Asymptotic normality of goodness-of-fit statistics for sparse Poisson data

Ursula U. Müller and Gerhard Osius
Department of Mathematics and Computer Sciences
University of Bremen
P.O. Box 33 04 40
D–28334 Bremen, Germany

December 21, 2001

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. Both are included in the family of Power-Divergence Statistics SD_λ which are asymptotically χ^2 distributed for the usual sampling schemes. We derive a limiting standard normal distribution for a standardization T_λ of SD_λ under Poisson sampling by considering an approach with an increasing number of cells. In contrast to the χ^2 asymptotics we do not require an increase of all expected values and thus meet the situation when data are sparse. Our limit result is useful even if a bootstrap test is used, because it implies that the statistic T_λ should be bootstrapped and not the sum SD_λ. The peculiarity of our approach is that the models under test only specify associations. Hence we have to deal with an infinite number of nuisance parameters. We illustrate our approach with an application.

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

1 Introduction

In this article we consider goodness-of-fit tests for discrete data with parameters to be estimated. For those tests, observed and expected counts for a given parametric model are compared by applying a certain "distance measure". This should be small if the model is true and large if it is not. Of course, the distribution of the distance under the null hypothesis, i.e., when the model holds, is needed in order to check the goodness-of-fit. The best known statistics are Pearson's χ^2 and the Likelihood Ratio Statistic ("deviance"). Cressie and Read **1**, **2** have embedded them in a family of "Power-Divergence Statistics" SD_λ ($\lambda \in \mathbf{R}$). Each

member SD_λ is a sum over all deviations between observed and expected counts:

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

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

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

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

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

Cressie and Read **2** further suggested $\lambda = 2/3$ as an intermediate value. To allow zero observations, which are typical when data are sparse, we will consider only values $\lambda \in (-1,\infty)$.

The data consist of a $J \times K$ contingency table (see Table 1) of observed counts X_{jk} of objects (Z, D) belonging to group j and category k. The J groups are often represented by different values z_j of a covariate vector $Z \in \mathbf{R}^M$; the variable D denotes K categories.

Table 1: $J \times K$ contingency table

	categories					
group/code			D			
(covariates)	1	$\cdots$	k	$\cdots$	K	sum
$1\,(z_1)$	X_{11}	$\cdots$	X_{1k}	$\cdots$	X_{1K}	X_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$j\,(z_j)$	X_{j1}	$\cdots$	X_{jk}	$\cdots$	X_{jK}	X_{j+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$J\,(z_J)$	X_{J1}	$\cdots$	X_{Jk}	$\cdots$	X_{JK}	X_{J+}
sum	X_{+1}	$\cdots$	X_{+k}	$\cdots$	X_{+K}	X_{++}

We consider Poisson sampling, i.e., $X_{11}, \ldots, X_{JK}$ are independent Poisson distributed random variables. Typical applications are epidemiological studies where the variable D stands for different states of a disease and all available data (Z, D) are collected within a fixed period. Besides its practical meaning, the Poisson model plays a key role for theoretical investigations of contingency tables: The usual sampling schemes derive from the Poisson

model through fixing of certain marginal sums. Important examples are case-control and cohort studies. Both consider product-multinomial tables (the columns respectively 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 distributions of covariates and categories. Thus, the models to be tested specify dependencies between covariates and categories by a finite-dimensional parameterized model and leave the marginal distributions arbitrary. Since the distribution of a contingency table is uniquely determined through marginal distributions and odds ratios, the actual models of interest are "odds ratio models", i.e., only the odds ratios are specified. We will consider models for the conditional distribution of the categories given the covariate groups. They will be shown to cover the class of odds ratio models.

It is known that for increasing sample size SD_λ is asymptotically χ^2 distributed for 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 violated, especially when the expectations of the cells are small ("sparse data"). Since χ^2 tests are standard tools in statistical software packages, they are used even so, especially when odds ratio models are to be checked. We argue that in the sparse data situation a different asymptotic approach with an increasing number of covariate groups J is more realistic. In this article we derive a limiting normal distribution of SD_λ for these asymptotics. In particular, we do not require all expectations to be large. The difficulty in proving such a result is that our models under test do not specify the marginal distribution of the covariate groups. Hence we have to deal with an asymptotically *infinite* number of nuisance parameters.

Our limiting result provides a new tool for goodness-of-fit testing. However, the tests derived are not designed for specific directions of departure and thus are likely to be too optimistic. Hence, the result is, in the first place, useful for comparisons, in particular, to check how the classical χ^2 approximation (and the critical region for the test statistic) relocates when our different asymptotic approach is appropriate.

The problem of χ^2 testing in the sparse data situation was mentioned early, for example by Haldane **3** in the context of a simple parametric model: He derived the exact conditional mean and variance of Pearson's statistic (conditional on marginal sums) for the model of homogeneity, i.e. all odds ratios are one. The possibility of taking a normal rather than a χ^2 approximation using the increasing cells approach was discussed later. Morris **4**, who did not consider estimated but given expected values, proved the asymptotic normality of Pearson's χ^2 and the Likelihood Ratio Statistic for multinomial sampling. He made explicit use of the fact that the multinomial distribution is a special conditional Poisson distribution. McCullagh **5** considered 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, being specified by a *finite*-dimensional parametric model. Osius **5** derived the asymptotic normality of SD (for a general distance measure) under binomial sampling and later extended these results to the case when the underlying model fails. Rojek **7** generalized and strengthened these arguments for product-multinomial sampling with the *rows* being J *independent* multinomials (e.g., cohort studies). His limiting result for the null hypothesis, an outline of its derivation and further supplementary information are given in Osius and Rojek **8**. In terms of expected values, Osius and Rojek examined the same models as this article. Hence several results could be adapted for our approach. Because of the underlying sampling scheme, though — the group sizes are given, they did not have to deal with an increasing number of nuisance parameters. The same applies to the work of Koehler **9** and Dale **10** who studied product-multinomial sampling as well. Koehler derived the limiting normality for the Likelihood Ratio Statistic when the dimension of the parameter vector increases as well. Dale considered Likelihood Ratio and Pearson's χ^2 Statistic, assuming the expectations of all cells to be bounded.

This article has its origin in the thesis of Müller **11** which treats the described subject in detail and, beyond Poisson sampling, suggests an approach to derive goodness-of-fit tests for product-multinomial sampling with the *columns* being independent multinomials (case-control studies). We will focus on Poisson distribution and present a more general result than that thesis by admitting an extended class of models under test.

