

MATHEMATIK-ARBEITSPAPIERE

ASYMPTOTIC NORMALITY OF GOODNESS-OF-FIT
STATISTICS FOR SPARSE POISSON DATA

URSULA MÜLLER
GERHARD OSIUS

PREPRINT No. 51

JANUARY 1998



MATHEMATIK-ARBEITSPAPIERE

A: MATHEMATISCHE FORSCHUNGSPAPIERE

ASYMPTOTIC NORMALITY OF GOODNESS-OF-FIT
STATISTICS FOR SPARSE POISSON DATA

URSULA MÜLLER
GERHARD OSIUS

PREPRINT NO. 51

JANUARY 1998

FACHBEREICH MATHEMATIK UND INFORMATIK
UNIVERSITÄT BREMEN

Bibliothekstraße
D-28359 Bremen
Germany

Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Poisson Data

Ursula Müller and Gerhard Osius
Department of Mathematics and Computer Sciences
University of Bremen

Abstract

Goodness-of-fit tests for discrete data and models with parameters to be estimated are usually based on Pearson's χ^2 or the Likelihood Ratio Statistic, which both are included in the family of Power-Divergence Statistics SD_λ . It is known that SD_λ is asymptotically χ^2 distributed for the common sampling schemes, which yield contingency tables being Poisson or conditional Poisson, e.g. product-multinomial, and for an asymptotic approach with the number of cells being fixed. Here a limiting normal distribution of SD_λ for Poisson distributed $J \times K$ tables is presented considering an increasing cells approach, i.e. beside the total size the number of covariable groups J increases, whereas the number of categories K and the number of model parameters remains fixed. In contrast to the "fixed cells" asymptotics an increase of all expected values is not required — the expectations of the cells may be large but need not be, which allows an application of the deduced tests to sparse data. The peculiarity of the here considered approach is that the underlying class of models to test does not specify the marginal distributions of the (covariable) groups and categories — only the associations, i. e. the odds ratios, are modelled with a finite number of parameters. One thus has to deal with an asymptotically infinite number of nuisance parameters.

Key words: goodness-of-fit, Power Divergence Statistics, contingency tables, Poisson data, sparse data, odds ratios.

1 Introduction

The subject of this article are goodness-of-fit tests for discrete data with parameters to be estimated. For this purpose, observed and expected counts for a given parametric model will be compared applying a certain "distance measure", which should be small if the model is true and large if it is not. Of course, the distribution of the distance under the nullhypothesis, i.e. the case that the model holds, is needed in order to check the goodness-of-fit. The best known statistics usually taken for those tests are Pearson's χ^2 and the Likelihood Ratio Statistic ("Deviance"). Cressie and Read (1984) have embedded them in a family of "Power-Divergence Statistics" SD_λ ($\lambda \in \mathbf{R}$), where each member SD_λ is a sum over all deviations between observed and expected counts:

$$SD_\lambda = \sum_{\text{cells}} a_\lambda(\text{observed}, \text{expected})$$

with distance function $a_\lambda : [0, \infty) \times (0, \infty) \rightarrow [0, \infty)$,

$$(x, \mu) \mapsto a_\lambda(x, \mu) = \frac{2 \cdot x}{\lambda(\lambda + 1)} \cdot \left(\left(\frac{x}{\mu} \right)^\lambda - 1 \right) - \frac{2}{\lambda + 1}(x - \mu) \geq 0.$$

The values $\lambda = 0$, where a_0 is defined by continuity, $\lambda = -1/2$ and $\lambda = 1$ indicate known goodness-of-fit statistics:

$$\begin{aligned} a_{-1/2}(x, \mu) &= 4(\sqrt{x} - \sqrt{\mu})^2 && \text{(Freeman-Tukey),} \\ a_0(x, \mu) &= 2\left(x \log x/\mu - (x - \mu)\right) && \text{(Likelihood Ratio),} \\ a_1(x, \mu) &= (x - \mu)^2/\mu && \text{(Pearson's } \chi^2\text{).} \end{aligned}$$

To allow zero observations, which are typical when data are sparse, only values $\lambda \in (-1, \infty)$ will be considered.

The observed data consist of a $J \times K$ contingency table:

group/code (covariables)	categories					sum
	1	...	k	...	K	
1 (z_1)	X_{11}	...	X_{1k}	...	X_{1K}	X_{1+}
⋮	⋮		⋮		⋮	⋮
j (z_j)	X_{j1}	...	X_{jk}	...	X_{jK}	X_{j+}
⋮	⋮		⋮		⋮	⋮
J (z_J)	X_{J1}	...	X_{Jk}	...	X_{JK}	X_{J+}
sum	X_{+1}	...	X_{+k}	...	X_{+K}	X_{++}

with J groups usually represented by different values z_j of a vector of covariables $Z \in \mathbf{R}^M$, K categories $D \in \{1, \dots, K\}$ and observed counts $X_{jk} \in \mathbf{N}_0$ of objects (Z, D) belonging to group j and category k . The sampling scheme considered here is Poisson, i.e. X_{11}, \dots, X_{JK} are independent Poisson distributed random variables, which typically occurs when all available data (Z, D) are collected within a fixed period. Since this proceeding is convenient by the realization of concrete studies, e.g. in epidemiology with D representing different states of a disease, the Poisson model is of particular interest for applications. Above all, it is of central importance for theoretical investigations of contingency tables in general, because the other relevant sampling schemes are conditional Poisson such as multinomial or product-multinomial sampling and can be derived from the Poisson model through fixing of certain marginal sums. Case-control and cohort studies, for example, which are frequently performed in epidemiology, both consider product-multinomial tables (the columns resp. rows are independent multinomials) and hence distribution models being conditional Poisson.

The main interest in the investigation of contingency tables lies in the description of associations within a table rather than in the marginal distribution of covariables and categories. Thus, the models to be tested specify dependencies between covariables and categories by a finite-dimensional parametrized model and leave the marginal distribution arbitrary. Since the distribution of a contingency table is uniquely determined through the marginal distributions and the odds ratios, the actual models of interest turn out to be “odds ratio models”, i.e. only the odds ratios are specified. In this paper now models for the conditional distribution of the categories given the covariable groups will be considered, which will be shown to cover the class of odds ratio models.

It is commonly held that for increasing sample size SD_λ is asymptotically χ^2 distributed under the common sampling schemes. This approach assumes the number of cells $J \cdot K$ to be fixed — hence the number of parameters is finite — and, moreover, an increase of all expected values. These assumptions are often not given, especially not when the data are sparse. The aim of this article is to meet this situation using an “increasing-cells” approach,

i.e. the number of categories K is fixed and the number of groups J increases, and to derive a limiting normal distribution of SD_λ . In particular, the expected values of each cell may be small but need not be. One difficulty in proving such a limiting result is that the models considered do not specify the marginal distribution of the covariable groups. So when the number of groups tends towards infinity, one has to deal with an asymptotically *infinite* number of nuisance parameters.

A number of authors have also discussed the possibility of taking a normal rather than a χ^2 approximation using the increasing-cells approach, e.g. Carl Morris, Peter McCullagh, Gerhard Osius and Dieter Rojek. Morris (1975), who is not concerned with parameter estimation but with given expected values, proves the asymptotic normality of Pearson's χ^2 and the Likelihood Ratio Statistic for multinomial sampling, i.e. for a J -dimensional multinomial vector where J increases to infinity, by making use of the fact that the multinomial distribution is a special conditional Poisson distribution. McCullagh (1986) considers Pearson's χ^2 and the Likelihood Ratio Statistic for Poisson and binomial sampling with all expected values and hence the distribution of the table specified by a *finite*-dimensional parametric model. Osius (1985) derived the asymptotic normality of SD (for a general distance measure) under binomial sampling and later extended these results to the case where the underlying model fails. Rojek (1989) generalized and strengthened these arguments to product-multinomial sampling with the *rows* being J *independent* multinomials (e.g. cohort studies). His limiting result for the nullhypothesis, an outline of its derivation and further supplementary information are given in Osius and Rojek (1992). In terms of expected values, Osius and Rojek examine the same models as this article. Because of the underlying sampling schemes, though — the group sizes are given, they do not have to deal with an increasing number of nuisance parameters. Nevertheless, their work is of special importance for the here considered approach since it provided valuable ideas for the general proceeding. This article will now be based on the dissertation of U. Müller (1997), which treats the described subject in detail. In contrast to that thesis here a slightly generalized result will be presented — an extended model class is admitted, which, however, requires only a few minor modifications. The thesis moreover considers a certain approach to derive goodness-of-fit tests for “column-multinomial” sampling, which is product-multinomial with the *columns* being independent multinomials (case-control studies). This will, however, not be touched here.

