



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 79*

On the Measurement of Model Fit for Sparse Categorical Data

KATRIN KRAUS



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012

ISSN 1652-9030
ISBN 978-91-554-8394-4
urn:nbn:se:uu:diva-173768

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Thursday, June 14, 2012 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Kraus, K. 2012. On the Measurement of Model Fit for Sparse Categorical Data. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 79. 23 pp. Uppsala. ISBN 978-91-554-8394-4.

This thesis consists of four papers that deal with several aspects of the measurement of model fit for categorical data. In all papers, special attention is paid to situations with sparse data.

The first paper concerns the computational burden of calculating Pearson's goodness-of-fit statistic for situations where many response patterns have observed frequencies that equal zero. A simple solution is presented that allows for the computation of the total value of Pearson's goodness-of-fit statistic when the expected frequencies of response patterns with observed frequencies of zero are unknown.

In the second paper, a new fit statistic is presented that is a modification of Pearson's statistic but that is not adversely affected by response patterns with very small expected frequencies. It is shown that the new statistic is asymptotically equivalent to Pearson's goodness-of-fit statistic and hence, asymptotically chi-square distributed.

In the third paper, comprehensive simulation studies are conducted that compare seven asymptotically equivalent fit statistics, including the new statistic. Situations that are considered concern both multinomial sampling and factor analysis. Tests for the goodness-of-fit are conducted by means of the asymptotic and the bootstrap approach both under the null hypothesis and when there is a certain degree of misfit in the data. Results indicate that recommendations on the use of a fit statistic can be dependent on the investigated situation and on the purpose of the model test. Power varies substantially between the fit statistics and the cause of the misfit of the model. Findings indicate further that the new statistic proposed in this thesis shows rather stable results and compared to the other fit statistics, no disadvantageous characteristics of the fit statistic are found.

Finally, in the fourth paper, the potential necessity of determining the goodness-of-fit by two sided model testing is adverted. A simulation study is conducted that investigates differences between the one sided and the two sided approach of model testing. Situations are identified for which two sided model testing has advantages over the one sided approach.

Keywords: goodness-of-fit, sparseness, model fit, categorical data, fit statistic, sparse contingency table

Katrin Kraus, Uppsala University, Department of Statistics, SE-751 20 Uppsala, Sweden.

© Katrin Kraus 2012

ISSN 1652-9030

ISBN 978-91-554-8394-4

urn:nbn:se:uu:diva-173768 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-173768>)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I Kraus, K. (2012) On the Computation of Pearson's Goodness-of-Fit Statistic for Sparse Contingency Tables.
- II Kraus, K. (2012) An Alternative to Pearson's Goodness-of-Fit Statistic when Expected Frequencies are Small.
- III Kraus, K. (2012) Measurement of Fit in Categorical Data Analysis.
- IV Kraus, K. (2012) A Note on Two Sided Goodness-of-Fit Testing.

Contents

1	Introduction	7
1.1	Categorical Data	7
1.2	Pearson's Goodness-of-Fit Statistic	9
1.3	Sparseness	10
1.4	Contributions of the Thesis	13
2	Summary of Papers	15
2.1	Paper I: On the Computation of Pearson's Goodness-of-Fit Statistic for Sparse Contingency Tables	15
2.2	Paper II: An Alternative to Pearson's Goodness-of-Fit Statistic for Sparse Contingency Tables	17
2.3	Paper III: Measurement of Fit in Categorical Data Analysis	18
2.4	Paper IV: A Note on Two Sided Goodness-of-Fit Testing	19
3	Acknowledgements	21
	References	23

1. Introduction

1.1 Categorical Data

Social scientists often deal with categorical data. Categorical variables can have two types of scales, nominal and ordinal. Nominal variables have categories that can not be naturally ordered. Examples for nominal variables are gender (female, male), hair color (brown, blonde, red, brunette, etc.), favorite subject in school (math, English, science, history, etc.), or occupation. No intrinsic ordering of the categories can be made. The order of the categories can be changed without any loss of information. When there is a natural order of the categories, a variable is called ordinal. Examples for ordinal variables are educational experience (elementary school graduate, high school graduate, college graduate, university graduate), liking of a product (dislike very much, dislike moderately, dislike slightly, neither like nor dislike, like slightly, like moderately, like very much), or agreement to a statement (disagree strongly, disagree, agree, agree strongly). The categories of an ordinal variable can be ordered but, as for nominal variables, numerical values can not be assigned to the categories because there is no meaningful interpretation of these values. The spacing between the categories of an ordinal variable might not be equal for the levels of the variable. The difference between 'neither like nor dislike' and 'like slightly' may not be the same as the difference between 'like moderately' and 'like very much'. Categorical variables have in common that the number of categories is limited. A researcher investigating categorical data has to base her analysis on the observed frequencies of the categories. Statistical inference based on categorical data requires furthermore some assumptions on the data generating process, as the distribution of the investigated variables.