The paper is organized as follows. Section 2 and 3 provide the general background. We will describe sampling scheme and asymptotics, explain the models under test and specify the estimators. In Section 4 we state our main result, namely the asymptotic normality of the Power-Divergence Statistics under the null hypothesis, i.e., when the model holds. Further, the decision rule for a goodness-of-fit test will be given. Section 5 contains a discussion of the assumptions. It turns out that the marginal distribution of the covariate groups will have to satisfy certain conditions. In particular, the data may not be extremely sparse. In Section 6 we sketch the derivation of the limiting result. Since the considered approach requires comprehensive calculations, the proof cannot be given in full length. Readers interested in technical details should refer to **11**. In Section 7 we discuss our approach by means of a real data application. Section 8 concludes the paper with some final remarks.

2 Stochastic model and asymptotics

In view of the comparatively complex asymptotics, let us describe the stochastic model first. This is necessary to explain the way of grouping and to clarify the meaning of the number of (covariate) groups to be a non-stochastic quantity. The asymptotic will now be indicated through a running index n.

We consider contingency tables $(X^n_{jk})_{j,k}$ with observed counts of objects belonging to group $j \in \{1, \ldots, J^n\}$ and category $k \in \{1, \ldots, K\}$. Each row of a table represents a group and each column a category (see Sec. 1, Table 1). In applications, a group is usually characterized by a certain range of values of a covariate vector $Z \in \mathbf{R}^M$ where Z consists of explanatory variables such as gender, weight or blood pressure. Formally speaking, we consider for each $n \in \mathbf{N}$ a disjoint decomposition of the image space of Z, namely

$$Im\, Z = \bigcup_{j=1}^{J^n} I^n_j \quad \text{with} \quad I^n_1, \ldots, I^n_{J^n} \text{ pairwise disjoint.}$$

Hence, in the presence of covariates, the groups are specified by the partitions $I^n_1, \ldots, I^n_{J^n}$. In practice the groups are often characterized by some representative covariate value z^n_j. For example, if Z is a continuously distributed random variable and I^n_j some interval, then z^n_j might be taken to be the middle of I^n_j.

The natural sampling scheme leading to Poisson distributed contingency tables $(X^n_{jk})_{j,k}$ is an independent sample of N^n individuals (Z_i, D_i), $i = 1, \ldots, N^n$, each characterized through a covariate vector Z and a category D, which arrive by chance within a fixed period. Then, under nearby assumptions, the counts of each cell are Poisson distributed with expected value μ^n_{jk},

$$\begin{aligned} &X^n_{jk} = |\{1 \le i \le N^n | Z_i \in I^n_j, D_i = k\}| \sim Pois(\mu^n_{jk}) \text{ for every } j, k, n, \\ &X^n_{11}, \ldots, X^n_{J^n K} \text{ stochastically independent.} \end{aligned}$$

This applies when the distribution of the counts is given through independent Poisson processes (see, for example, Billingsley **12**, Section 23). The marginal sums, e.g., X^n_{j+}, and the total sample size $N^n = X^n_{++}$ also are Poisson, with parameter μ^n_{j+} and μ^n_{++}, respectively (The subscript "$\cdot$" will always denote the vector and "+" the sum over the corresponding index).

For the asymptotic we assume:

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

For applications, the running index n can be regarded as the *realized* sample size. Since the sample size is stochastic, we treat n as a formal index which increases proportionally to the expected sample size μ^n_{++}, i.e., $\mu^n_{++} = n \cdot const + O(1)$, where *const* denotes some positive constant.

Additionally, and in contrast to the classical χ^2 asymptotics, we suppose:

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

One basic requirement to accomplish the increasing cells approach is the existence of a sequence of decompositions $(\bigcup_{j=1}^{J^n} I_j^n)_n$ where the number of partitions increases and, in particular, each partition has a chance to be filled, i.e., $P(Z \in I_j^n) > 0$ for all j, n. This is, for example, not given if the distribution of the covariates is discrete with finite domain. Even stronger, we will have to demand that asymptotically all groups be filled with probability one (see cond. (LC0), Sec. 5). Further some conditions on the unspecified marginal distribution of the covariates and the way of grouping will have to be satisfied. These are given in Section 5.

3 Parametric modeling and estimation

Keeping the notation of the last sections now the models under test will be described. We consider the table of expectations $(\mu_{jk}^n)_{j,k} \in (0,\infty)^{J^n \times K}$ and the conditional probabilities $\pi_{jk}^{\circ n} = P(D = k | Z \in I_j^n)$ which are of particular interest for applications. In the Poisson model they equal the ratios μ_{jk}^n / μ_{j+}^n,

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

We will model them in dependence on a finite-dimensional parameter vector $\theta \in \Theta$, where Θ is some open parameter space, and call the models briefly $\pi_{jk}^n(\theta)$. The functions $\pi_{jk}^n(\cdot)$ should depend on the category k and on the covariate group j: If, for example, D denotes a disease state, we expect that more people from risk groups will belong to the category "infected" than people who are not exposed. Parametric models for typical hypotheses like this should, of course, be covered by our general class of models $\pi_{jk}^n(\cdot)$. Since the covariate groups or covariate vectors $z_j^n \in \mathbf{R}^M$ may vary with n, we further allow our modeled probabilities to depend on n. In the presence of covariate vectors z_j^n, a simple example of such a model is

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

where $F_1, \ldots, F_K$ denote given functions. Here the dependence on j and n only comes in through the covariates. In particular, for each category (disease state) k a different dependence on the covariates is assumed.

Besides the parameter vector θ for the modeled ratios, we take the expected row sums $\mu_{1+}^n, \ldots, \mu_{J^n+}^n$ into account as additional parameters. This means, in particular, that we do not specify the marginal distribution of the covariate groups. Hence we have $\mu_{jk}^n(\theta) = \mu_{j+}^n \pi_{jk}^n(\theta)$, $\theta \in \Theta$, as a model for the cell expectations.

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

$$H_0: \text{ For each } n \text{ there exists } \theta_0^n \in \Theta \text{ such that } \mu_{jk}^n = \mu_{j+}^n \pi_{jk}^n(\theta_0^n) \text{ for all } j, k. \tag{1}$$

Since $\mu_{jk}^n = \mu_{j+}^n \pi_{jk}^{\circ n}$ we can, equivalently, formulate H_0 in terms of probabilities, i.e., $\pi_{jk}^{\circ n} = \pi_{jk}^n(\theta_0^n)$. The reason why the true parameter vector θ_0^n should depend on n will probably be explained best by the next important example, the odds ratio model.