In the following an informative presentation of the results will be given. Section 2 and 3 provide the general background by explaining the sampling scheme, the “increasing cells” approach, the parametric models of interest and the estimation, i.e. the way of fitting the data to an assumed model. In section 4 the main result, namely the asymptotic normality of the Power-Divergence Statistics under the nullhypothesis, i.e. the model holds, and the decision rule for the tests deduced will be presented. Thereafter all assumptions will be listed and discussed in part (sec. 5). Section 6 sketches the derivation of the limiting result. Since the considered approach requires comprehensive calculation and argumentation, this can, for reasons of brevity, not be done in full length. Thus only the most important steps will be summarized without going into mathematical details. Readers, who are interested in the complete proofs, should refer to U. Müller (1997).

2 Stochastic Model and Asymptotics

In regard to the comparatively complex asymptotics, first a suitable stochastic model will be presented. This is indispensable in order to explain the way of grouping and to clarify the meaning of the number of (covariable) groups to be a nonstochastic quantity.

As in the introduction let contingency tables $(X_{jk}^n)_{j,k}$ with observed counts of objects

belonging to group $j \in \{1, \dots, J^n\}$ and category $k \in \{1, \dots, K\}$ be given. The asymptotics will now be indicated through a running index n . In the presence of covariables $Z \in \mathbf{R}^M$, which is the normal situation, for each $n \in \mathbf{N}$ a disjoint decomposition of the image space of Z will be considered:

$$Im Z = \bigcup_{j=1}^{J^n} I_j^n \quad \text{with} \quad I_1^n, \dots, I_{J^n}^n \text{ pairwise disjoint.}$$

Hence the covariables groups are specified through the partitions $I_1^n, \dots, I_{J^n}^n$. The natural sampling scheme in order to achieve Poisson distributed contingency tables $(X_{jk}^n)_{j,k}$ is a sample of N^n objects (Z_i, D_i) , $i = 1, \dots, N^n$, each characterized through a covariable vector Z and a category D . These typically arrive by chance within a fixed period and thus yield the size N^n to be a random variable. Proceeding on this sampling scheme and using the notation from above, one gets under nearby assumptions for each cell of a table a Poisson distribution with expected value μ_{jk}^n ,

$$X_{jk}^n = |\{1 \leq i \leq N^n | Z_i \in I_j^n, D_i = k\}| \sim Pois(\mu_{jk}^n) \text{ for every } j, k, n,$$

that is then, when the distribution of the counts is given through independent Poisson processes (see for example Billingsley, 1986, section 23). In particular, all entries of a table are thus stochastically independent. Further, the marginal sums, e.g. X_{j+}^n , and the total size X_{++}^n are again Poisson distributed, now with parameter μ_{j+}^n resp. μ_{++}^n (The subscript “.” will always denote the vector and “+” the sum over the corresponding index). Summing up, in the following Poisson distributed $J^n \times K$ contingency tables $(X_{jk}^n)_{j,k}$ will be studied with

$$\begin{aligned} X_{jk}^n &\sim Pois(\mu_{jk}^n) \text{ for all } j \in \{1, \dots, J^n\}, k \in \{1, \dots, K\}, n \in \mathbf{N}, \\ X_{11}^n, \dots, X_{J^n K}^n &\text{ stochastically independent.} \end{aligned}$$

For the here considered asymptotics will just like for the common “fixed-cells approach” be assumed:

- The expected total sample size tends towards infinity, $\mu_{++}^n \longrightarrow \infty$,
- the dimension M of the covariable vector is fixed,
- the number K of categories is fixed.

The running index n will in practice certainly be identified with the *realized* sample size. Because this quantity is in fact a random variable, it is more useful for theoretical investigations to choose n as a formal index, which increases proportionally to μ_{++}^n , i.e. $\mu_{++}^n = nc + o(1)$ with constant $c > 0$. Additionally will now be supposed:

- The number of groups J^n increases, $J^n \longrightarrow \infty$.

Considering the grouping just described, one basic requirement to accomplish the increasing-cells approach is the existence of a sequence of decompositions $\bigcup_{j=1}^{J^n} I_j^n$, which increase in number and where for each partition the probability of getting filled is positive, i.e. $P(Z \in I_j^n) > 0$ for all j, n . This is, for example, not given if the distribution of the covariables is discrete with finite domain. As a stronger condition one will even have to demand that asymptotically all groups be filled with probability one (see cond. (LC0), sec. 5). Because the marginal distribution concerning the covariables will not be specified by a parametric model, further conditions concerning grouping resp. the underlying probabilities will have to be satisfied, which will be given in section 5.

3 Parametric Modelling and Estimation

Keeping the notation of the last sections, now the actual models of interest, for which the goodness-of-fit shall be tested, will be described. For this purpose let the table of expectations $(\mu_{jk}^n)_{j,k} \in (0, \infty)^{J^n \times K}$ be considered. Of primary interest in the study of contingency tables is the conditional probability $\pi_{jk}^n = P(D = k | Z \in I_j^n)$, which, in the here considered case of Poisson distribution, equals the ratio μ_{jk}^n / μ_{j+}^n :

$$\pi_{jk}^n = P(D = k | Z \in I_j^n) = \mu_{jk}^n / \mu_{j+}^n \quad \text{for } j = 1, \dots, J^n, k = 1, \dots, K, n \in \mathbf{N}.$$

These probabilities will be modelled in dependence on a finite-dimensional parameter vector θ^n , i.e. $\pi_{jk}^n = \pi_{jk}^n(\theta^n)$. In applications, where the single groups are typically represented by covariable vectors $z_j^n \in \mathbf{R}^M$, in general explicitly

$$\pi_{jk}^n(\theta^n) = F_k(z_j^n, \theta^n) \quad \text{for every } j, k, n$$

is assumed with F_1, \dots, F_K being given functions.

If beside the modelled ratios the expected row sums $\mu_{1+}^n, \dots, \mu_{J^n+}^n$ are taken into account as additional parameters — hence the marginal distribution of the covariables $\mathcal{L}(Z)$ is not specified — one obtains the following representation of the expectations (let $\Theta \subset \mathbf{R}^S$ be an open parameter space):

$$\mu_{jk}^n(\theta^n) = \mu_{j+}^n \pi_{jk}^n(\theta^n) \quad \text{with } \pi_{jk}^n(\theta^n) = \frac{\mu_{jk}^n(\theta^n)}{\mu_{j+}^n}, \theta^n \in \Theta. \quad (1)$$

The hypothesis to check with a goodness-of-fit test now says that the model is true:

$$H_0 : \exists \theta_0^n \in \Theta : \mu_{jk}^n = \mu_{j+}^n \pi_{jk}^n(\theta_0^n) \quad \text{for all } j, k, n. \quad (2)$$

Models for dependencies within contingency tables are of special interest in epidemiology, where the π_{jk}^n are typically disease risks and the categorical variable D indicates different disease groups. For those investigations, usually a more specific parametrization is chosen, which will now be briefly described. Starting point is a *log linear model* with linear predictor $\eta_{jk}^n = \log \mu_{jk}^n$ and the following parametrization of the complete model:

$$\log \mu_{jk}^n = \eta_{jk}^n = \alpha^n + \rho_j^n + \gamma_k^n + \psi_{jk}^n.$$