One of the most important distributions in the context of categorical variables is the multinomial distribution. Consider the following situation: In each of a series of n identical and independent trials, exactly one of C mutually independent outcomes is observed. C denotes the fixed finite number of categories for variable Y . Let $\pi_c = P(Y_{ic} = 1)$ be the probability of outcome in category c for each trial. The vector $\pi = (\pi_1, \pi_2, \dots, \pi_C)$ summarizes these probabilities. One multinomial trial can be expressed as $y_i = (y_{i1}, y_{i2}, \dots, y_{iC})$ with $y_{ic} = 1$ if category c is the outcome for trial i and $y_{ic} = 0$ otherwise. The observed frequency for category c is denoted as n_c with $n_c = \sum_i y_{ic}$ being the number of trials that result in an outcome in category c . The vector (n_1, n_2, \dots, n_C) is said to follow a multinomial distribution with parameters n and π . For a multinomial distributed variable it applies that the expected number of times category

Table 1.1. A 2×3 contingency table. (Imaginary data.)

		Y			total
		$j = 1$	$j = 2$	$j = 3$	
X	$i = 1$	10	12	8	30
	$i = 2$	6	12	14	32
total		16	24	22	

c is observed is $E(n_c) = n\pi_c$ with $\text{var}(n_c) = n_c\pi_c(1 - \pi_c)$. The binomial distribution is a special case of the multinomial distribution when the considered variable has only two categories, that is, the variable is dichotomous.

The relationship between two categorical variables can be illustrated by means of a two-dimensional contingency table. Let X be a variable with I categories and let Y be a variable with with J categories. Let the I rows of a table represent the I categories of variable X and the J columns of the table represent the J categories of variable Y . Let further the cells of the table present the IJ outcomes of the joint distribution of X and Y . When the table contains the frequencies of the cells, the table is called a contingency table. The totals for the rows and columns represent the marginal distributions of X and Y . Table 1.1 shows a 2×3 contingency table for two variables X and Y with $I = 2$ categories for variable X and $J = 3$ categories for variable Y . Variable X could for example be a treatment variable where the two categories are 'treatment' and 'no treatment'. Variable Y could be a result variable with categories 'improvement', 'no change', and 'deterioration'. The table shows imaginary data for 62 participants. It can be seen that ten participants of the study got treatment and an improvement was observed. For eight participants, deterioration was observed despite getting the treatment.

In practise, often more than two categorical variables are considered at the same time. For example in surveys, respondents are asked several questions on some concepts as their mood status, their liking of a new product or political attitudes. An example questionnaire on mood status may consist of three items or statements: 'I feel happy today', 'Today is a good day', and 'I feel excited'. Response alternatives could then be given as 'disagree strongly', 'disagree', 'agree' and 'agree strongly'. The respondent is asked to choose a response alternative for each question. In the considered situation, the respondent can choose from four response alternatives for each of the three questions. In total, the respondent has $4 \times 4 \times 4 = 64$ possibilities to respond to the statements. It can also be said that there are 64 response patterns in the present situation. Usually, a survey consists of many more than three questions. The number of possible response patterns increases substantially when the number of questions or response alternatives increases. Consider z observed variables X_i , $i = 1, \dots, z$, that allow for $c_i = 1, \dots, C_i$ response alternatives. The total

number of response patterns is then

$$R = \prod_{i=1}^z C_i.$$

When ten variables are considered with four response alternatives each, the number of response patterns increases to 1,048,576. All inference in the context of categorical data is based on the frequencies of the response pattern. On one hand, frequencies for the response patterns are observed in the sample. The probabilities of the response pattern can, on the other hand, for example in the context of factor analysis, be modelled. Model implied probabilities of the response patterns can then be expressed as functions of the parameters of the model. It is easy to understand that the computational burden for maximum likelihood increases with the number of response patterns for full information estimation methods (e.g., Jöreskog & Moustaki, 2001). Note, however, the conceptual similarities between the situations of one, two, or more categorical variables. The R response patterns in a situation with z categorical variables can be seen as the cells of a z -dimensional contingency table. The response patterns, or the cells of a contingency table can also be regarded as categories of a multinomial variable.