In applications typically the odds ratios are parameterized. They can be derived from a *log linear model* with linear predictor $\eta_{jk}^n = \log \mu_{jk}^n$,

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

Interpretation and uniqueness of the parameters are given through suitable marginal conditions, e.g., $\rho_1^n = \gamma_1^n = 0$, $\psi_{j1}^{\circ n} = 0$, $\psi_{1k}^{\circ n} = 0$. Under these constraints the parameters $\psi_{jk}^{\circ n}$ equal the log odds ratios,

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

Log odds ratio models assume $\psi_{jk}^{\circ n} = \psi_{jk}^n(\beta)$. A popular example is the log linear odds ratio model with the odds ratios given by the scalar product, $\psi_{jk}^n(\beta) = (z_j^n, \beta_k)$, $\beta_k \in \mathbf{R}^M$. Here $\beta_1 = 0$ and, without loss of generality, $z_1^n = 0$ is set in order to meet the marginal conditions. For each category we again assumed different dependencies on z_j^n by writing β_k. Rewriting shows that the log odds ratio models accomplishes our model for the conditional probabilities from above: Under the null hypothesis we have for the general log odds ratio model

$$\pi_{jk}^{\circ n} = \pi_{jk}^n(\theta_0^n) = \frac{\exp(\gamma_k^n + \psi_{jk}^{\circ n})}{\sum_{l=1}^K \exp(\gamma_l^n + \psi_{jl}^{\circ n})} = \frac{\exp(\gamma_k^n + \psi_{jk}^n(\beta))}{\sum_{l=1}^K \exp(\gamma_l^n + \psi_{jl}^n(\beta))} \tag{2}$$

with $\theta_0^n = (\gamma_2^n, \ldots, \gamma_K^n, \beta)$. In the explicit log *linear* odds ratio model with $\psi_{jk}^n(\beta) = (z_j^n, \beta_k)$ the parameter vector states $\theta_0^n = (\gamma_2^n, \ldots, \gamma_K^n, \beta_2, \ldots, \beta_K) \in \mathbf{R}^S$, $S = (K-1) \cdot (1+M)$. Besides the odds ratio parameter vector β of interest, θ_0^n consists of further nuisance parameters, namely

$$\gamma_k^n = \log \frac{\mu_{1k}^n}{\mu_{11}^n} = \log \Big(\frac{\mu_{1k}^n}{\mu_{1+}^n} \cdot \frac{\mu_{1+}^n}{\mu_{11}^n} \Big) = \log \frac{P(D = k | Z \in I_1^n)}{P(D = 1 | Z \in I_1^n)}, \quad k = 1, \ldots, K. \tag{3}$$

At this point it becomes clear why we let θ_0^n depend on n: We are interested only in associations within a table. Hence we let γ_k^n (and thus θ_0^n) vary with n in order to to avoid additional restrictions concerning the distribution within the first row of the table.

For the estimation of θ_0^n we will take the maximum likelihood estimator $\hat{\theta}^n$, or some equivalent estimation function in regard to the approximability through information matrix and scores (see Sec. 5). The log likelihood function $l^n(\theta)$ and the score vector $U^n(\theta)$ are given by

$$\begin{aligned} l^n(\theta) &= \sum_{j=1}^{J^n} \Big(\sum_{k=1}^K X_{jk}^n \log \mu_{j+}^n + \sum_{k=1}^K X_{jk}^n \log \pi_{jk}^n(\theta) - \mu_{j+}^n - \sum_{k=1}^K \log X_{jk}^n! \Big), \\ U^n(\theta) &= \sum_{j=1}^{J^n} U_j^n(\theta) = \sum_{j=1}^{J^n} \sum_{k=1}^K X_{jk}^n D_\theta^T \log \pi_{jk}^n(\theta) = D_\theta^T l^n(\theta). \end{aligned}$$

The information matrix under the null hypothesis is obtained by simple calculus:

$$I^n(\mu^n_{\cdot+}, \theta^n_0) = Cov(U^n(\theta^n_0)) = \sum_{j=1}^{J^n} \mu^n_{j+} \sum_{k=1}^{K} \frac{1}{\pi^n_{jk}(\theta^n_0)} D^T_\theta \pi^n_{jk}(\theta^n_0) \cdot D_\theta \pi^n_{jk}(\theta^n_0).$$

Since for the vector of expected row sums $\mu^n_{\cdot+}$ no structure is specified, the observed counts in every covariate group will be used for its estimation: $\hat{\mu}^n_{j+} = X^n_{j+}$ for $j = 1, \ldots, J^n$. They are easily seen to be maximum likelihood estimators. In conclusion, the estimators for the expectations in the model will be $\hat{\mu}^n_{j+}\pi^n_{jk}(\hat{\theta}^n)$ for all j, k, n.

4 Limit theorem and goodness-of-fit test

With the fitted expectations being specified, the test statistic $SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)$ is given ($\hat{\mu}^n_{\cdot+}$ is the vector of row sums as stipulated: $\hat{\mu}^n_{\cdot+} = (\hat{\mu}^n_{j+})_j = (X^n_{j+})_j$):

$$SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) \quad = \quad \sum_{j=1}^{J^n} \sum_{k=1}^{K} a_\lambda(X^n_{jk}, \hat{\mu}^n_{j+}\pi^n_{jk}(\hat{\theta}^n)).$$

We now state the main result, the limiting normal distribution of the statistic for the increasing cells asymptotics. Having the central limit theorem in mind, this behavior is not surprising: We consider an increasing sum with J^n nearly independent components — correlations arise since the parameter estimator $\hat{\theta}^n$ is involved. The idea of proof is a standard one: We derive an approximation of the test statistic which does not depend on estimators any more and apply the central limit theorem. The standardization terms in Theorem 1 are expected value and standard error of the approximated statistic, evaluated at its estimates.

The result will be given for the null hypothesis, i.e., when the assumed model holds. A similar normal limit can be derived for arbitrary alternatives. This will, however, not be carried out here.

Theorem 1 *Consider the increasing cells asymptotics described in Section 2 and assume that condition (LC0) – (LC3), (RC0) – (RC3), (MD0) – (MD2) and (VC) given in Section 5 are satisfied. If the null hypothesis (1) holds, i.e., the model fits, the family of Power-Divergence Statistics SD^n_λ ($\lambda > -1$) has a limiting normal distribution as follows:*

$$T^n_\lambda \quad = \quad \frac{SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) - m^{*n}_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} \quad \xrightarrow{\mathcal{L}} \quad N(0,1) \qquad (n \to \infty).$$