Interpretability and uniqueness of the parameters are given through suitable marginal conditions. Choosing especially $\rho_1^n = \gamma_1^n = 0$, $\psi_{j1}^n = 0$, $\psi_{1k}^n = 0$ for all j, k, n , the parameters ψ_{jk}^n turn out to be the “log odds ratios”,

$$\psi_{jk}^n = \eta_{jk}^n + \eta_{11}^n - \eta_{j1}^n - \eta_{1k}^n = \log \frac{\mu_{jk}^n \cdot \mu_{11}^n}{\mu_{j1}^n \cdot \mu_{1k}^n},$$

which, as already mentioned in the introduction, contain the whole information about the associations between covariables and categories. Of special importance are further the parameters $\gamma_2^n, \dots, \gamma_K^n$, which have the following interpretation ($k = 2, \dots, n$):

$$\gamma_k^n = \log \frac{\mu_{1k}^n}{\mu_{1+}^n} = \log \left(\frac{\mu_{1k}^n}{\mu_{1+}^n} \cdot \frac{\mu_{1+}^n}{\mu_{11}^n} \right) = \log \frac{P(D = k | Z \in I_1^n)}{P(D = 1 | Z \in I_1^n)}. \quad (3)$$

A suitable parametric model in order to accomplish (1) is now as follows:

$$\eta_{jk}^n = \alpha^n + \rho_j^n + \gamma_k^n + \psi_{jk}^n(\beta) \quad \text{for all } j, k, n \quad (4)$$

(“odds ratio model”) with the commonly considered special case

$$\eta_{jk}^n = \alpha^n + \rho_j^n + \gamma_k^n + \langle z_j^n, \beta_k \rangle \quad (\beta_k \in \mathbf{R}^M) \quad \text{for all } j, k, n$$

(“log linear odds ratio model”). Here $\rho_1^n = \gamma_1^n = 0$, $\beta_1 = 0$ and further (without loss of generality) $z_1^n = 0$ are set in order to meet the marginal conditions. This particular odds ratio model, $\psi_{jk}^n = \langle z_j^n, \beta_k \rangle$, assumes linear dependencies between the covariables and the different stages of a disease. An equivalent rewriting now reveals (4) to be a *multivariate logit model*:

$$\text{logit}_k \pi_j^n := \log \pi_{jk}^n - \log \pi_{j1}^n = \gamma_k^n + \psi_{jk}^n(\beta).$$

In particular, the ratio $\pi_{jk}^n = \mu_{jk}^n / \mu_{j+}^n$ states in accordance with (1) explicitly as follows:

$$\pi_{jk}^n = \pi_{jk}^n(\theta^n) = \frac{\exp(\gamma_k^n + \psi_{jk}^n(\beta))}{\sum_{l=1}^K \exp(\gamma_l^n + \psi_{jl}^n(\beta))}. \quad (5)$$

In the log linear odds ratio model this clearly results in

$$\pi_{jk}^n = \pi_{jk}^n(\theta^n) = \frac{\exp(\gamma_k^n + \langle z_j^n, \beta_k \rangle)}{\sum_{l=1}^K \exp(\gamma_l^n + \langle z_j^n, \beta_l \rangle)} \quad (6)$$

with $\theta^n = (\gamma_2^n, \dots, \gamma_K^n, \beta_2, \dots, \beta_K) \in \mathbf{R}^S$ ($S = (K-1) \cdot (M+1)$) and $\beta = (\beta_2, \dots, \beta_K)$ being the odds ratio parameters of interest. Considering the interpretation of γ_k^n ($k = 2, \dots, n$) given in (3), these parameters — and thus θ^n — should be allowed to depend on n to avoid additional restrictions, which especially concern the choice of the first covariable groups I_1^n ($n \in \mathbf{N}$).

For the estimation of θ_0^n , in the following the maximum likelihood (ML) estimator $\hat{\theta}^n$, or some equivalent estimation function in regard to the approximability through information matrix and scores, will be taken. The loglikelihood function $l^n(\theta)$ and the score vector $U^n(\theta)$ are given as follows:

$$l^n(\theta) = \sum_{j=1}^{J^n} \left(\sum_{k=1}^K X_{jk}^n \log \mu_{j+}^n + \sum_{k=1}^K X_{jk}^n \log \pi_{jk}^n(\theta) - \mu_{j+}^n - \sum_{k=1}^K \log X_{jk}^n! \right)$$

with loglikelihood kernel $\sum_{j=1}^{J^n} \sum_{k=1}^K X_{jk}^n \log \pi_{jk}^n(\theta)$ (with respect to θ),

$$U^n(\theta) = \sum_{j=1}^{J^n} U_j^n(\theta) = \sum_{j=1}^{J^n} \sum_{k=1}^K X_{jk}^n D_\theta^T \log \pi_{jk}^n(\theta) = D_\theta^T l^n(\theta).$$

The ML estimator $\hat{\theta}^n$ there clearly has to fulfil $U^n(\hat{\theta}^n) = 0$. The information matrix under the nullhypothesis is also obtained by simple calculations:

$$I^n(\mu_{+}^n, \theta_0^n) = \text{Cov}(U^n(\theta_0^n)) = \sum_{j=1}^{J^n} \mu_{j+}^n + \sum_{k=1}^K \frac{1}{\pi_{jk}^n(\theta_0^n)} D_\theta^T \pi_{jk}^n(\theta_0^n) \cdot D_\theta \pi_{jk}^n(\theta_0^n).$$

Since for the vector of expectations μ_{+}^n no structure is specified, the observed counts in every group, namely the row sums X_{j+}^n ($j = 1, \dots, J^n$), will be used for its estimation. These sums are easily seen to be ML estimators, too.

In conclusion, the following estimators will be considered:

- $\hat{\theta}^n$ maximum likelihood or some asymptotic equivalent estimator for θ_0^n with
 - $\sqrt{n}(\hat{\theta}^n - \theta_0^n) = O_p(1)$,
 - $\hat{\theta}^n - \theta_0^n = (I^n(\mu_{+}^n, \theta_0^n))^{-1} \cdot U^n(\theta_0^n) + O_p(n^{-1})$,
- $\hat{\mu}_{j+}^n = X_{j+}^n$ ($j \in \{1, \dots, J^n\}$) maximum likelihood estimator for μ_{j+}^n .

The estimators for the expectations in the model thus will be $\hat{\mu}_{j+}^n \pi_{jk}^n(\hat{\theta}^n)$ for all j, k, n .

4 Limit Theorem and Goodness-of-Fit Test

With the model fit now being specified, the test statistic $SD_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n)$ is given ($\hat{\mu}_+^n$ is the vector of row sums as stipulated: $\hat{\mu}_+^n = (\hat{\mu}_{j+}^n)_j = (X_{j+}^n)_j$):

$$SD_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n) = \sum_{j=1}^{J^n} \sum_{k=1}^K a_\lambda(X_{jk}^n, \hat{\mu}_{j+}^n \pi_{jk}^n(\hat{\theta}^n)).$$

For this statistic, suitably scaled and centered, the asymptotic normality under the increasing cells approach will now be presented as the main result of this article. This limiting result is intuitively nearby, since an increasing sum is considered with its components $\sum_{k=1}^K a_\lambda(X_{jk}^n, \hat{\mu}_{j+}^n \pi_{jk}^n(\hat{\theta}^n))$ being nearly independent — the dependencies between the rows just come in through the parameter estimation $\hat{\theta}^n$. The result will, as in U. Müller (1997), be given under the nullhypothesis, i.e. the model holds. A similar limiting normality can also be shown for arbitrary alternatives (the given model fails), which will, however, not be carried out here.