1.2 Pearson's Goodness-of-Fit Statistic

Inference based on categorical data is done in two steps. First, parameter values are estimated according to a prespecified structural model, for example, a factor model or a latent class model. Then, the goodness of the model fit has to be evaluated to see how well the model with the estimated parameter values fits the data. A test for the goodness of a model fit is conducted to compare the model implied frequencies of the response patterns with the corresponding observed frequencies in the sample. The null hypothesis is that the observed frequencies equal the model implied frequencies of the response patterns. When the null hypothesis is true, the expected values of the observed frequencies n_r , $r = 1, \dots, R$, called expected frequencies, are $E(n_r) = N\hat{\pi}_r$. The most widely used statistic for the measurement of fit for categorical data is Pearson's goodness-of-fit statistic. It is defined as:

$$GF = \sum_{r=1}^R \frac{(n_r - N\hat{\pi}_r)^2}{N\hat{\pi}_r},$$

where N denotes the sample size, n_r is the observed frequency of response pattern r , $r = 1, \dots, R$. The expected frequency of response pattern r is denoted as $N\hat{\pi}_r$ where the probability $\hat{\pi}_r = \pi(\hat{\theta})$ is a function of the estimated parameter values, summarized in vector $\hat{\theta}$. Under the null hypothesis, Pearson's

goodness of-fit statistic is asymptotically chi-square distributed with $R - 1 - k$ degrees of freedom:

$$GF \sim \chi_{R-1-k}^2,$$

where k denotes the number of parameters that have to be estimated. Hence, the null hypothesis is rejected, that is, the model is said not to fit the data, when the observed value of GF exceeds a critical value derived from the chi-square distribution with $df = R - 1 - k$.

Other fit statistics have been proposed that use different measures of the difference between the observed and the expected frequencies of response patterns. The most prominent alternative to Pearson's goodness-of-fit statistic is the likelihood ratio test statistic. The two fit statistics are asymptotically equivalent and the likelihood ratio test statistic is often regarded as on a par with Pearson's goodness-of-fit statistic. Irrespective of the applied fit statistic it is implicitly assumed that the differences between the observed and the expected frequencies of the response patterns are smallest when the null hypothesis is true, that is, when the model fits the data. Hence, it is assumed that the value of the fit statistic is smallest under the null hypothesis and that the value of the fit statistic increases when the data generating model and the null model differ. Tests for the goodness-of-fit are therefore conducted one sided.

1.3 Sparseness

It was mentioned above that the number of response patterns increases rapidly when the number of considered variables or the number of categories of the variables increase. Only six items with five response alternatives each allow for 15,625 response alternatives. Ten items with five response alternatives each allow for 9,765,625 response patterns. In practise, the number of response patterns might be considerably larger than the sample size. In these situations, the data are said to be sparse. In the literature, sparseness is often described as small observed or expected frequencies due to small sample sizes in comparison to the number of response patterns. According to Agresti and Yang (1987), "contingency tables are said to be *sparse* when the ratio of the sample size to the number of cells is relatively small" (p. 9). In this thesis, a broader definition of sparseness is applied, based on Agresti (2002) who states that "contingency tables having small cell counts are said to be sparse" (p. 391). Sparseness will here refer to any situation that causes observed or expected frequencies of response patterns to be small.

Sparse data usually contain a considerable number of response patterns with observed frequencies that equal zero. There are mainly two reasons for observed frequencies of zero. The first is of structural nature. When the sample size is smaller than the number of response patterns, it is not possible to have observed frequencies of at least one for each response pattern. The second

reason for observed frequencies of zero is sampling. The number of sampling zeros is partly due to the model and parameters, and partly due to pure sampling. When strong models underlie a sampling process, for example, in factor analysis when factor loadings are high, some response patterns are expected to occur often whereas other response patterns are expected to occur very seldom under the null hypothesis. The amount of sampling zeros can therefore be an indicator for the strength of a model that is underlying the data. On the other hand, respondents might, however, choose some of the response patterns with very small expected frequencies by chance or because of similarities to other response patterns.

Consequences of sparseness for the evaluation of the model fit have been widely investigated and discussed in the literature (e.g., Larntz, 1978; Koehler & Larntz, 1980; Agresti & Yang, 1987; Reiser & VandenBerg, 1994; Jöreskog & Moustaki, 2001). The main issue is that the chi-square approximation does not hold for the distribution of the fit statistics when data are sparse. Pearson's goodness-of-fit statistic is inflated and empirical rejection rates are too high. This is due to very large contributions of response patterns with small expected frequencies and observed frequencies of at least one. Unfortunately, "there is a lack of universal agreement on what constitutes a small expected frequency" (Reiser and VandenBerg, 1994, p. 87). Hence, there is no universal recommendation on when the chi-square approximation can be applied for the test of a model fit and when the chi-square approximation does not hold. The most famous recommendation is the one of Cochran (1950) that most of the response patterns should have expected frequencies of at least five to use Pearson's goodness-of-fit statistic with the asymptotic chi-square approximation. Reiser and VandenBerg give a brief overview of other suggestions for the use of the chi-square approximation. Guidelines are also given by Koehler and Larntz (1980) and Agresti and Yang (1987). Other fit statistics, as the likelihood ratio test statistic are not defined for response patterns with observed frequencies of zero. These response patterns do not contribute to the value of the likelihood ratio test statistic irrespective of the expected frequencies of the response patterns. Some authors, as Koehler and Larntz (1980), describe the contribution of response patterns with observed frequencies of zero to the likelihood ratio test statistic as a limit that equals $2N\hat{\pi}_r$. Other authors, as Reiser and VandenBerg (1994) set their contribution to zero.