The asymptotic expectation m^{*n}_λ is

$$m^{*n}_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) \quad = \quad m^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) - J^n$$

where m^n_λ denotes the expectation of SD^n_λ,

$$m^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \quad = \quad E(SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)) \; = \; \sum_{j=1}^{J^n} \sum_{k=1}^{K} E(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}(\theta^n_0))),$$

and is evaluated at the estimate $(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)$. Additionally, m^{*n}_λ involves the number of groups J^n which is the expectation of Pearson's statistic for the row sums, i.e., $\sum_{j=1}^{J^n} a_1(X^n_{j+}, \mu^n_{j+})$. This term is part of the approximation of $SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)$ and must be incorporated in order to handle the bias caused by the estimation of the marginal distribution of the covariate groups (see Sec. 6 for details). Of course, this also effects the asymptotic variance, i.e., the variance of the approximation, which computes to

$$
\begin{aligned}
&\sigma^{n2}_\lambda(\mu^n_{\cdot+}, \theta^n_0) \\
&= Var\Big(SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) - \sum_{j=1}^{J^n} a_1(X^n_{j+}, \mu^n_{j+})\Big) - c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)^T \qquad (4) \\
&= \sum_{j=1}^{J^n}\sum_{k=1}^{K} Var(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}(\theta^n_0))) + 2J^n + \sum_{j=1}^{J^n} \frac{1}{\mu^n_{j+}} \\
&\quad - 2\sum_{j=1}^{J^n} \frac{1}{\mu^n_{j+}} \sum_{k=1}^{K} Cov(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}(\theta^n_0)), (X^n_{jk})^2) \\
&\quad + 4\sum_{j=1}^{J^n}\sum_{k=1}^{K} \pi^n_{jk}(\theta^n_0) Cov(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}(\theta^n_0)), X^n_{jk}) \\
&\quad - c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)^T.
\end{aligned}
$$

The quadratic form at the end of the formula comes from the estimation of the parameter vector θ^n_0. It involves the S-dimensional vector of covariances between SD^n_λ and the score vector U^n,

$$
\begin{aligned}
c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) &= Cov\Big(SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0), U^n(\theta^n_0)\Big) \\
&= \sum_{j=1}^{J^n}\sum_{k=1}^{K} D_\theta \log \pi^n_{jk}(\theta^n_0) \cdot Cov(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}(\theta^n_0)), X^n_{jk}),
\end{aligned}
$$

and the inverse of the information matrix $I^n(\mu^n_{\cdot+}, \theta^n_0)$.

Since large deviations between observed and fitted expectations, i.e., large values of SD^n_λ, speak against the null hypothesis, the limiting result suggests the following one-sided level α test (z_α denotes the upper α–quantile of the standard normal distribution $N(0,1)$):

$$
\text{rejection of } H_0 \quad \Leftrightarrow \quad T^n_\lambda > z_\alpha.
$$

Except for integer valued λ, the moments involved in T^n_λ (e.g., m^n_λ, c^n_λ) cannot be stated explicitly. Nevertheless, numerical computation of those Poisson expectations as a sum over all relevant outcomes is straightforward. Most convenient in regard to computational efforts is Pearson's statistic ($\lambda = 1$),

$$
SD^n_1(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) = \sum_{j=1}^{J^n}\sum_{k=1}^{K} \frac{(X^n_{jk} - \hat{\mu}^n_{j+}\pi^n_{jk}(\hat{\theta}^n))^2}{\hat{\mu}^n_{j+}\pi^n_{jk}(\hat{\theta}^n).}
$$

Its expectation $m_1^n = E(SD_1^n)$ needs no estimation since it does not depend on the model, $m_1^n = J^n \cdot K$. Hence we have the asymptotic expectation $m_1^{*n} = m_1^n - J^n = J^n(K-1)$. Writing X_{j+}^n instead of $\hat{\mu}_{j+}^n$, the test statistic states in this case:

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

with

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

and

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

Let us compare T_1^n and the Pearson statistic of Osius and Rojek **8**, who consider the same setting as this article, but row–multinomial sampling, i.e., the J^n rows are independent multinomials and the row sums are given, $X_{j+}^n = \mu_{j+}^n = n_j$. Their test statistic states

$$\tilde{T}_1^n = \frac{1}{\tilde{\sigma}_1^n(X_{\cdot+}^n, \hat{\theta}^n)}\Big(\sum_{j=1}^{J^n}\sum_{k=1}^{K}\frac{(X_{jk}^n - X_{j+}^n\pi_{jk}^n(\hat{\theta}^n))^2}{X_{j+}^n\pi_{jk}^n(\hat{\theta}^n)} - \tilde{m}_1^n\Big)$$

with

$$\begin{aligned}\tilde{m}_1^n &= J^n(K-1),\\ \tilde{\sigma}_1^n(X_{\cdot+}^n, \hat{\theta}^n) &= \Big(2J^n(K-1) + \sum_{j=1}^{J^n}\frac{1}{X_{j+}^n}\Big(\sum_{k=1}^{K}\frac{1}{\pi_{jk}^n(\hat{\theta}^n)} - K^2 - 2(K-1)\Big)\\ &\quad - c_1^n(\hat{\theta}^n) I^n(X_{\cdot+}^n, \hat{\theta}^n)^{-1} c_1^n(\hat{\theta}^n)^T\Big)^{1/2}.\end{aligned}$$

The quadratic form at the end of the variance formula has the same analytic form as the corresponding term in our formula. Obviously, the centering terms are the same for both distribution models, namely $J^n(K-1)$. The variances are different. Since we estimate the marginal covariate distribution, the variance in the Poisson model is, as expected, bigger than the variance in the row–multinomial model:

$$\sigma_1^n(X_{\cdot+}^n, \hat{\theta}^n) = \tilde{\sigma}_1^n(X_{\cdot+}^n, \hat{\theta}^n) + (K^2-1)\sum_{j=1}^{J^n}\frac{1}{X_{j+}^n}.$$

This will lead to smaller values of the test statistic. Hence, we expect that tests for Poisson data will be less powerful than tests for row–multinomial sampling.

5 Sufficient conditions

We now give the conditions for our limit theorem, Theorem 1, and discuss them briefly. One basic assumption concerns the estimators $\hat{\mu}^n_{j+} = X^n_{j+}$ $(j = 1, \ldots, J^n)$ of the expected row sums. They are required to be asymptotically nonzero with probability 1,

(LC0) $\quad P(\hat{\mu}^n_{j+} > 0 \; \forall \, j \in \{1, \ldots, J^n\}) \longrightarrow 1,$

i.e., all covariate groups have to be filled. Hence (LC0) corresponds well with practice where only those groups are taken into account which have at least one observation. In particular, it explains the meaning of J^n as the number of *observed* groups. From a technical point of view, (LC0) is in need, because $a_\lambda(\cdot, \mu)$ is not defined for $\mu = 0$.