Using the following notation, which will be explained later, namely

$$\begin{aligned} m_\lambda^n(\mu_+^n, \theta_0^n) &= E(SD_\lambda^n(\mu_+^n, \theta_0^n)) = \sum_{j=1}^{J^n} \sum_{k=1}^K E(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n))), \\ c_\lambda^n(\mu_+^n, \theta_0^n) &= \sum_{j=1}^{J^n} \sum_{k=1}^K D_\theta \log \pi_{jk}^n(\theta_0^n) \cdot Cov(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n)), X_{jk}^n), \\ (\sigma_\lambda^n(\mu_+^n, \theta_0^n))^2 &= \sum_{j=1}^{J^n} \sum_{k=1}^K Var(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n))) + 2J^n + \sum_{j=1}^{J^n} \frac{1}{\mu_{j+}^n} \\ &\quad - 2 \sum_{j=1}^{J^n} \frac{1}{\mu_{j+}^n} \sum_{k=1}^K Cov(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n)), (X_{jk}^n)^2) \\ &\quad + 4 \sum_{j=1}^{J^n} \sum_{k=1}^K \pi_{jk}^n(\theta_0^n) Cov(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n)), X_{jk}^n) \\ &\quad - c_\lambda^n(\mu_+^n, \theta_0^n) (I^n(\mu_+^n, \theta_0^n))^{-1} (c_\lambda^n(\mu_+^n, \theta_0^n))^T, \end{aligned}$$

the main result can now be given:

Theorem 4.1 *Keeping the terminology from above and the preceding sections, then for the increasing cells approach described in section 2, the family of Power-Divergence Statistics SD_λ^n ($\lambda > -1$) is asymptotically normal under the nullhypothesis (2), i.e. the model holds, as follows:*

$$T_\lambda^n = \frac{SD_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n) - m_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n) + J^n}{\sigma_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n)} \xrightarrow{\mathcal{L}} N(0, 1) \quad (n \rightarrow \infty),$$

provided the conditions (LC0) – (LC3), (RC0) – (RC3), (MD0) – (MD2) and (VC) given in section 5 are satisfied.

Since large deviations between observed and fitted expectations and hence large values of a_λ resp. SD_λ speak against the nullhypothesis, this limiting result suggests the following one-sided level α test:

$$\text{rejection of } H_0 \quad \Leftrightarrow \quad T_\lambda^n > z_\alpha$$

(z_α is the upper α -quantile of the standard normal distribution $N(0, 1)$). The derivation of the result will be sketched in section 6.

Considering the test statistic from Theorem 4.1, the derived centering term obviously consists not only of the expectation of SD_λ^n , namely m_λ^n evaluated at the estimate $m_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)$. Additionally the number of groups J^n has to be added, which is the expectation of Pearson's χ^2 Statistic for the row sums, i.e. $\sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n)$. By the derivation of a suitable approximation (for which later the asymptotic normality is shown) this term has to be incorporated in order to handle the nuisance arising through the estimation of $\mu_{\cdot+}^n$ respectively of the not specified marginal distribution. Of course it also has to be regarded by computing the asymptotic variance, which turns out to be

$$\begin{aligned} (\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^2 &= \text{Var}\left(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n)\right) \\ &\quad - c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)(I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1}(c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^T, \end{aligned} \quad (7)$$

evaluated at the estimate $\hat{\mu}_{\cdot+}^n, \hat{\theta}^n$. In the formula for $(\sigma_\lambda^n)^2$ given before Theorem 4.1 $\text{Var}(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n))$ is stated in detail. Hence $(\sigma_\lambda^n)^2$ is the variance of the difference between SD_λ^n and Pearson's χ^2 Statistic for the row sums reduced by a quadratic form. The latter one comes into being in order to handle the estimation of the finite-dimensional model parameters and involves the S -dimensional covariance vector of SD_λ^n and the derivative of the loglikelihood (score vector),

$$\begin{aligned} c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) &= \text{Cov}\left(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n), U^n(\theta_0^n)\right) \\ &= \sum_{j=1}^{J^n} \sum_{k=1}^K D_\theta \log \pi_{jk}^n(\theta_0^n) \cdot \text{Cov}(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta_0^n)), X_{jk}^n), \end{aligned}$$

as well as the inverse of the information matrix

$$I^n(\mu_{\cdot+}^n, \theta_0^n) = \text{Cov}(U^n(\theta_0^n)).$$

For Pearson's χ^2 Statistic

$$SD_1^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = \sum_{j=1}^{J^n} \sum_{k=1}^K \frac{(X_{jk}^n - \hat{\mu}_{j+}^n \pi_{jk}^n(\hat{\theta}^n))^2}{\hat{\mu}_{j+}^n \pi_{jk}^n(\hat{\theta}^n)}$$

the standardization terms can be given explicitly, especially $m_1^n = J^n \cdot K$ holds. Writing X_{j+}^n for the estimated expectations of the row sums μ_{j+}^n ($j = 1, \dots, J^n$), in this case ($\lambda = 1$) the test statistic states as follows:

$$T_1^n = \frac{1}{\sigma_1^n(X_{\cdot+}^n, \hat{\theta}^n)} \left(\sum_{j=1}^{J^n} \sum_{k=1}^K \frac{(X_{jk}^n - X_{j+}^n \pi_{jk}^n(\hat{\theta}^n))^2}{X_{j+}^n \pi_{jk}^n(\hat{\theta}^n)} - J^n(K-1) \right)$$

with

$$\begin{aligned} \sigma_1^n(X_{\cdot+}^n, \hat{\theta}^n) &= \left(2J^n(K-1) + \sum_{j=1}^{J^n} \frac{1}{X_{j+}^n} \left(\sum_{k=1}^K \frac{1}{\pi_{jk}^n(\hat{\theta}^n)} + 1 - 2K \right) \right. \\ &\quad \left. - c_1^n(\hat{\theta}^n)(I^n(X_{\cdot+}^n, \hat{\theta}^n))^{-1}(c_1^n(\hat{\theta}^n))^T \right)^{\frac{1}{2}} \end{aligned}$$

and

$$c_1^n(\hat{\theta}^n) = \sum_{j=1}^{J^n} \sum_{k=1}^K D_\theta \log \pi_{jk}^n(\hat{\theta}^n).$$

Except for integer valued λ , $\lambda = 1, 2, 3, \dots$, there is unfortunately no simple expression for the involved terms $m_\lambda^n = E(SD_\lambda^n)$, $Var(SD_\lambda^n)$ and the covariances (e.g. c_λ^n) available. In regard to computational efforts thus Pearson's statistic ($\lambda = 1$) seems preferable — in particular because the expectation $m_1^n = E(SD_1^n)$ does not depend on the model and thus needs not to be estimated.

5 Sufficient Conditions

In the following the conditions for the limit theorem (Th. 4.1) will be listed and briefly discussed. One basic assumption concerns the estimators $\hat{\mu}_{j+}^n = X_{j+}^n$ ($j = 1, \dots, J^n$) for the expectations of the row sums and requires that asymptotically with probability 1 all groups are filled:

$$(LC0) \quad P(\hat{\mu}_{j+}^n > 0 \forall j \in \{1, \dots, J^n\}) \longrightarrow 1.$$

Hence it is directed towards the distribution of the covariables and the way of grouping them. In regard to concrete applications, (LC0) complies with the usual proceeding by the investigation of contingency tables, that is to take only groups into account with at least one observation. It thus clarifies the meaning of J^n , which is by definition a nonstochastic quantity (see sec. 2), to be the number of *observed* groups. From a technical point of view, (LC0) is in need, because $a_\lambda(\cdot, \mu)$ is not defined for $\mu = 0$. Although this can be circumvented by putting the respective cases to zero, there would, however, by an approach without (LC0), several additional difficulties arise.