A variety of remedies have been suggested to solve the issues that emerge when data are sparse. Agresti and Yang (1987), as well as Reiser and VandenBerg (1994), and Jöreskog and Moustaki (2001) discuss a variety of ideas that have been suggested in the literature to improve the measurement of model fit for sparse contingency tables. These ideas include (i) adding constants to cells, (ii) collapsing cells, (iii) considering only cells with observed or expected frequencies that exceed a certain values, and (iv) deriving the small sample distribution of a fit statistic by means of the bootstrap.

One aspect of inference that is not discussed in this thesis is the estimation of parameter values that can be problematic for sparse data. In the literature it has been suggested to add small constants to the cells of a contingency table. Goodman (1970) recommends to add small constants to each cell. Grizzle, Starmer, & Koch (1969) suggest adding to each cell with an observed frequency of zero the reverse of the number of cells in a contingency table. Agresti and Yang (1987) investigate the effects of adding cell constants on fit statistics and find that this approach causes havoc for the use of Pearson's goodness-of-fit statistic. They interpret the findings as that the adding of cell constants has a smoothing effect towards independence. Hence, the statistic is affected to be more conservative, rejecting the null hypothesis too seldom. They conclude that "if adding a constant is necessary to ensure existence of estimates, it may be preferable to select a very small constant and it is wise to try constants of various size to assess the dependence on the result on that choice" (Agresti & Yang, 1987, p. 20).

Another widespread attempt to decrease the impact of small expected frequencies on the value of the fit statistic is to collapse categories or response patterns. The number of response patterns can be reduced by collapsing two or more categories, for example, 'agree' and 'agree strongly', for one or several variables of a survey. By combining categories for all variables, the number of response patterns can be reduced substantially and the expected frequencies of the individual response patterns increase. However, there is no guarantee that the expected frequencies increase to such an amount that the fit statistic will not be affected by small expected frequencies of individual response patterns. When there is a large number of response patterns from the beginning, it is likely that the number is not reduced enough to eliminate all very small expected frequencies. Reiser and Vandenberg (1994) note that "combining cells is more successful when the extent of sparseness is not widespread" (p. 87). A further reason against collapsing categories or response patterns is that the choice which pairs of categories or response patterns to collapse is rather arbitrary. Note also that the categories should be collapsed before parameter estimation to assure that the model is fitted to the same data as it is tested against. This means that a considerable amount of information that is given by the data will not be taken into account because the diversity in the data is reduced.

Considering only response patterns with observed or expected frequencies that exceed a certain value can substantially reduce the number of response patterns that are taken into account. For instance, Jöreskog and Moustaki (2001) calculate Pearson's goodness-of-fit statistic only over response patterns that have observed frequencies of at least one. They also reduce the degrees of freedom by the number of response patterns that have been excluded from the evaluation of the model fit, that is, the response patterns with observed frequencies of zero. Reiser and Vandenberg (1994) investigate this approach of reducing the degrees of freedom for the likelihood ratio test statistic. This

statistic is not defined for response patterns with observed frequencies of zero and, hence, these response patterns do not contribute to the total value of the fit statistic. They found the empirical rejection rates to be considerably too high for this approach and conclude that the approach of reducing the degrees of freedom by the number of response patterns with observed frequencies of zero is not useful when sparseness is severe. Jöreskog and Moustaki investigate further the approach of calculating the fit statistic only over response patterns with observed frequencies that exceed a certain value. The number of response patterns that is included in the evaluation of the model fit can be reduced considerably by this approach. They find evidence for their data set that sparseness adversely affected the value of the fit statistics. However, they criticize the arbitrariness of the choice of a limit for the expected frequencies for the exclusion from the calculation. Further they explain that the fit statistics are no longer chi-square distributed because the parameter estimation was done based on another set of response patterns than the model evaluation.