The following "limiting conditions" are standard assumptions with (LC2) and (LC3) being satisfied by the maximum likelihood estimator under mild conditions:

(LC1) $\quad n^{-1} I^n(\mu^n_{\cdot+}, \theta^n_0) \longrightarrow I_\infty$ positive definite,

(LC2) $\quad n^{1/2}(\hat{\theta}^n - \theta^n_0) = O_p(1),$

(LC3) $\quad (\hat{\theta}^n - \theta^n_0) = I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} U^n(\theta^n_0) + O_p(n^{-1}).$

For the modeled ratios we need some "regularity conditions". First of all, the sequence of true parameters must be asymptotically stable:

(RC0) $\quad \theta^n_0 = O(1).$

This condition guarantees the existence of a convex compact subset $\bar{W} \subset \Theta$ which contains almost all θ^n_0 and, due to the assumed consistency (LC2), almost all $\hat{\theta}^n$. We will need $\bar{W}$ for the proofs, particularly in order to formulate the technical conditions (RC2) and (RC3). In the log odds ratio model (2) condition (RC0) is satisfied. Here θ^n_0 consists of the (constant) odds ratio parameter vector β and of the parameters $\gamma^n_k = \log(\mu^n_{1k}/\mu^n_{11}) = \log(\pi^{\circ n}_{1k}/\pi^{\circ n}_{11})$ $(k = 2, \ldots, K$, see Eq. (3)). The γ^n_k's are bounded by the subsequently formulated condition (RC2) which, in particular, requires that every cell has a chance to be filled and which we need anyway.

We assume that the following regularity conditions hold:

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

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

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

b) $\|D^2_\theta \pi^n_{jk}(\theta)\| < M \quad$ for all $j, k, n, \theta \in \bar{W}.$

In order to understand (RC3), let us consider the parametric model introduced in Sec. 3, $\pi^n_{jk}(\theta) = F_k(z^n_j, \theta)$ with given functions $F_1, \ldots, F_K$ and covariates $z^n_1, \ldots, z^n_{J^n}$. The func-

tions F_k are typically continuously differentiable. Hence, (RC3) is obviously fulfilled if the covariates have a natural bound which is the normal situation in applications. (For an illustration consider the logit model (2) with $\psi_{jk}^n(\beta) = (z_j^n, \beta_k)$, for which the derivatives can be determined by simple calculations.) It should be mentioned that (RC3) could be relaxed to some extent since the proofs only consider sums over j and hence certain means. This would, however, amount to several additional technical conditions, and will, for reasons of clarity, not be carried out here.

For the expected row sums, and hence for the (not modeled) marginal distribution of the covariate groups, we need a bounding condition:

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

This condition, combined with (RC2), particularly implies that the cell expectations μ_{jk}^n must be bounded away from zero. Those bounding conditions will be seen to be essential for our proofs where we will need several auxiliary results about the order of Poisson moments, regarded as a function of the expected value.

So far we only formulated assumptions which are typically satisfied in practice. Now we state two assumptions concerning the marginal distribution of the covariate groups which should be checked carefully. We require

(MD1) $$(J^n)^{-1/2} \sum_{j=1}^{J^n} \Big(\frac{\mu_{j+}^n}{\mu_{++}^n}\Big)^{1/2} \longrightarrow 0,$$

(MD2) $$(J^n)^{-1/2} \sum_{j=1}^{J^n} (\mu_{j+}^n)^{-1/2} \longrightarrow 0.$$

These conditions arise when we replace the estimated by the true covariate distribution and represent the approximation error, which, of course, must disappear in the limit. Both conditions, (MD1) and (MD2), can be stated equivalently in terms of sample means. Consider the p–th mean

$$M_p(\mu_{\cdot+}^n) \;=\; \Big(\frac{1}{J^n} \sum_{j=1}^{J^n} (\mu_{j+}^n)^p\Big)^{1/p}, \quad p \in \mathbf{R},$$

where $p = 1, 1/2$ and $-1/2$ denote arithmetic, square root and inverse square root mean. With this notation we obtain

(MD1)' $$\Big((J^n)^{-1/2} \sum_{j=1}^{J^n} \Big(\frac{\mu_{j+}^n}{\mu_{++}^n}\Big)^{1/2}\Big)^{-2} = \frac{M_1(\mu_{\cdot+}^n)}{M_{1/2}(\mu_{\cdot+}^n)} \to \infty,$$

(MD2)' $$\Big((J^n)^{-1/2} \sum_{j=1}^{J^n} (\mu_{j+}^n)^{-1/2}\Big)^{-2} = \frac{M_{-1/2}(\mu_{\cdot+}^n)}{J^n} \to \infty.$$

The latter condition states that the inverse square root mean of the row sums must increase faster than the number of rows. This clarifies that our approach is intermediate between the

classical "fixed cells" approach, where *each* row sum must increase, and an "extreme sparse increasing cells" approach, where all row sums have a fixed bound. Since the p–th mean is increasing in p, (MD2)' implies $M_1(\mu^n_{\cdot+})/J^n = \mu^n_{++}/(J^n)^2 \to \infty$. Hence, due to the choice of n, $(J^n)^2/n \to 0$ must be satisfied.

Condition (MD1)' requires an increase of the ratio $M_1/M_{1/2}$. This ratio is a measure of dispersion for the roots $r^n_j = (\mu^n_{j+})^{1/2}$ of the row sums. In fact the variance (with respect to point mass $1/J^n$) is given by

$$1/J^n \sum_{j=1}^{J^n} (r^n_j - \bar{r}^n)^2 \quad = \quad M_1(\mu^n_{\cdot+}) - M_{1/2}(\mu^n_{\cdot+}).$$

Hence (MD1)' requires that this variance increases faster than the square $(\bar{r}^n)^2 = M_{1/2}(\mu^n_{\cdot+})$ of the mean,

$$(M_1(\mu^n_{\cdot+}) - M_{1/2}(\mu^n_{\cdot+}))/M_{1/2}(\mu^n_{\cdot+}) \quad \to \quad \infty.$$

In summary, condition (MD2) requires that the inverse square root mean of the row sums increases suitably fast whereas condition (MD1) demands an increasing *variation* of the row sums. The latter is, for example, not given when all row sums of a table are equally sized. Particularly when $\mu^n_{j+} = (J^n)^{1+\alpha}$ for every j ($\alpha > 0$), one easily checks that (MD1) fails and (MD2) is satisfied. A simple example for the opposite situation is the following: Assume that only the first row sum increases, say $\mu^n_{1+} = (J^n)^2$, whereas $\mu^n_{j+} = 1$ for all $j > 1$. In this case (MD1) is satisfied but (MD2) fails.

Since asymptotic conditions can never be verified for concrete samples, a simple rule of thumb would certainly be desirable. Such a rule is, unfortunately, not yet available. For this purpose one needs more evidence, in particular elaborate empirical and simulation studies. This goes beyond the scope of this paper, but would certainly be interesting, also in view of the fact that it is not yet clear whether the sufficient conditions (MD1) and (MD2) are also necessary. We recommend to compute the left-hand side of (MD1)' and (MD2)' for the observed row sums $X^n_{j+} = \hat{\mu}^n_{j+}$ to see whether they may be considered large. An illustration for one particular data situation and a further discussion is given in Section 7.