The following “limiting conditions” are standard assumptions with (LC2) and (LC3) generally being met by the maximum-likelihood estimator:

$$(LC1) \quad \frac{1}{n} I^n(\mu_{\cdot+}^n, \theta_0^n) \longrightarrow I_\infty \text{ positive definite,}$$

$$(LC2) \quad \sqrt{n}(\hat{\theta}^n - \theta_0^n) = O_p(1),$$

$$(LC3) \quad (\hat{\theta}^n - \theta_0^n) = (I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1} U^n(\theta_0^n) + O_p\left(\frac{1}{n}\right).$$

Further some “regularity conditions” for the modelled ratios will be required. First of all the sequence of true parameters must be asymptotically stable:

$$(RC0) \quad \theta_0^n = O(1).$$

This, in particular, has to be satisfied in order to guarantee the existence of a convex compact neighborhood $\bar{W} \subset \Theta$ of almost all θ_0^n . In the odds ratio model (5), where θ_0^n consists of $\gamma_2^n, \dots, \gamma_K^n$ and the odds ratio parameters β , this condition obviously only concerns the γ_k^n having the interpretation $\gamma_k^n = \log \pi_{1k}^n / \pi_{11}^n$ (cp. (3)). The boundedness of these parameters will, however, be guaranteed by the subsequently formulated condition (RC2). (RC0) can thus be dropped in this case.

With \bar{W} chosen as just described, the other regularity conditions state as follows:

$$(RC1) \quad \pi_{jk}^n(\theta) \text{ is continuously differentiable twice in } \theta \text{ for all } j, k, n,$$

$$(RC2) \quad \exists \epsilon > 0 : \pi_{jk}^n(\theta) \geq \epsilon \quad \text{for all } j, k, n, \theta \in \bar{W},$$

$$(RC3) \quad \exists M > 0 : \quad \text{a) } \|D_\theta \pi_{jk}^n(\theta)\| < M \quad \text{for all } j, k, n, \theta \in \bar{W}, \\ \quad \quad \quad \text{b) } \|D_\theta^2 \pi_{jk}^n(\theta)\| < M \quad \text{for all } j, k, n, \theta \in \bar{W}.$$

In the presence of covariables z_1^n, \dots, z_n^n generally $\pi_{jk}^n(\theta) = F_k(z_j^n, \theta)$ with given functions F_1, \dots, F_K (cp. sec. 3) is assumed. These functions and their derivatives are typically continuous in z_j^n and — in accordance with (RC1) — continuous in θ . Hence (RC3) is clearly

fulfilled if the covariables are bounded (For an illustration consider the multivariate logit model (6), for which the derivatives can be determined by simple calculations). Such a requirement is certainly not necessary in this strength. This is also made clear by the fact that the proofs only consider sums over j and hence certain means. In this regard (RC3) could also be relaxed to some extent.

For the expected row sums, and hence for the (not modelled) marginal distribution of the groups, a bounding condition is needed:

$$(MD0) \quad \exists \epsilon > 0 : \mu_{j+}^n \geq \epsilon \quad \text{for all } j, n.$$

Since (RC2) requires similarly $\pi_{jk}^n(\theta_0^n) \geq \epsilon$ for some $\epsilon > 0$ and all j, k, n , both conditions together immediately show the necessity of the true expectations to be bounded away from zero:

$$\exists \epsilon > 0 : \mu_{jk}^n = \mu_{j+}^n \pi_{jk}^n(\theta_0^n) \geq \epsilon \quad \text{for all } j, k, n.$$

Further regarding the marginal distribution respectively the way of grouping, the following assumptions are required to hold:

$$(MD1) \quad \frac{1}{\sqrt{J^n}} \sum_{j=1}^{J^n} \sqrt{\frac{\mu_{j+}^n}{\mu_{++}^n}} \longrightarrow 0,$$

$$(MD2) \quad \frac{1}{\sqrt{J^n}} \sum_{j=1}^{J^n} \frac{1}{\sqrt{\mu_{j+}^n}} \longrightarrow 0.$$

These conditions can also be expressed in terms of sample means. For this purpose, consider the p -th mean

$$M_p(\mu_{.+}^n) = \left(\frac{1}{J^n} \sum_{j=1}^{J^n} (\mu_{j+}^n)^p \right)^{\frac{1}{p}} \quad \text{with } \mu_{.+}^n = (\mu_{1+}^n, \dots, \mu_{J^n+}^n) \in \mathbf{R}^{J^n}, p \in \mathbf{R}.$$

For example, the case $p = -1$ denotes the harmonic, $p = 1$ the arithmetic and $p = 2$ the quadratic mean. With this notation, (MD2) states equivalently as follows:

$$\left(\frac{1}{\sqrt{J^n}} \sum_{j=1}^{J^n} \frac{1}{\sqrt{\mu_{j+}^n}} \right)^{-2} = \frac{1}{J^n} M_{-\frac{1}{2}}(\mu_{.+}^n) \longrightarrow \infty.$$

If the terms of the sum are positive, the p -th mean M_p is increasing in p thus giving

$$\frac{\mu_{++}^n}{(J^n)^2} = \frac{1}{J^n} M_1(\mu_{.+}^n) \geq \frac{1}{J^n} M_p(\mu_{.+}^n) \geq \frac{1}{J^n} M_{-\frac{1}{2}}(\mu_{.+}^n) \quad \text{for all } p \in \left(-\frac{1}{2}, 1\right).$$

(MD2) thus requires a fast increase of all p -th means with $p \geq -1/2$, and due to the choice of n especially $(J^n)^2/n \longrightarrow 0$.

Condition (MD1) now requires

$$\left(\frac{1}{\sqrt{J^n}} \sum_{j=1}^{J^n} \sqrt{\frac{\mu_{j+}^n}{\mu_{++}^n}} \right)^2 = \left(\frac{1}{J^n} M_1(\mu_{.+}^n) \right)^{-1} \cdot \frac{1}{J^n} M_{\frac{1}{2}}(\mu_{.+}^n) \rightarrow 0.$$

Outgoing from (MD2), which implies $(1/J^n M_1(\mu_{.+}^n))^{-1} \rightarrow 0$ and $1/J^n M_{1/2}(\mu_{.+}^n) \rightarrow \infty$, (MD1) thus demands the increase of the arithmetic mean $M_1(\mu_{.+}^n)$ to be much faster than that of the $1/2$ -th mean. Beside that, it is obviously an explicit requirement concerning the distribution of the covariable groups. For an illustration let p_j^n ($j = 1, \dots, J^n$) denote the marginal probabilities of interest: $p_j^n = \mu_{j+}^n / \mu_{++}^n = P(Z \in I_j^n)$. (MD1) is clearly fulfilled if for the limiting distribution given by $(p_j^\infty)_j$ holds $\sum_{j=1}^\infty \sqrt{p_j^\infty} < \infty$. This is, as a simple

example, satisfied if the p_j^∞ denote Poisson probabilities, i.e. $p_j^\infty = P(Z = j) = e^{-\mu} \mu^j / j!$ ($\mu > 0$), and can immediately be verified applying the ratio test for convergence.

In view of applications, (MD1) and (MD2) are approximatively met, if — first of all — the total sample size is much larger than the number of groups. Further, the scaled $-1/2$ -th mean $1/J^n M_{-1/2}(\mu_{\cdot+}^n)$ has to be large, which can in the application case be checked inserting the estimated expectations of the row sums $X_{\cdot+}^n = (X_{j+}^n)_j$. Finally, the arithmetic mean $M_1(\mu_{\cdot+}^n) = \mu_{\cdot+}^n / J^n$ respectively n/J^n should be notably larger than the $1/2$ -th mean $M_{1/2}(\mu_{\cdot+}^n)$.

The last assumption made concerns the variance:

$$(VC) \quad \frac{J^n}{(\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^2} = O(1).$$

It can, however, be relaxed to some extent. Considering formula (7), i.e.

$$\begin{aligned} (\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^2 &= \text{Var} \left(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n) \right) \\ &\quad - c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) (I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1} (c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^T, \end{aligned}$$

it can be shown that the quadratic term disappears in the limit: It holds $c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) = O(J^n)$ and, combined with (LC1), thus $c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) (I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1} (c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^T = O((J^n)^2/n)$, which must converge to zero by (MD2) as seen in the preceding discussion. Hence, instead of (VC), it would suffice to require

$$(VC)' \quad \frac{J^n}{\text{Var}(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n))} = O(1).$$

For Pearson's χ^2 Statistic ($\lambda = 1$), the variance condition can even be dropped, since the relevant part of $(\sigma_1^n)^2$ computes to $2J^n(K-1) + \sum_{j=1}^{J^n} 1/\mu_{j+}^n (\sum_{k=1}^K 1/\pi_{jk}^n(\theta_0^n) + 1 - 2K)$, which obviously satisfies (VC)'.