The chi-square approximation for the distribution of Person's goodness-of-fit statistic, and other statistics, holds for the asymptotic situation, that is, when the sample size goes to infinity for a fixed number of response patterns. In practise, the sample size and the number of response patterns use to be fixed. Hence, it is likely that the asymptotic approximation does not hold. The bootstrap has therefore been suggested to derive the distribution of a fit statistic for small sample situations (e.g., Bollen & Stine, 1992; Collins, Fidler, Wugalter, & Long, 1993). Langeheine, Pannekoek, and van de Pol (1996) argue that resampling from the original sample is not the appropriate way to derive the distribution of fit statistics under the null hypothesis. They explain why parametric bootstrap should be applied instead in the context of model fit evaluation. Von Davier (1997) reports that the bootstrap approach works well for Pearson's goodness-of-fit statistic even for very sparse data.

1.4 Contributions of the Thesis

This thesis deals with several aspects of the measurement of fit when data are sparse. The first concerns the computational burden of calculating Pearson's goodness-of-fit statistic for situations where many response patterns have observed frequencies that equal zero. A simple solution is presented that allows for the computation of the total value of Pearson's goodness-of-fit statistic when the expected frequencies of response patterns with observed frequencies of zero are unknown. Further reasons are discussed that the statistic should be computed based on the complete set of response patterns and not only on response patterns that have observed frequencies of at least one. Additionally, a new fit statistic is presented that is a modification of Pearson's statistic but that is not adversely affected by response patterns with very small expected frequencies. It is shown that the new statistic is asymptotically equivalent to

Pearson's goodness-of-fit statistic and hence, asymptotically chi-square distributed. Further, comprehensive simulation studies are conducted that compare seven asymptotically equivalent fit statistics, including the new statistic. Situations that are considered concern both multinomial sampling and factor analysis. Tests for the goodness-of-fit are conducted by means of the asymptotic and the bootstrap approach both under the null hypothesis and when there is a certain degree of misfit in the data. Results indicate that recommendations on the use of a fit statistic can be dependent on the investigated situation and on the purpose of the model test. In general, the bootstrap approach works well for most of the statistics and results from the literature are confirmed that it has to be advised against the use of the asymptotic approach when data are sparse. It is found that the power varies substantially between the fit statistics and the cause of the misfit of the model. Findings indicate further that the new statistic proposed in this thesis shows rather stable results and compared to the other fit statistics, no disadvantageous characteristics of the fit statistic are found. Finally, in this thesis, the potential necessity of determining the goodness-of-fit by two sided model testing is adverted. Situations are discussed where the value of Pearson's goodness-of-fit statistic might be smaller under an alternative data generating model than under the null model. A simulation study is conducted that investigates differences between the one sided and the two sided approach of model testing. Situations are identified for which two sided model testing has advantages over the one sided approach.

2. Summary of Papers

2.1 Paper I: On the Computation of Pearson's Goodness-of-Fit Statistic for Sparse Contingency Tables

Pearson's goodness-of-fit statistic is probably the most widely used statistic when the model fit for a contingency table has to be derived. The statistic is asymptotically chi-square distributed with $R - 1$ degrees of freedom where R denotes the number of cells of the contingency table. The degree of freedom is further reduced by the number of parameters when parameter values have to be estimated. The most influential rule of thumb states that expected frequencies should exceed five (Cochran, 1954) for the chi-square approximation to hold. In practice, this rule of thumb is often not met. Especially in the context of analyzing z categorical variables that can be summarized in z -dimensional contingency tables, sample sizes are usually smaller than the number of cells of the contingency table. The contingency table is then said to be sparse (Agresti & Yang, 1987). Even when the sample is larger than the number of cells, a contingency table can contain cells with observed frequencies that equal zero, also called empty cells. This can be caused by the underlying data generating model that implies skewness in the data or simply by sampling. In this paper, the term sparse table refers to all kinds of contingency tables that contain empty cells. Besides the issue of inflation of Pearson's goodness-of-fit statistic when contingency tables are sparse and the expected frequencies of some cells are very small, there are two approaches for the handling of empty cells found in the literature. The first approach is the complete approach. The fit statistic is calculated over all cells. This approach is often found in studies when the number of cells is relatively small. Agresti & Yang (1987), for example, base their investigation on a maximum of ten dichotomous variables giving a total of 1,024 cells and Reiser and VandenBerg (1994) regard a 10×10 contingency table as the largest model. The second approach that is found in the literature is based only on cells with observed frequencies of at least one (Jöreskog & Moustaki, 2001; 2006; Jöreskog & Sörbom, 2006). Empty cells are ignored when the fit statistic is calculated and the degrees of freedom is reduced by the number of empty cells. This approach has one major advantage. The computational burden of the calculation of the model fit can be very high when the expected frequencies of a large number of cells have to be computed (Jöreskog & Moustaki, 2001). For the reduced

version of Pearson's goodness-of-fit statistic, expected frequencies need only to be computed for cells with observed frequencies of at least one.

