 by the relatively large number of counts for married homeowners but the latter association is less evident.

A dimensional interpretation of the plot yields further insight. Dimensional interpretations are not strictly determined by statistical considerations; rather they emerge from the application of substantive knowledge to appropriately constructed correspondence plots. In this sense they are well suited to the task of hypothesis generation. It is noteworthy in Figure 1 that along the horizontal line through the origin (the 0-0 coordinates) the married point and separated/divorced point are most distant from one another. The plot also suggests that the home-owning point and the senior housing/assisted living point are more distant from one another than from the other housing points. Assuming that being separated or divorced is associated with some degree of emotion turmoil and that assisted living situations suggest some degree of physical infirmity, the first dimension in the plot might be interpreted as a general measure of emotional and physical distress. The second, vertical dimension shows less difference between the marital status points. Regarding the housing status points it suggests that the largest difference is between living in the home of a relative and living in an apartment. A sensitivity analysis using an asymmetric plot scaled to focus interest on housing status affirmed the reported relationships. Assuming that practical and emotional support are more forthcoming in a relative's home than in an apartment, and this assumption would appear to be especially plausible for unmarried apartment dwellers, this second

dimension might be interpretable as indicating degree of available support for older persons. The percentage of variance accounted for by this second dimension (8.0%) is much less than by the first dimension (91.8%) and so should be given less weight in formulating an explanatory hypothesis from this data.

In an analysis involving two variables a multiplicative analysis is not possible but even a simple mosaic plot makes contingency table data more accessible and insightful. A mosaic plot resizes the cells in a contingency table so that the area of a cell is proportional to the percentage of persons in the entire dataset who occupy that cell. Figure 2 shows the cell for married homeowners as having the largest area. The second largest area is accorded to widower's who own their own home. These two cells are visually differentiated in two ways, by the coloring and by the outlining of the cells. The visual features are functions of standardized residuals for the cells obtained from a log-linear independence model. Bluer coloring represents larger absolute residuals and solid versus dashed outlining indicates respectively whether the residuals are positive or negative. Thus, the pink and dashed-outlined cell for married apartment dwellers indicates that, upon the assumption of an independence log-linear model, there are slightly fewer counts in this cell than would be expected given the marginal distributions for the table. The dark blue color and solid outline of the cell for married home-owners indicate that there are moderately more counts in this cell than would be expected given the marginal distributions. The utility of mosaic plots for visualizing categorical data becomes greater when they reflect a more complicated process of log-linear modeling as will be evident in the second illustration.

An explanatory hypothesis suggested by the most striking features of this correspondence plot is that transitions in marital status are reflected in transitions in housing status in later life and that some such transitions might be associated with diminished well-being. Separated or divorced persons—and to a lesser degree, widowers—living in senior housing or assisted living situations might be especially vulnerable. The mosaic plot shows that these groups are not inconsiderable in number.

The facts revealed by the correspondence and mosaic plots are not entirely unrepresented in the corresponding contingency table. The plots, though, are more readily interpretable given their visual representation, and they are more statistically incisive given their respective analytical annotations. The clinical importance of the relationship between late life marital and housing transitions is that such changes are probably related to health changes. Finally, there is also statistical importance to confirming the hypothesis that housing locations are associated with mental and physical health among older persons because study samples are often gathered according to housing sites. Meaningful clustering of data according to sites should be accounted for in statistical inference.

The EPIC Study—Descriptive characteristics of study participants upon admission to the ICU are provided in Table 2. This cohort mostly consists of the “young old,” with 61.8% between the ages of 60 and 75 years. The majority are female (58.6%), white (81.6%), and high school-educated (78.2%). Some 69.1% of study participants were smokers at some time prior to ICU admission.

The prevalence of each of the five health conditions assessed one month after discharge from the ICU is represented in Figure 3 with an accompanying 95% credible interval. Having at least one ADL disability is the most common health condition (57.2%) and experiencing delirium is the least common (7.2%). PTSD is slightly more prevalent (7.9%) than delirium, and depression is slightly less prevalent (48.0%) than ADL disability.