6 Derivation of the Limit Theorem

In the following an outline of the derivation of Theorem 4.1, which in fact requires comprehensive argumentation, will be given. For reasons of clarity the conditions listed in section 5 will be assumed throughout and not be stated each time needed. An exception will only be made for the variance condition (VC) and the conditions concerning the marginal distribution (MD1) and (MD2), which will only be used for the final conclusions.

Starting point is the centered statistic

$$Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = SD_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - m_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n),$$

which will be gradually approximated through the “true” statistic $Z_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)$ and additional correction terms. The derived approximation does not depend on the estimators $\hat{\mu}_{\cdot+}^n$ and $\hat{\theta}^n$ anymore and thus in particular represents a sum of J^n independent random variables. For this statistic, scaled and recentered — the correction terms have to be regarded — the asymptotic normality can be shown applying the central limit theorem. The standardization terms hence turn out to be the variance and the expected value of the approximated recentered statistic.

In the first step a first order Taylor expansion in $\hat{\theta}^n$ around θ_0^n gives for the centered goodness-of-fit statistic:

$$Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) + D_\theta Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n) + O_p(1). \quad (8)$$

Considering the gradient $D_\theta Z_\lambda^n$ first, which still depends on the estimator for the nuisance parameters, in the second step $\hat{\mu}_+^n$ will be substituted by μ_+^n , using Taylor expansion again. This gives

$$D_\theta Z_\lambda^n(\hat{\mu}_+^n, \theta_0^n) = D_\theta Z_\lambda^n(\mu_+^n, \theta_0^n) + \sum_{j=1}^{J^n} O_p(\sqrt{\mu_{j+}^n})$$

and with assumption $\sqrt{n}(\hat{\theta}^n - \theta_0^n) = O_p(1)$ (LC2) and $n/\mu_{++}^n = O(1)$ thus

$$D_\theta Z_\lambda^n(\hat{\mu}_+^n, \theta_0^n)(\hat{\theta}^n - \theta_0^n) = D_\theta Z_\lambda^n(\mu_+^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n) + \sum_{j=1}^{J^n} O_p(\sqrt{\mu_{j+}^n/\mu_{++}^n}). \quad (9)$$

In the next step, the derived vector can be replaced by its expectation:

$$\begin{aligned} D_\theta Z_\lambda^n(\hat{\mu}_+^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n) &= E(D_\theta Z_\lambda^n(\mu_+^n, \theta_0^n)) \cdot (\hat{\theta}^n - \theta_0^n) + O_p(1) \\ &= -c_\lambda^n(\mu_+^n, \theta_0^n)(\hat{\theta}^n - \theta_0^n) + O_p(1), \end{aligned} \quad (10)$$

where further the conformity of $c_\lambda^n(\mu_+^n, \theta_0^n) = Cov(SD_\lambda^n(\mu_+^n, \theta_0^n), U^n(\theta_0^n))$ as defined in sec. 4 and the negative expectation $-E(D_\theta Z_\lambda^n(\mu_+^n, \theta_0^n))$ was used.

The assumed approximability of the parameter difference through information matrix and scores, i.e. $(\hat{\theta}^n - \theta_0^n) = (I^n(\mu_+^n, \theta_0^n))^{-1}U^n(\theta_0^n) + O_p(1/n)$ (LC3), now makes it possible to pass on to an expression without depending on $\hat{\theta}^n$ anymore:

$$c_\lambda^n(\mu_+^n, \theta_0^n)(\hat{\theta}^n - \theta_0^n) = c_\lambda^n(\mu_+^n, \theta_0^n)(I^n(\mu_+^n, \theta_0^n))^{-1}U^n(\theta_0^n) + O_p(1). \quad (11)$$

Here the error is stochastically bounded, since the order of c_λ^n computes to $c_\lambda^n = O(J^n)$ and already the meaning of J^n as the number of filled groups asserts $J^n/n = O(1)$.

Resuming the preceding approximation steps, thus for the gradient, which is the correction term concerning the parameter estimation $\hat{\theta}^n$ derived in (8), an approximation through a sum of independent variables is given. Going back to (8), the last missing step from $Z_\lambda^n(\hat{\mu}_+^n, \theta_0^n)$ to $Z_\lambda^n(\mu_+^n, \theta_0^n)$ is probably the most difficult one of the whole approximation. To get the error small, a second order Taylor expansion of a_λ in both components is necessary. This yields Pearson's χ^2 Statistic ($\lambda = 1$) for the row sums, namely $\sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n) = \sum_{j=1}^{J^n} (X_{j+}^n - \mu_{j+}^n)^2/\mu_{j+}^n$, as an additional correction term as follows:

$$Z_\lambda^n(\hat{\mu}_+^n, \theta_0^n) = Z_\lambda^n(\mu_+^n, \theta_0^n) - \sum_{j=1}^{J^n} \frac{(X_{j+}^n - \mu_{j+}^n)^2}{\mu_{j+}^n} + \sum_{j=1}^{J^n} O_p(\sqrt{1/\mu_{j+}^n}). \quad (12)$$

Summarizing the approximation steps (8) – (12) and writing explicitly $SD_\lambda^n - m_\lambda^n$ instead of Z_λ^n , leads to the desired sum of independent variables:

$$\begin{aligned} &SD_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n) - m_\lambda^n(\hat{\mu}_+^n, \hat{\theta}^n) \\ &= SD_\lambda^n(\mu_+^n, \theta_0^n) - m_\lambda^n(\mu_+^n, \theta_0^n) - \sum_{j=1}^{J^n} \frac{(X_{j+}^n - \mu_{j+}^n)^2}{\mu_{j+}^n} \\ &\quad - c_\lambda^n(\mu_+^n, \theta_0^n)(I^n(\mu_+^n, \theta_0^n))^{-1}U^n(\theta_0^n) \\ &\quad + O_p(1) + \sum_{j=1}^{J^n} O_p(1)(\sqrt{\mu_{j+}^n/\mu_{++}^n} + \sqrt{1/\mu_{j+}^n}). \end{aligned} \quad (13)$$

Considering this approximated statistic, now, as mentioned, a recentering is required. By definition holds $E(SD_\lambda^n) = m_\lambda^n$ and since $E(U^n) = 0$ the expectation of the last term equals zero. Further, however, $E(\sum_{j=1}^{J^n} (X_{j+}^n - \mu_{j+}^n)^2/\mu_{j+}^n) = J^n$ has also to be taken into account.