A simple solution to the issue of computational burden for large sparse tables is shown in the paper. The calculation of the values of Pearson's goodness-of-fit statistic can be decomposed in two parts concerning cells with observed frequencies of at least one and empty cells:

$$GF = \sum_{r \in I_1} \frac{(n_r - N\hat{\pi}_r)^2}{N\hat{\pi}_r} + \sum_{r \in I_0} N\hat{\pi}_r, \quad (2.1)$$

where N denotes the sample size, n_r is the observed frequency and $\hat{\pi}_r$ the expected probability of cell r , with $\hat{\pi}_r = \pi(\hat{\theta})$. $\sum_{r \in I_1}$ and $\sum_{r \in I_0}$ denote the sum over the cells with observed frequencies of one or zero respectively. The first part of Equation (2.1) corresponds to the reduced version of GF that is calculated only over the observed cells. The second part of Equation (2.1) can be rewritten as:

$$\sum_{r \in I_0} N\hat{\pi}_r = N - \sum_{r \in I_1} N\hat{\pi}_r.$$

This shows that the contribution of all empty cells to the value of Pearson's goodness-of-fit statistic can easily be derived from the expected frequencies of the cells with observed frequencies of at least one. The computational burden for the calculation of the complete version of Pearson's goodness-of-fit statistic is therefore not noticeable higher than for the reduced version of the fit statistic.

A small simulation study is conducted to investigate whether there are other reasons for the use of the reduced version of Pearson's goodness-of-fit statistic in favor of the complete version. Findings show that the empirical rejection rates are far too high when the reduced version of the fit statistic is used and the sample size is considerably smaller than the number of cells. The results can be explained by the character of the difference between the two versions of Pearson's goodness-of-fit statistic. The value of the fit statistic is reduced by less than the sample size for the reduced version compared to the complete version of the statistic. The degree of freedom, on the other hand, is reduced at least by the difference between the number of cells and the sample size. For very large sparse tables and small sample sizes this can explain the findings of the simulation study. In sum, it is shown in the first paper that there is no reason for the use of the reduced version of Pearson's goodness-of-fit statistic. On contrary, it is advised against the approach of calculating the fit statistic only over cells with observed frequencies of at least one and reducing the degrees of freedom by the number of empty cells.

2.2 Paper II: An Alternative to Pearson's Goodness-of-Fit Statistic for Sparse Contingency Tables

In the second paper, a new statistic is introduced that is asymptotically equivalent to Pearson's goodness-of-fit statistic GF . However, in contrast to GF , the new statistic is not adversely affected by cells with observed frequencies of at least one and very small expected frequencies. The statistic is a modification of Pearson's goodness-of-fit statistic containing both the expected and the observed frequency of a cell in the denominator of the statistic. The statistic is defined as:

$$K = 2 \sum_{r=1}^R \frac{(n_r - N\hat{\pi}_r)^2}{n_r + N\hat{\pi}_r},$$

where n_r denotes the observed frequency and $\hat{\pi}_r$ the expected probability of a cell r , $r = 1, \dots, R$, that is a function of the estimated parameter values of the model, $\hat{\pi}_r = \pi_r(\hat{\theta})$. The sample size is denoted as N . By including both the expected and the observed frequency of a cell in the denominator of the statistic, it is obviated that the denominator becomes less than one for cells with observed frequencies of at least one, and the contribution of a cell is bounded. In the paper, it is shown that K is asymptotically equivalent to Pearson's goodness-of-fit statistic. Hence, K is under the null hypothesis asymptotically chi-square distributed with $R - 1 - k$ degrees of freedom:

$$K \sim \chi_{R-1-k}^2,$$

where k denotes the number of parameter values that have to be estimated.

Three examples are discussed that describe the similarities and differences between the new statistic and Pearson's goodness-of-fit statistic. The third example concerns a real data example from factor analysis. The value of Pearson's goodness-of-fit statistic is very much accounted for by just a few response patterns that have been chosen by only six of the 392 respondents of the sample. It is argued that in practise, a researcher would exclude these response patterns from the analysis. Then, the value of K would exceed the reduced value of Pearson's goodness-of-fit statistic. In practise, it can be difficult to motivate the exclusion of individual response patterns from the determination of the model fit. The decision of which response patterns to exclude may be highly subjective and may have a large impact on the decision considering the fit of a model. The advantage of the new statistic K is that a researcher does not need to decide whether a response pattern that is occasionally observed in the data and that has a small expected frequency should be excluded from the determination of the model fit because the new statistic is not adversely affected by such response patterns.