Our last condition finally concerns the variance which must increase sufficiently fast:

$$\text{(VC)} \qquad \frac{J^n}{\sigma^{n2}_\lambda(\mu^n_{\cdot+}, \theta^n_0)} = O(1).$$

This condition only affects the first part of the variance given in (4) since the quadratic term in that formula has the order $(J^n)^2/n$, and, due to the discussion above, is asymptotically negligible. For Pearson's χ^2 Statistic ($\lambda = 1$) the variance is explicitly given (see end of Sec. 4). Condition (VC) is obviously satisfied in this case.

6 Derivation of the limit theorem

We now sketch the proof of Theorem 1. The idea of proof is to approximate the centered statistic

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

gradually through a statistic which does not depend on the estimators anymore. This approximation is a sum of J^n independent random variables. Hence we can apply the central limit theorem to its standardization. Since the variance must be estimated, too, we finally need the consistency of the variance estimation. In conclusion, this establishes the asserted normality of the Power–Divergence family.

In the following we will assume throughout that all covariate groups be filled with probability 1 (LC0) and that the sequence of true parameters θ_0^n is asymptotically stable (RC0). The last condition implies that there is a convex compact subset $\bar{W}$ of Θ which contains almost all θ_0^n. Since the approximation of the test statistic is based on Taylor expansions, we will also assume that the modeled probabilities are differentiable (RC1).

6.1 Approximation of the test statistic

We start with a first order Taylor expansion in $\hat{\theta}^n$ around θ_0^n which gives for the centered statistic:

$$Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) + D_\theta Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n) + O_p(1). \tag{5}$$

The idea of proof is the following. Let be $\bar{W} \subset \Theta$ a convex compact neighborhood with $\theta_0^n \in \bar{W} \subset \Theta$ for all n. This can, due to (RC0), be assumed without loss of generality. For reasons of definiteness let all observed row sums $\hat{\mu}_{j+}^n = X_{j+}^n$ be nonzero and consider $\hat{\theta}^n \in \bar{W}$. The case that some row sums are zero or that $\hat{\theta}^n \notin \bar{W}$ is asymptotically negligible by assumption (LC0) and the assumed consistency of $\hat{\theta}^n$. The approximation error in (5) computes to

$$\begin{aligned}
&\|Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) - D_\theta Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n)\| \\
&= \|(\hat{\theta}^n - \theta_0^n)^T \int_0^1 (1-z) D_\theta^2 Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_z^n) dz \, (\hat{\theta}^n - \theta_0^n)\| \\
&\leq \|\hat{\theta}^n - \theta_0^n\|^2 \cdot \sum_{j=1}^{J^n} \sum_{k=1}^{K} \sup_{\theta \in \bar{W}} \Big\| D_\theta^2 \Big(a_\lambda(X_{jk}^n, \hat{\mu}_{j+}^n \pi_{jk}^n(\theta)) - e_\lambda(\hat{\mu}_{j+}^n, \pi_{jk}^n(\theta)) \Big) \Big\|.
\end{aligned}$$

We wrote briefly θ_z^n for $\theta_0^n + z(\hat{\theta}^n - \theta_0^n)$. The function e_λ denotes the expected value of a_λ, $e_\lambda(\mu_{j+}^n, \pi_{jk}^n(\theta)) = E(a_\lambda(X_{jk}^n, \mu_{j+}^n \pi_{jk}^n(\theta)))$, where X_{jk}^n is Poisson distributed with expected value $\mu_{j+}^n \pi_{jk}^n(\theta)$. If now for all j, k, n holds

$$\sup_{\theta \in \bar{W}} \Big\| D_\theta^2 \Big(a_\lambda(X_{jk}^n, \hat{\mu}_{j+}^n \pi_{jk}^n(\theta)) \Big) \Big\| \leq \hat{\mu}_{j+}^n \cdot const \quad \text{for all } j, k, n, \tag{6}$$

$$\sup_{\theta \in \bar{W}} \|D_\theta^2 e_\lambda(\hat{\mu}_{j+}^n, \pi_{jk}^n(\theta))\| \leq \hat{\mu}_{j+}^n \cdot const \quad \text{for all } j, k, n, \tag{7}$$

we obtain

$$\begin{aligned}
&\|Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) - D_\theta Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) \cdot (\hat{\theta}^n - \theta_0^n)\| \\
&\leq \|\hat{\theta}^n - \theta_0^n\|^2 \cdot \sum_{j=1}^{J^n} \sum_{k=1}^{K} \hat{\mu}_{j+}^n \cdot const \\
&= \|\hat{\theta}^n - \theta_0^n\|^2 \cdot \hat{\mu}_{++}^n \cdot K \cdot const \\
&= \|n^{1/2}(\hat{\theta}^n - \theta_0^n)\|^2 \cdot \frac{\hat{\mu}_{++}^n}{n} \cdot K \cdot const.
\end{aligned}$$

Since n increases proportionally with μ_{++}^n and $\hat{\mu}_{++}^n/\mu_{++}^n \xrightarrow{P} 1$, we have $\hat{\mu}_{++}^n/n = O_p(1)$. This and condition (LC2), $n^{1/2}(\hat{\theta}^n - \theta_0^n) = O_p(1)$, then establishes (5).

The proof of (6) is straightforward. Besides the generally assumed conditions (LC0, RC0, RC1) only analytical arguments and condition (RC2) apply, i.e. the probabilities must be bounded away from zero on $\bar{W}$ where $\bar{W}$ is a convex subset of Θ as above and exists due to (RC0). Inequality (7) will be verified by showing that for each j, k, n holds

$$\sup_{\theta \in \bar{W}} \|D_\theta^2 e_\lambda(\mu_{j+}^n, \pi_{jk}^n(\theta))\| \leq \mu_{j+}^n \cdot const \quad \text{for all } \mu_{j+}^n \in [\epsilon, \infty)$$

where ϵ is some constant on $(0, 1)$. This suffices since the row sums in (7) were assumed to be nonzero, i.e., $\hat{\mu}_{j+}^n = X_{j+}^n \geq 1$ for all j, k, n. Simple calculus gives

$$\begin{aligned}
&D_\theta^2 e_\lambda(\mu_{j+}^n, \pi_{jk}^n(\theta)) \\
&= \mu_{j+}^n \Big(\frac{\mu_{j+}^n}{\mu_{jk}^n(\theta)} \cdot D_\theta^T \pi_{jk}^n(\theta) \cdot \mu_{jk}^n(\theta) \frac{\partial^2}{(\partial \mu_{jk}^n(\theta))^2} E(a_\lambda(X_{jk}^n, \mu_{jk}^n(\theta))) \cdot D_\theta \pi_{jk}^n(\theta) \Big) \\
&\quad + \mu_{j+}^n \Big(\frac{\partial}{\partial \mu_{jk}^n(\theta)} E(a_\lambda(X_{jk}^n, \mu_{jk}^n(\theta))) \cdot D_\theta^2 \pi_{jk}^n(\theta) \Big).
\end{aligned}$$