If now the stochastic terms are denoted by $\Psi_{\lambda+}^n$, i.e.

$$\begin{aligned}\Psi_{\lambda+}^n &= SD_{\lambda}^n(\mu_{+}^n, \theta_0^n) - \sum_{j=1}^{J^n} \frac{(X_{j+}^n - \mu_{j+}^n)^2}{\mu_{j+}^n} - c_{\lambda}^n(\mu_{+}^n, \theta_0^n)(I^n(\mu_{+}^n, \theta_0^n))^{-1}U^n(\theta_0^n) \\ &= \sum_{j=1}^{J^n} \Psi_{\lambda j}^n\end{aligned}$$

with

$$\Psi_{\lambda j}^n = \sum_{k=1}^K a_{\lambda}(X_{jk}^n, \mu_{j+}^n, \pi_{jk}^n(\theta_0^n)) - \frac{(X_{j+}^n - \mu_{j+}^n)^2}{\mu_{j+}^n} - c_{\lambda}^n(\mu_{+}^n, \theta_0^n)(I^n(\mu_{+}^n, \theta_0^n))^{-1}U_j^n(\theta_0^n),$$

then, in consideration of the expectation J^n of the correction term derived in (12), the approximation (13) can be equivalently rewritten as follows:

$$\begin{aligned}SD_{\lambda}^n(\hat{\mu}_{+}^n, \hat{\theta}^n) - m_{\lambda}^n(\hat{\mu}_{+}^n, \hat{\theta}^n) + J^n \\ = \Psi_{\lambda+}^n - E(\Psi_{\lambda+}^n) + O_p(1) + \sum_{j=1}^{J^n} O_p(1)(\sqrt{\mu_{j+}^n/\mu_{++}^n} + \sqrt{1/\mu_{j+}^n})\end{aligned}\quad (14)$$

with $E(\Psi_{\lambda+}^n) = m_{\lambda}^n(\mu_{+}^n, \theta_0^n) - J^n$. For this statistic, scaled with its standard error $\sigma_{\lambda}^n = \sigma_{\lambda}^n(\mu_{+}^n, \theta_0^n) = (Var(\Psi_{\lambda+}^n))^{1/2}$, a limiting normal distribution

$$\frac{\Psi_{\lambda+}^n - E(\Psi_{\lambda+}^n)}{\sigma_{\lambda}^n} \xrightarrow{\mathcal{L}} N(0, 1)\quad (15)$$

can be shown checking Ljapounov's condition for the central limit theorem, namely

$$\sum_{j=1}^{J^n} E(|\Psi_{\lambda j}^n|^{2+\delta})/(\sigma_{\lambda}^n)^{2+\delta} \longrightarrow 0$$

for some $\delta > 0$. The proof of this, as carried out in U. Müller (1997) for the simple case $\delta = 2$, requires additionally

$$\frac{(J^n)^2 \cdot \sum_{j=1}^{J^n} \sum_{k=1}^K (\mu_{j+}^n \pi_{jk}^n(\theta_0^n))^2}{(\mu_{++}^n)^4} = \left(\frac{J^n}{\mu_{++}^n}\right)^2 \cdot \frac{\sum_{j=1}^{J^n} \sum_{k=1}^K (\mu_{j+}^n \pi_{jk}^n(\theta_0^n))^2}{(\mu_{++}^n)^2} = o(1).$$

This is, however, satisfied, if assumption (MD2) holds, which is needed for the concluding arguments. As discussed in section 5, (MD2) requires an increase of the arithmetic mean, namely $\mu_{++}^n/J^n \rightarrow \infty$ respectively $J^n/\mu_{++}^n \rightarrow 0$. This condition, combined with $\sum_{j=1}^{J^n} \sum_{k=1}^K (\mu_{j+}^n \pi_{jk}^n(\theta_0^n))^2 = \sum_{j=1}^{J^n} \sum_{k=1}^K (\mu_{jk}^n)^2 \leq (\mu_{++}^n)^2$, immediately asserts the desired zero convergence.

The last missing argument, the consistency of $(\sigma_{\lambda}^n(\hat{\mu}_{+}^n, \hat{\theta}^n))^2$ as an estimator for the variance $(\sigma_{\lambda}^n(\mu_{+}^n, \theta_0^n))^2$, can be shown by proving

$$\frac{\sigma_{\lambda}^n(\hat{\mu}_{+}^n, \hat{\theta}^n)}{\sigma_{\lambda}^n(\mu_{+}^n, \theta_0^n)} \xrightarrow{P} 1.\quad (16)$$

For this purpose similar arguments as for the approximation steps (8), (9) and (12) apply thus considering the transitions from $\hat{\theta}^n$ to θ_0^n and $\hat{\mu}_{+}^n$ to μ_{+}^n separately. By this proceeding (cp. U. Müller, 1997, Lemma 6.3), $1/J^n \sum_{j=1}^{J^n} 1/\sqrt{\mu_{j+}^n} \longrightarrow 0$ will be needed as another argument, which is again fulfilled because of the stronger assumption (MD2). Both conditions

concerning the marginal distribution,

$$(MD1) \sum_{j=1}^{J^n} \sqrt{\mu_{j+}^n / \mu_{++}^n} = o(\sqrt{J^n}) \quad \text{and} \quad (MD2) \sum_{j=1}^{J^n} 1 / \sqrt{\mu_{j+}^n} = o(\sqrt{J^n}),$$

which are necessary to handle the error terms caused by the approximation steps concerning the nuisance parameters (9) and (12), will now be used for the final conclusions. Further the variance condition $J^n / (\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n))^2 = O(1)$ (VC) will be needed, which, as mentioned in section 5, is not necessary for Pearson's statistic ($\lambda = 1$). Passing on to the approximation (14) in the first step, using (MD1), (MD2) and (VC) in the second and the asymptotic normality of the approximation (15) combined with the consistency of the variance estimation (16) in the final argument, gives

$$\begin{aligned} & \frac{SD_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - m_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) + J^n}{\sigma_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)} \\ = & \frac{\Psi_{\lambda+}^n - E(\Psi_{\lambda+}^n)}{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)} \cdot \frac{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)}{\sigma_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)} \\ & + \frac{O_p(\sum_{j=1}^{J^n} \sqrt{\mu_{j+}^n / \mu_{++}^n}) + O_p(\sum_{j=1}^{J^n} \sqrt{1 / \mu_{j+}^n}) + O_p(1)}{\sqrt{J^n}} \cdot \frac{\sqrt{J^n}}{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)} \cdot \frac{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)}{\sigma_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)} \\ = & \left(\frac{\Psi_{\lambda+}^n - E(\Psi_{\lambda+}^n)}{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)} + o_p(1) \cdot O(1) \right) \cdot \frac{\sigma_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)}{\sigma_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)} \\ & \xrightarrow{\mathcal{L}} N(0, 1). \end{aligned}$$

Hence the limiting result stated in Theorem 4.1 is established.

In conclusion, it may be said that through the consideration of an “increasing cells” approach, goodness-of-fit tests for the important case of Poisson sampling could be derived, which meet the common situation when data are sparse. Hence a basis for further investigations of the deducing distribution models (*conditional* Poisson models) is provided with case-control studies, i.e. column-multinomial sampling, probably being most interesting in this context — since the distribution of the covariables is not given there and only associations shall be modelled, one has to deal with an asymptotically infinite number of nuisance parameters, too. Beside those extensions, it would certainly also be desirable to derive tests based on higher order approximations such as edgeworth- and saddlepoint-approximations. Instead of relying on asymptotic distributional results, a goodness-of-fit test may also be deduced from the bootstrapped distribution of the statistic T_λ^n , which is (asymptotically) pivotal. Those approaches were, for example, carried out by Osius (1994) for the row-multinomial case and would be advisable in order to improve the goodness of the tests derived.

References

- [1] Billingsley, Patrick (1986): *Probability and Measure*, John Wiley & Sons, New York.
- [2] Cressie, N.A.C. and Read T.R.C. (1984): “Multinomial Goodness-of-Fit Tests”, *Journal of the Royal Statistical Society*, Ser. B, 46, No. 3, 440 – 464.
- [3] Cressie, N.A.C. and Read T.R.C. (1988): *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.
- [4] Habermann, S.J. (1974): *The Analysis of Frequency Data*, The University of Chicago Press, Chicago and London.