Table 2.1. *The seven asymptotically equivalent fit statistics that are included in the study and their computation. Each fit statistic is a function of the observed (n_r) and expected ($N\hat{\pi}_r$) frequencies of the response patterns $r = 1, \dots, R$.*

Statistic	Computation
GF	$= \sum \frac{(n_r - N\hat{\pi}_r)^2}{N\hat{\pi}_r}$
K	$= 2 \sum \frac{(n_r - N\hat{\pi}_r)^2}{n_r + N\hat{\pi}_r}$
LR	$= 2 \sum n_r \ln \left(\frac{n_r}{N\hat{\pi}_r} \right)$
CR	$= \frac{2}{\lambda(\lambda+1)} \sum n_r \left[\left(\frac{n_r}{N\hat{\pi}_r} \right)^\lambda - 1 \right]; \lambda = 2/3$
NM	$= \sum \frac{(n_r - N\hat{\pi}_r)^2}{n_r}$
FT	$= 4 \sum (\sqrt{n_r} - \sqrt{N\hat{\pi}_r})^2$
mFT	$= \sum (\sqrt{n_r} + \sqrt{n_r + 1} - \sqrt{4N\hat{\pi}_r + 1})^2$

2.3 Paper III: Measurement of Fit in Categorical Data Analysis

It is the aim of the third paper to compare seven asymptotically equivalent fit statistics for the measurement of fit for categorical data for situations with sparse data. It is of special interest to investigate the performance of the statistic K , proposed in Paper II, in a variety of situations and in comparison to other fit statistics for categorical data. The seven fit statistics that are included in the study are given in Table 2.1.

Three simulation studies are conducted that investigate the performances of the fit statistics for a variety of situations concerning (i) model tests when the model is true or false, (ii) the asymptotic and the bootstrap approach for the determination of the critical values for a decision on the rejection of the null hypothesis and (iii) multinomial sampling and factor analysis. The first study is an extension and continuation of the first study in Larntz (1978) investigating the performances of the fit statistics for simple multinomial sampling for several degrees of sparseness. The study examines empirical Type I error levels when the null hypothesis is true both with the asymptotic approach and the bootstrap approach. The study by Larntz is extended by considering seven instead of three asymptotically equivalent fit statistics, providing a larger variety of skewnesses, larger number of cells and hence, more severe sparseness and smaller minimum cell expectations, and taking both the asymptotic and the bootstrap approach into account. The second simulation study focuses on the power of tests with the seven fit statistics when the bootstrap approach is applied. The sensitivity of the fit statistics to misfit is investigated by generating some of the data according to some prespecified null model and some of the data according to an alternative data generating model, the equiprobable model. Several null models are considered that differ in their skewness parameter and, hence, from the alternative data generating model. The third study

investigates the performances of the seven fit statistics in the context of factor analysis for ordinal data both when the null hypothesis is true and false, and with the asymptotic and the bootstrap approach.

Findings from the present paper confirm results from the literature that the asymptotic approach should not be used for testing the goodness of a model when the data are sparse. The bootstrap approach worked in general well for the situations considered in the three simulation studies. Results indicate that the context of the model test might be important for the choice of the fit statistic. The statistics *GF*, *CR*, and *LR* were found to be extremely sensitive to misfit in the data when the misfit is due to some observations of response patterns with very small expected frequencies. On the other hand, power was found to be rather low for *GF* and *CR* in the context of factor analysis. Based on the findings in the present paper it has to be advised against the use of the fit statistics *NM* and *FT* because of some ambiguous findings. Results for the fit statistic *K* indicate that it has none of the disadvantages of *GF* and *NM* that it can be seen as a modification of. Findings for *K* are rather stable compared to the other fit statistics. However, the chi-square approximation does not hold for *K* either when the data are sparse. Further simulation studies are suggested to investigate conditions for the use of the fit statistics or combinations of them and that investigate the question how differences between the fit statistics can be interpreted.

2.4 Paper IV: A Note on Two Sided Goodness-of-Fit Testing

Statistics that measure the goodness of a model fit for categorical data are usually constructed as measures of the discrepancy between the observed and the expected frequencies. It is implicitly assumed that a function of the difference between the observed and the expected frequencies is smallest under the null hypothesis. When data are generated according to an alternative model, it is assumed that, overall, the discrepancy between the observed and the expected frequencies is larger than under the null model. It is therefore a matter of course that tests for the goodness-of-fit are conducted one sided.