Figure 4 shows the correlations among the five conditions. In general, these conditions have a consistently modest level of positive pair-wise correlation with only delirium and PTSD having a level of correlation that falls below 0.15. Cognitive impairment and delirium are most highly correlated ($\tau_{ab} = 0.32$) while PTSD and depression are correlated at the next highest level ($\tau_{ab} = 0.26$). ADL disability—the one condition that is not primarily a mental health condition—has a fairly consistent level of correlation with the other four variables, its correlation with delirium being the lowest ($\tau_{ab} = 0.19$) and with depression the highest ($\tau_{ab} = 0.24$).

The results of a multiple correspondence analysis are displayed in Figure 5. A two-dimensional plot more thoroughly satisfies proposed criteria [8] [23] [24] for determining the number of plot dimensions than do one and three-dimensional alternatives; it also promotes parsimony and ease of interpretation. This two-dimensional correspondence plot is symmetric. From the point of view of visual interpretation, symmetric plots have both advantages and disadvantages relative to asymmetric plots, with the extent of these relative merits varying according to the joint distributions of the data. Advantageous is the fact that symmetric plots are less likely to show data points that are closely “bunched up” [9:70] in the middle of the plot. To avoid such bunching a symmetric plot was used in this analysis. A visual drawback of symmetric plots is that distances between row and column points are undefined and so clustering of data points for different variables in multiple correspondence plots are less revealing of associations. For this reason, when bunching is not a problem, an asymmetric plot can often be more helpful for the process of hypothesis generation recommended here.

This symmetric correspondence plot of Figure 5 shows a clear differentiation between health conditions being present and being absent: data points representing the occurrence of health conditions appear on the right side of the plot and data points representing their nonoccurrence appear on the left side. Thus they are all differentiated visually along the horizontal axis for Dimension 1. Additionally, the data points for cognitive impairment and delirium are clustered together in the plot’s lower-right quadrant and the data points for PTSD and depression are clustered in the upper-right quadrant. While in symmetric correspondence plots the distances between points from different variables are undefined, the co-occurrence of points in the same quadrant, thereby having a similar orientation to the plot’s origin, provides some indication of clustering relationships [8:642]. For Dimension 2 there is no strictly statistical interpretation, but the clinical nature of the various conditions suggests that this dimension might be differentiating between cognitive (cognitive impairment and delirium) [17] and emotional disorders (PTSD and depression). Upon this interpretation ADL disability is grouped with the emotional disorders—a grouping that makes some sense in light of research showing the impact of depressive symptoms on functional disability [25, 26].

In trying to fit a log-linear model to data consisting of the five health conditions of interest in this cohort, almost half (15/32) of the table cells had zero counts. The analysis and plots of these data would not be very informative so attention was concentrated on PTSD, depression, and ADL disability, i.e. the three variables sharing the upper-right quadrant in the correspondence plot. Still, when the cross-tabulation of PTSD by depression was further broken down by ADL disability status a cell with a zero count occurred. The occurrence of such a zero count in a three-dimensional table—one involving three variables—indicates insufficient degrees of freedom to fit a saturated model with a three-way interaction. Practically, this implies that whether associations between two variables differ across the levels of a third cannot be effectively investigated. Since the zero cell was empirical rather than structural in nature, i.e., it was attributable to chance occurrence in this specific sample and not to a measurement or design constraint that made a count for this cell impossible, a

single count was added to the empty cell and a saturated model was fit. This enabled estimation of a three-way interaction. Since this interaction contributed little to the model, analysis was conducted without this addition.

Results from log-linear models for PTSD, depression, and ADL disability are represented in the mosaic plots of panels a and b of Figure 6. Panel 6a shows the results from a mutual independence model, i.e., one that assumes that the distribution of cell counts is what one would expect by chance given the marginal distributions of the variables. Large absolute residuals occur where this assumption fails. In panel 6a large positive residuals occur in the upper-left and lower-right rectangles, indicating that more than expected counts occur when PTSD, depression, and ADL disability values were all positive or all negative. This supports the results of the correspondence analysis that showed their co-occurrence in the upper-right quadrant of Figure 5.