- [5] McCullagh, Peter (1985): “Sparse Data and Conditional Tests”, *Bulletin of the International Statistics Institute*, Proceedings of the 45th Session of ISI (Amsterdam), Invited Paper 28.3, 1 – 10.
- [6] McCullagh, Peter (1986): “The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data”, *Journal of the American Statistical Association*, Vol. 81, No. 393, 104 – 107.
- [7] Morris, Carl (1975): “Central Limit Theorems for Multinomial Sums”, *The Annals of Statistics*, Vol. 3, No. 1, 165 – 188.
- [8] Müller, Ursula (1997): “Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Poisson and Case Control Data”, Ph.D. thesis, University of Bremen, Germany.
- [9] Osius, Gerhard (1985): “Goodness-of-Fit Tests for Binary Data With (Possible) Small Expectations but Large Degrees of Freedom”, *Statistics & Decision*, Suppl. No. 2, 213-224.
- [10] Osius, Gerhard and Rojek, Dieter (1992): “Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom”, *Journal of the American Statistical Association*, Vol. 87, No. 420, 1145 – 1152.
- [11] Osius, Gerhard (1994): “Evaluating the Significance Level of Goodness-of-Fit Statistics for Large Discrete Data”. In: Dirschedl, P. and Ostermann, R. (Eds.), *Computational Statistics*, Physica-Verlag Heidelberg, p. 393-417.
- [12] Rojek, Dieter (1989): “Asymptotik für Anpassungstests in Produkt-Multinomialmodellen bei wachsendem Freiheitsgrad”, Ph.D. thesis, University of Bremen, Germany.

Vertrieb der Hefte 4, 14, 23, 26 durch Universitätsbuchhandlung, Bibliothekstr. 3, D-28359 Bremen. Vertrieb der übrigen Hefte (soweit nicht vergriffen) durch die Autoren oder FB 3 Mathematik/Informatik Universität Bremen, Postfach 330440, D-28334 Bremen.

1. Ulrich Krause (1976): Strukturen in unendlichdimensionalen konvexen Mengen, 74 S.
2. Fritz Colonius, Diederich Hinrichsen (1976): Optimal control of hereditary differential systems. Part I, 66 S.
3. Günter Matthiessen (1976): Theorie der heterogenen Algebren, 88 S.
4. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1976): Skript zur Analysis, Band 1 (11. Auflage 2000), 286 S.
5. Wolfgang Schröder (1977): Operator-algebraische Ergodentheorie für Quantensysteme, 59 S.
6. Rolf Röhrig, Michael Unterstein (1977): Analyse multivariabler Systeme mit Hilfe komplexer Matrixfunktionen, 216 S.
7. Horst Herrlich, Hans-Eberhard Porst, Rudolf-Eberhard Hoffmann, Manfred Bernd Wischnewsky (1976): Nordwestdeutsches Kategorienseminar, 193 S.
8. Fritz Colonius, Diederich Hinrichsen (1977): Optimal Control of Hereditary Differential Systems. Part II: Differential State Space Description, 36 S.
9. Ludwig Arnold (1977): Differentialgleichungen und Regelungstheorie, 185 S.
10. Rudolf Lorenz (1977): Iterative Verfahren zur Lösung großer, dünnbesetzter symmetrischer Eigenwertprobleme, 104 S.
11. Konrad Behnen, Hans-Peter Kinder, Gerhard Osius, Rüdiger Schäfer, Jürgen Timm (1977): Dose-Response-Analysis, 206 S.
12. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1978): Minimalbasen polynomialer Moduln, Strukturindizes und BRUNOVSKY-Transformationen, 53 S.
13. Konrad Behnen (1978): Vorzeichen-Rangtests mit Nullen und Bindungen, 53 S.
14. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer, Eberhard Oeljeklaus (1978): Skript zur Linearen Algebra, Band 1 (13. Auflage 2000), 249 S.
15. Günter Ludyk (1978): Abtastregelung zeitvarianter Einfach- und Mehrfachsysteme, 54 S.
16. Momme Johs Thomsen (1977): Zur Theorie der Fastalgebren, 146 S.
17. Klaus Horneffer, Horst Diehl (1978): Modellrechnungen zur anaeroben Reduktionskinetik des Cytochroms P-450, 34 S.
18. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1979): Structure of Topological Categories, 252 S.
19. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part I: Basic categories and functors, 28 S.
20. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part II: Moduletheoretic characterization and reachability, 28 S.
21. Eckart Beutler, Hans Kaiser, Günter Matthiessen, Jürgen Timm (1979): Biduale Algebren, 165 S.
22. Horst Diehl, Detlef Harbach, Jürgen Timm (1980): Planung und Auswertung von Atomabsorptions-Spektrometrie-Untersuchungen mit der Additionsmethode, 44 S.
23. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1981): Skript zur Analysis, Band 2 (6. Auflage 1996), 299 S.
24. Horst Herrlich (1981): Categorical Topology 1971-1981, 105 S.
25. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1981): Special Topics in Topology and Category Theory, 108 S.

26. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1984): Skript zur Linearen Algebra, Band 2 (7. Auflage 1999), 257 S.
27. Rudolf-Eberhard Hoffmann (1982): Continuous Lattices and Related Topics, 314 S.
28. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst (1987): Workshop on Category Theory, 169 S.
29. Harald Boehme (1987): Zur Berufspraxis des Diplommathematikers, 16 S.
30. Jürgen Timm (1986): Mathematische Modelle der Dosis-Wirkungsanalyse bei den experimentellen Untersuchungen der Arbeitsgruppe zur karzinogenen Belastung des Menschen durch Luftverunreinigung, 65 S.
31. Dieter Denneberg (1988): Mathematik für Wirtschaftswissenschaftler. I. Lineare Algebra, 97 S.
32. Peter E. Crouch, Diederich Hinrichsen, Anthony J. Pritchard, Dietmar Salamon (1988, previous edition University of Warwick 1981): Introduction to Mathematical Systems Theory, 244 S.
33. Gerhard Osius (1989): Some Results on Convergence of Moments and Convergence in Distribution with Applications in Statistics, 27 S.
34. Dieter Denneberg (1989): Verzerrte Wahrscheinlichkeiten in der Versicherungsmathematik, Quantilsabhängige Prämienprinzipien, 24 S.
35. Eberhard Oeljeklaus (1989): Birational splitting of homogeneous Albanese bundles, 30 S.
36. Gerhard Osius, Dieter Rojek (1989): Normal Goodness-of-Fit Tests for Parametric Multinomial Models with Large Degrees of Freedom, 38 S.
37. Dieter Denneberg (1990): Mathematik zur Wirtschaftswissenschaft. II. Analysis, 59 S.
38. Ulrich Krause, Cornelia Zahlten (1990): Arithmetik in Krull monoids and the cross number of divisor class groups, 29 S.
39. Dieter Denneberg (1990): Subadditive Measure and Integral, 39 S.
40. Ulrich Krause, Peter Ranft (1991): A limit set trichotomy for monotone nonlinear dynamical systems, 31 S.
41. Angelika van der Linde (1992): Statistical analyses with splines: are they well defined? 22 S.
42. Dieter Denneberg (1992): Lectures on non-additive measure and integral (new edition: Non-additive measure and integral. TDLB 27, Kluwer Academic, Dordrecht (1994)), 114 S.
43. Gerhard Osius (1993): Separating Agreement from Association in Log-linear Models for Square Contingency Tables With Applications, 23 S.
44. Hans-Peter Kinder, Friedrich Liese (1995): Bremen-Rostock Statistik Seminar, 5. - 7. März 1992, 110 S.
45. Dieter Denneberg (1995): Extension of a measurable space and linear representation of the Choquet Integral, 30 S.
46. Dieter Denneberg, Michael Grabisch (1996): Shapley value and interaction index, 20 S.
47. Angelika Bunse-Gerstner, Heike Faßbender (1996): A Jacobi-like method for solving algebraic Riccati equations on parallel computers, 24 S.
48. Hans-Eberhard Porst editor (1997): Categorical methods in algebra and topology - a collection of papers in honour of Horst Herrlich, 498 S.
49. Angelika van der Linde, Gerhard Osius (1997): Estimation of nonparametric risk functions In matched case-control studies, 28 S.
50. Angelika van der Linde (1997): Estimating the smoothing parameter in generalized spline-based regression, 46 S.
51. Ursula Müller, Gerhard Osius (1998): Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, 15 S.
52. Ursula Müller (1999): Nonparametric regression for threshold data, 18 S.
53. Gerhard Osius (2000): The association between two random elements – A complete characterization in terms of odds ratios, 32 S.