We have to show that the two terms in large parentheses are bounded for $\theta \in \bar{W}$. This is obvious for the (derived) probabilities by the assumed bounding conditions (RC2) and (RC3). In order to formulate bounding results for (derivatives of) Poisson expectations like $\partial/(\partial \mu_{jk}^n(\theta))\, E(a_\lambda(X_{jk}^n, \mu_{jk}^n(\theta)))$ for our special asymptotics, we show, in this example,

$$\Big| \frac{\partial}{\partial \mu} E\Big(a_\lambda(X, \mu)\Big) \Big| \quad \leq \quad const \quad \text{for all } \mu \geq \epsilon > 0 \tag{8}$$

where X is Poisson distributed with expected value μ. This suffices since we assume that for $\theta \in \bar{W}$ the cell expectations $\mu_{jk}^n(\theta)$ are bounded away from zero (MD0, RC1). Results like this are crucial for the derivation of our limit theorem and are needed several times. For reasons of brevity we will not go into technical details or state the full list of auxiliary results needed and refer to **11**, Sec. 4.

With Equation (5) being verified, we now want to exchange estimated and true row expectations in the right-hand side of (5). Consider the gradient $D_\theta Z_\lambda^n$ first. We will show

$$D_\theta Z_\lambda^n(\hat{\mu}_{\cdot+}^n, \theta_0^n) \quad = \quad D_\theta Z_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) + \sum_{j=1}^{J^n} O_p((\mu_{j+}^n)^{1/2}). \tag{9}$$

Then, using the $\sqrt{n}$–consistency of $\hat{\theta}^n$ (LC2) and $n/\mu^n_{++} = O(1)$,

$$D_\theta Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0)(\hat{\theta}^n - \theta^n_0) = D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \cdot (\hat{\theta}^n - \theta^n_0) + \sum_{j=1}^{J^n} O_p\Big((\frac{\mu^n_{j+}}{\mu^n_{++}})^{1/2}\Big). \tag{10}$$

The proof of (9) is done by Taylor expansion. By definition we have $D_\theta Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) = D_\theta SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) - D_\theta m^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0)$. Hence the result can be proved treating the two terms separately. Consider the derivative of the goodness-of-fit statistic first and remember that $SD_\lambda = \sum_{j,k} a_\lambda$. We show that for given $\delta \in (0,1)$ there exists a constant M_δ such that for almost all n holds

$$P\Bigg(\Big\|(\mu^n_{j+})^{-1/2} D_\theta \sum_{k=1}^{K} \Big(a_\lambda(X^n_{jk}, \hat{\mu}^n_{j+} \cdot \pi^n_{jk}(\theta^n_0)) - a_\lambda(X^n_{jk}, \mu^n_{j+} \cdot \pi^n_{jk}(\theta^n_0))\Big)\Big\| > M_\delta\Bigg) \leq \delta \tag{11}$$

for all $j \in \{1, \ldots, J^n\}$. Then the difference is stochastically bounded giving

$$D_\theta SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) - D_\theta SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) = \sum_{j=1}^{J^n} O_p((\mu^n_{j+})^{1/2}).$$

To prove (11) let any $j \in \{1, \ldots, J^n\}$ be given and consider $\hat{\mu}^n_{j+} > 0$. Using the positive homogeneity of the distance function a_λ, i.e. $c\, a_\lambda(x, \mu) = a_\lambda(cx, c\mu)$, we obtain

$$\begin{aligned}
&\Big\|(\mu^n_{j+})^{-1/2} D_\theta \sum_{k=1}^{K} \Big(a_\lambda(X^n_{jk}, \hat{\mu}^n_{j+} \cdot \pi^n_{jk}(\theta^n_0)) - a_\lambda(X^n_{jk}, \mu^n_{j+} \cdot \pi^n_{jk}(\theta^n_0))\Big)\Big\| \\
&\leq (\mu^n_{j+})^{1/2} \sum_{k=1}^{K} \Big\|D_\theta a_\lambda\Big(\frac{X^n_{jk}}{\mu^n_{j+}}, \frac{\hat{\mu}^n_{j+}}{\mu^n_{j+}} \cdot \pi^n_{jk}(\theta^n_0)\Big) - D_\theta a_\lambda\Big(\frac{X^n_{jk}}{\mu^n_{j+}}, \pi^n_{jk}(\theta^n_0)\Big)\Big\|.
\end{aligned}$$

Taylor expansion in $\hat{\mu}^n_{j+}/\mu^n_{j+}$ around its stochastic limit, 1, and application of the bounding conditions concerning the derived probabilities (RC3) shows that the term above is bounded. Since zero row sums $\hat{\mu}^n_{j+}$ appear with probability 0 by (LC0), we have (11). In order to verify

$$D_\theta m^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) = D_\theta m^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) + O_p(\sum_{j=1}^{J^n} (\mu^n_{j+})^{1/2}) \tag{12}$$

we show, componentwise, that for every j, k holds

$$(\mu^n_{j+})^{-1/2}\Big(\frac{\partial}{\partial \theta_s} e_\lambda(\hat{\mu}^n_{j+}, \pi^n_{jk}(\theta^n_0)) - \frac{\partial}{\partial \theta_s} e_\lambda(\mu^n_{j+}, \pi^n_{jk}(\theta^n_0))\Big) = O_p(1).$$

Since we need to differentiate with respect to the expected value and zero row sums can be neglected, we prove, instead,

$$\mathbf{1_N}(\hat{\mu}^n_{j+})\Big(\frac{\partial}{\partial \theta_s} e_\lambda(\hat{\mu}^n_{j+}, \pi^n_{jk}(\theta^n_0)) - \frac{\partial}{\partial \theta_s} e_\lambda(\mu^n_{j+}, \pi^n_{jk}(\theta^n_0))\Big) = O_p((\mu^n_{j+})^{1/2}). \tag{13}$$

Here $\mathbf{1_N}$ is an indicator function, i.e., $\mathbf{1_N}(x) = 1$ for positive integers x, e_λ denotes again the expectation of a_λ. For the proof we use the following Lemma which can be verified using standard properties of probability measures (see **11**, Lemma 7.3, for a proof).