Figure 6b is a mosaic plot of the same categorical data as represented in 6a; perforce, the pattern of colored rectangles remains the same. The standardized residual values, however, are quite different, reflecting a log-linear model containing all three pair-wise interactions in addition to the three health variables themselves. The small values of the residuals indicate that the model fits better than the independence model and that the associations between PTSD, depression, and ADL disability persist in this multivariate analytical context.

This second illustration investigates the prevalence and interrelationships of five health conditions in an older cohort of patients one month after discharge from an ICU. The most striking feature of the post-ICU health data is the co-occurrence of PTSD with other serious mental health conditions. In our cohort PTSD following ICU discharge most often occurs in the presence of two or more other conditions. What might explain this observation is that PTSD among older patients following ICU discharge is unlike PTSD attributable to experiences of violence in that it is not attributable primarily to a single traumatic event. Rather it follows upon an acute illness requiring treatment in the ICU that leads to a cascade of mental health decrements eventuating in a condition that satisfies the diagnostic criteria for PTSD. This explanatory hypothesis fulfills the modest clinical objective of this study, and it does so more plausibly and parsimoniously than expansion of the notion of a single traumatic event to include the multifarious dimensions of the ICU experience. Rigorously testing this hypothesis would require a separate, carefully designed, and considerably more complex study.

Discussion

Inferential statistics seek to move from old knowledge to new where old knowledge consists of a sample of data representative of some larger population and new knowledge consists of population parameter estimates. Such new knowledge is spurious when it is, in fact, attributable to chance, bias, or ambiguity; it is useless when incapable of meaningful generalizability. A relatively large number of missing data points—and especially ones that are not plausibly missing at random—makes statistical inference questionable and this is the case with the EPIC post-ICU cohort of older, critically ill persons. Also, key EPIC study variables were measured differently at ICU admission and at 1-month post-ICU, making longitudinal analyses problematic. Accordingly, p-values for longitudinal associations were not provided in these analyses.

Descriptive statistics seek to summarize data in ways that advance practical purposes. While they are not subject to inferential errors, descriptive statistics can be misleading. A major culprit here is what the logician Alfred North Whitehead called “the fallacy of misplaced concreteness” or the mistaking of abstract representations for the multiple concrete realities

they represent [27]. For instance, summing data points under single characteristics, or correlating them in pairs, are relatively abstract operations because the realities they represent actually occur in much more variegated interactive contexts. Examining multivariate statistical relationships is a partial remedy to the dangers of abstractionism in descriptive statistics. A second way in which the methods proposed here aim to clarify the complexities of multivariate statistical relationships is through use of graphical methods rather than numerical representations. The representation of multivariate relationships by correspondence and mosaic plots allows for perceptual identifications of patterns and relationships not always readily apparent in numerical representations.

There are limitations to the graphical strategy recommended here. Although commercial software is increasingly providing relevant resources, some time must be invested in generating graphical displays of categorical data. Also, while Bayesian credible intervals and Kendall tau-b statistics work fairly well with small samples, correspondence analysis and log-linear models are less informative for small studies. Finally, use of graphical methods as recommended here will not remedy potential biases from measurement error and missing data but only mitigate mistakes by making claims hypothetical. There is no rigorous way to ensure with these graphical methods that an explanatory hypothesis generated with them is better than alternatives or will be confirmed upon further investigation. The methods are intended as spurs to imagination and creativity. When combined with clinical investigators' wealth of relevant experience, and with statistical considerations mentioned in this study, these graphical methods can make limited data useful. Novel and productive hypotheses are an essential ingredient in important clinical research. Multivariate graphical methods of the sort presented in this study can help researchers more easily formulate them from limited categorical data.