Pearson's goodness-of-fit statistic is often used when the goodness of a model fit is to be determined for categorical data. It is, however, known that Pearson's goodness-of-fit statistic is inflated when contingency tables are sparse (e.g. Larntz, 1978; Agresti & Yang, 1987) and the chi-square approximation does not hold. The inflation of the statistic is due to cells that are observed in the data but that have very small expected frequencies (Jöreskog & Moustaki, 2001). In the literature, the bootstrap approach is suggested to derive the distribution of a fit statistic when data are sparse (Efron & Tibshirani, 1993; Langeheine, Pannekoek, & van de Pol, 1996). Critical values for the rejection of the null hypothesis are obtained by comparing the actual value

of the fit statistic to the bootstrap distribution of the statistic. Again, it is assumed that the values of the fit statistic are in general smallest under the null hypothesis. Model tests are conducted one sided.

In the fourth paper of this thesis, an example illustrates a situation where the values of Pearson's goodness-of-fit statistic are considerably smaller under the alternative hypothesis than under the null hypothesis. As a consequence, two sided model testing is suggested. A small simulation study is conducted that compares the approaches of one sided and two sided model testing for the situation of large, sparse tables when the bootstrap approach is applied. Results indicate that the two sided approach in some situations with very sparse tables is superior to the one sided approach of model testing. This is the case when the probability of observing the group of cells with very high contributions to the fit statistic is lower under the alternative hypothesis than under the null hypothesis.

3. Acknowledgements

Several people have in different ways contributed to this thesis. Without being able to mention all of them and what they have done in particular, I want to express my gratitude to all those who gave me the possibility to complete this thesis.

First and foremost I want to thank Dag Sörbom and KG Jöreskog. It has been a true privilege to have such experienced researchers as supervisors. Amongst many other things, their ability to see the essential in a problem and, hence, to make the complicated seem so easy has been inspirational to me. I am deeply grateful to Johan Lyhagen for all his encouragement and support. He has contributed to this thesis in the most versatile and dedicated way. Thank you!

I will always be grateful to Rolf Steyer and Christof Nachtigall for introducing me to the fascinating world of psychometrics and for believing in me. Discussions with Albert Satorra have been most inspiring for some of the papers in this thesis. Many ideas for this thesis are born during discussions with Joakim Ekström. His encouragement, enthusiasm, patience explaining mathematics to me, and his valuable comments on early versions of this thesis are gratefully acknowledged.

I am grateful to the Department of Statistics, its members and the participants in the seminars of the department for numerous insights, discussions and their comments on earlier versions of the papers in this thesis. A special thanks to my teachers Anders Christoffersson, Rolf Larsson and Anders Ågren. I have always highly enjoyed discussions about and around statistics with my colleagues Lisbeth Hansson, Bertil W. Andersson and Åsa Vernby. Further, I appreciated the companionship with my PhD-student colleagues, especially mentioned Daniel Preve and Nicklas Korsell who shared a large part of this experience with me. Thanks to the administrative staff at the department and to the computer support staff, mentioning in particular Pierre Hjälms who has more than once saved my data from becoming a part of the empty set.

Even aside from the pure scientific conditions, it has been a real pleasure to be a PhD-student at Uppsala University. Thanks to my colleagues from the Department of Information Science and to my friends at Gothenburg's Nation for many warm and exciting memories.

My friends and family have with their loving support contributed to this thesis both through recreational moments and encouraging discussions. The numerous phone calls, 'fikas', family meetings and exciting adventures have been very important to me.

I owe a special thanks to my parents for all their support and encouragement along the way: Papa, ich bin stark, nur mit Dir. Mama, Vertraute und Vorbild. Danke - für Alles!

Finally, the last months have been much easier and fun thanks to Joel who gave me the strength and inner calm to finish this thesis.

Katrin Kraus

Uppsala, May 5, 2012

References

- Agresti, A. (2002). *Categorical Data Analysis*. New Jersey: John Wiley & Sons, 2nd edition.
- Agresti, A. & Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis*, 5, 9–21.
- Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, 21, 205–229.
- Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics*, 10, 417–451.
- Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28, 375–389.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton, Florida: Chapman & Hall/CRC.
- Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226–256.
- Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489–504.
- Jöreskog, K. G. & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36, 347–387.
- Jöreskog, K. G. & Moustaki, I. (2006). Factor analysis of ordinal variables with full information maximum likelihood. Available at <http://www.ssicentral.com/lisrel/techdocs/orfiml.pdf>.
- Jöreskog, K. G. & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.
- Koehler, K. J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75, 336–344.
- Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24, 492–516.
- Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, 73, 253–263.
- Reiser, M. & Vandenberg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, 47, 85–107.
- von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data - results of a Monte Carlo study. *Methods of Psychological Research Online*, 2, 29–48.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 79*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences.



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012

Distribution: publications.uu.se
urn:nbn:se:uu:diva-173768