Acknowledgments

Dr. Van Ness was supported in part by grants from the Biostatistics Core of the Claude D. Pepper Older Americans Independence Center at Yale University School of Medicine (#P30AG021342). The authors thank Terri R. Fried for the use of LEAP data, Heather G. Allore for her support and assistance, and Linda Leo-Summers for her assistance with the creation of the correspondence and mosaic plots.

References

1. Tukey, JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing; 1977.
2. Cleveland, WS. *Visualizing Data*. Summit, NJ: Hobart Press; 1993.
3. Peirce, CS. Pragmatism as the logic of abduction. In: Houser, N., et al., editors. *The Essential Peirce: Selected Philosophical Writings*. Vol. 2. Bloomington, IN: Indiana University Press; 1998 (1903). p. 226-241.
4. Niiniluoto I. Defending abduction. *Philosophy of Science*. 1999; 66 (3 Supplement):S436-S451.
5. Kapitan, T. Peirce and the structure of abductive inference. In: Houser, N.; Roberts, DD.; Van Evra, J., editors. *Studies in the Logic of Charles Sanders Peirce*. Bloomington, IN: Indiana University Press; 1997. p. 477-496.
6. Lee, PM. *Bayesian Statistics: An Introduction*. 3. Hoboken, NJ: John Wiley & Sons; 2009.
7. Agresti, A. *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons; 1984. p. 166-67.
8. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunanathan S, Bandeen-Roche K, Beland F, Bergman H. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiology*. 2010; 63:638-646. 643. and especially. [PubMed: 19896800]
9. Greenacre, M. *Correspondence Analysis in Practice*. 2. Boca Raton, FL: Chapman & Hall/CRC; 2007.
10. Zelterman, D. *Advanced Log-Linear Models Using SAS®*. Cary, NC: SAS Institute Inc; 2002.
11. Hartigan JA, Kleiner B. A mosaic of television ratings. *American Statistician*. 1984; 38:32-35.

12. Theus M, Lauer SRW. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*. 1999; 8:396–412.
13. Friendly, M. *Visualizing Categorical Data*. Cary, NC: SAS Institute, Inc; 2000.
14. SAS/STAT 9.22 User's Guide. Cary, NC: SAS Institute Inc; 2010.
15. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2010.
16. Fried TR, Byers AL, Gallo WT, Van Ness PH, Towle VR, O'Leary JR, Dubin JA. Prospective study of health status preferences and changes in preferences over time in older adults. *Archives of Internal Medicine*. 2006; 166:890–895. [PubMed: 16636215]
17. Pisani MA, Murphy TE, Van Ness PH, Araujo KLB, Ely EW, Inouye SK. Characteristics associated with delirium in older patients in a medical intensive care unit. *Archives of Internal Medicine*. 2007; 167:1629–1634. [PubMed: 17698685]
18. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The Index of ADL: A standardized measure of biological and psychosocial function. *JAMA*. 1963; 185:914–919. [PubMed: 14044222]
19. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression. Two questions are as good as many. *Journal of General Internal Medicine*. 1997; 12:439–445. [PubMed: 9229283]
20. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*. 1975; 12:189–198. [PubMed: 1202204]
21. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegel AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. [see comment]. *Annals of Internal Medicine*. 1990; 113:941–948. [PubMed: 2240918]
22. Stoll C, Kapfhammer HP, Rothenhausler HB, Haller M, Briegel J, Schmidt M, Krauseneck T, Durst K, Schelling G. Sensitivity and specificity of a screening test to document traumatic experiences and to diagnose post-traumatic stress disorder in ARDS patients after intensive care treatment. *Intensive Care Medicine*. 1999; 25:697–704. [PubMed: 10470573]
23. Higgs NT. Practical and innovative uses of correspondence analysis. *Journal of the Royal Statistical Society Series D (The Statistician)*. 1991; 40:183–194. 186. and especially.
24. Clausen, S-E. *Applied Correspondence Analysis: An Introduction*. Thousand Oaks, CA: Sage Publications, Inc; 1998. p. 23-25.
25. Penninx BW, Deeg DJ, van Eijk JT, Beekman AT, Guralnik JM. Changes in depression and physical decline in older adults: a longitudinal perspective. *Journal of Affective Disorders*. 2000; 61:1–12. [PubMed: 11099735]
26. Bruce ML, Seeman TE, Merrill SS, Blazer DG. The impact of depressive symptomatology on physical disability: MacArthur Studies of Successful Aging. *American Journal of Public Health*. 1994; 84:1796–1799. [PubMed: 7977920]
27. Whitehead, AN. *Science and the Modern World*. New York: Free Press; 1997 (1925). p. 51

What is new?

1. A historical and logical rationale for using multivariate graphical methods as means for hypothesis generation is provided.
2. A sequence of graphical procedures is described and illustrated and recommendations for their application are given.
3. The multivariate graphical methods illustrated here should be considered as alternatives or complements to traditional descriptive statistics, especially when trying to formulate explanatory hypotheses from limited categorical data.

Figure 1.
Asymmetric Correspondence Plot of Marital Status by Housing Status (LEAP Study)

Figure 2.
Mosaic Plot of Marital Status by Housing Status from a Mutual Independence Log-Linear Model (LEAP Study)

Figure 3.
Prevalence Estimates with 95% Bayesian Credible Intervals of Five Health Conditions
Observed One Month after ICU Discharge (n=152) (EPIC Study)

Figure 4.
Kendall Tau-b Correlation Matrix with Asymptotic Standard Errors for Five Health Conditions Observed One Month after ICU Discharge (n=152) (EPIC Study)

Figure 5.
Symmetric Multiple Correspondence Plot of Five Health Conditions Observed One Month after ICU Discharge (EPIC Study)

Figure 6a

Figure 6b

Figure 6.

Figure 6a. Mosaic Plot for a Mutual Independence Log-Linear Model of Three Health Conditions Observed One Month after ICU Discharge (EPIC Study)

Figure 6b. Mosaic Plot for an All Two-Way Interaction Log-Linear Model of Three Health Conditions Observed One Month after ICU Discharge (EPIC Study)

Table 1

Cross-tabulation of Marital Status by Housing Status (LEAP Study)

Marital Status	Housing Status				
	Frequency	Percentage	Row Percentage	Column Percentage	Total
Married	100	7	17	6	130
	44.25	3.10	7.52	2.65	57.52
	76.92	5.38	13.08	4.62	
	74.07	25.93	56.67	17.65	
Separated/Divorced	6	7	6	11	30
	2.65	3.10	2.65	4.87	13.27
	20.00	23.33	20.00	36.67	
	4.44	25.93	20.00	32.35	
Widowed	23	12	4	14	53
	10.18	5.31	1.77	6.19	23.45
	43.40	22.64	7.55	26.42	
	17.04	44.44	13.33	41.18	
Never Married	6	1	3	3	13
	2.65	0.44	1.33	1.33	5.75
	46.15	7.69	23.08	23.08	
	4.44	3.70	10.00	8.82	
Total	135	27	30	34	226
	59.73	11.95	13.27	15.04	100.0
					0

Statistic	DF	Value	P-value
Chi-Square	9	57.4522	<.0001
WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Table Probability (P)	4.363E-19
Probability <=P	1.573E-10

Table 2

Participant ICU Admission Characteristics (n=152) (EPIC Study)

	n (%)
Age in categories of years	
60–65	38 (25.0)
66–75	56 (36.8)
76–84	44 (29.0)
85–92	14 (9.2)
Female gender	89 (58.6)
White race	124 (81.6)
Married	71 (46.7)
High School Education	115 (78.2)
From Nursing Home	20 (13.2)
Alcohol Use	61 (40.7)
Tobacco Use	105 (69.1)
ICU Admission Diagnosis	
Respiratory Illness	78 (51.3)
Gastrointestinal Illness	31 (20.4)
Sepsis	22 (14.5)
Other	21 (13.8)

Missing data present for some subjects. For high school education missing=5; Alcohol use missing=2.