Lemma 1 *Let $(X^n)_{n\in\mathbf{N}}$ be a sequence of nonnegative integer valued random variables and $(u^n)_{n\in\mathbf{N}}$ an arbitrary vector valued sequence ($u^n \in \mathbf{R}^m, m \in \mathbf{N}$). Assume $E(X^n) = \mu^n \in [\epsilon, \infty)$, $0 < \epsilon \leq 1$, and $Var(X^n) \leq \mu^n$ for all $n \in \mathbf{N}$. Consider a function $g : [0,\infty) \times \mathbf{R^m} \to \mathbf{R}$, $(X, u) \mapsto g(X, u)$, which is continuously differentiable in the first component on $(0,\infty)$. Let r be some nonnegative integer. If*

$$\mu^r \cdot \Big|\frac{\partial}{\partial\mu} g(\mu, u^n)\Big| \leq c \quad \text{for each } n \in \mathbf{N} \text{ and } \mu \geq \epsilon \quad (c > 0)$$

then for the asymptotics $n \to \infty$ holds

$$\mathbf{1_N}(X^n) \cdot \Big(g(X^n, u^n) - g(\mu^n, u^n)\Big) = O_p(\frac{1}{\mu^{r-1/2}}).$$

Since $(\hat{\mu}^n_{j+})_n$ is a sequence of Poisson variables with expected value $\mu^n_{j+} \in [\epsilon, \infty)$, we can apply Lemma 1 with $r = 0$ to establish (13). The function $\partial e_\lambda/\partial\theta_s$ corresponds with function g from Lemma 1. We only must show that for every j, k, n holds

$$|\frac{\partial}{\partial\mu}\frac{\partial}{\partial\theta_s} e_\lambda(\mu, \pi^n_{jk}(\theta^n_0))| \quad \leq \quad const \quad \text{for all } \mu \geq \epsilon.$$

By definition of e_λ, this holds if

$$|\frac{\partial}{\partial\mu^n_{j+}}\frac{\partial}{\partial\theta_s} e_\lambda(\mu^n_{j+}, \pi^n_{jk}(\theta^n_0))| \quad \leq \quad const \quad \text{for all } \mu^n_{j+} \geq \epsilon.$$

The proof can be done by differentiating $e_\lambda(\mu^n_{j+}, \pi^n_{jk}(\theta^n_0)) = E(a_\lambda(X^n_{jk}, \mu^n_{j+}\pi^n_{jk}))$ and applying further auxiliary results about the order of Poisson expectations, such as (8), combined with all bounding conditions concerning the probabilities and row expectations (RC2, RC3, MD0). This gives (9) and thus (10). Inserting in (5) yields

$$Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) = Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) + D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \cdot (\hat{\theta}^n - \theta^n_0) + \sum_{j=1}^{J^n} O_p\Big((\frac{\mu^n_{j+}}{\mu^n_{++}})^{1/2}\Big).$$

Now we approximate the gradient by its expected value,

$$D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \cdot (\hat{\theta}^n - \theta^n_0) \quad = \quad E(D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)) \cdot (\hat{\theta}^n - \theta^n_0) + O_p(1).$$

This follows from (LC2), $n^{1/2}(\hat{\theta}^n - \theta^n_0) = O_p(1)$, and

$$\begin{aligned} n^{-1/2}(D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) - E(D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0))) &= n^{-1/2}(D_\theta SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) - E(D_\theta SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0))) \\ &= O_p(1), \end{aligned}$$

which is shown by simple calculations and applying Chebyshev's inequality. We used, in particular, that the variance of $D_\theta SD^n_\lambda$ has the order n which is another auxiliary result about Poisson expectations (cf. (8)) and needs the bounding conditions (RC2),(RC3) and (MD0). Further calculations show that the covariance $c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)$ between SD^n_λ and the

score vector U^n (see Sec. 4) equals the negative expectation $-E(D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0))$. Hence we have

$$D_\theta Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)(\hat{\theta}^n - \theta^n_0) \quad = \quad -c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)(\hat{\theta}^n - \theta^n_0) + O_p(1). \tag{14}$$

We now replace the estimator in (14) through information matrix and score vector. This is possible by assumption (LC3), i.e., $(\hat{\theta}^n - \theta^n_0) = I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} U^n(\theta^n_0) + O_p(1/n)$. We obtain

$$c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)(\hat{\theta}^n - \theta^n_0) = c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} U^n(\theta^n_0) + O_p(1). \tag{15}$$

The error term in the formula is stochastically bounded since the order of c^n_λ computes to $c^n_\lambda = O(J^n)$, using the same bounding conditions as above. Already the meaning of J^n as the number of filled groups asserts $J^n/n = O(1)$. Using (14) and (15) we can rewrite (9):

$$\begin{aligned} Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) \quad &= \quad Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) - c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1} U^n(\theta^n_0) \\ &\quad + O_p(1) + \sum_{j=1}^{J^n} O_p\Big((\frac{\mu^n_{j+}}{\mu^n_{++}})^{1/2} \Big). \end{aligned} \tag{16}$$

Finally, we want to exchange the first term of the right-hand side of (16), $Z^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0)$, by $Z^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)$. This step is crucial: Since we replace the estimated marginal distribution of the covariate groups by their true distribution, we have to regard the bias that arises. The idea is as follows. We treat SD_λ and its expectation m_λ separately. For m_λ, which is simpler to handle, we obtain

$$m^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) = m^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) + O_p\Big(\sum_{j=1}^{J^n} (\mu^n_{j+})^{-1/2} \Big). \tag{17}$$

This follows by analogous arguments as in the proof of (12), i.e., we apply Lemma 1 and use the same bounding conditions. Let us now consider the transition from $SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0)$ to $SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)$. Using the positive homogeneity of a_λ and writing briefly $\mu^n_{jk} = \mu^n_{j+}\pi^n_{jk}(\theta^n_0)$, we have

$$\begin{aligned} SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) - SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \quad &= \quad \sum_{j=1}^{J^n} \sum_{k=1}^{K} \Big(a_\lambda(X^n_{jk}, \hat{\mu}^n_{j+}\pi^n_{jk}(\theta^n_0)) - a_\lambda(X^n_{jk}, \mu^n_{jk}) \Big) \\ &= \quad \sum_{j=1}^{J^n} \sum_{k=1}^{K} \Big(a_\lambda(X^n_{jk}, \hat{\mu}^n_{j+} \cdot \frac{\mu^n_{jk}}{\mu^n_{j+}}) - a_\lambda(X^n_{jk}, \mu^n_{jk}) \Big) \\ &= \quad \sum_{j=1}^{J^n} \sum_{k=1}^{K} \mu^n_{jk} \Big(a_\lambda(\frac{X^n_{jk}}{\mu^n_{jk}}, \frac{\hat{\mu}^n_{j+}}{\mu^n_{j+}}) - a_\lambda(\frac{X^n_{jk}}{\mu^n_{jk}}, 1) \Big). \end{aligned}$$

The key idea now is a second order Taylor expansion of the difference of the two a_λ's in both components around $(1, 1)$. Since a_λ has the appealing property that its first partial derivatives and a_λ itself vanish in $(1, 1)$, we obtain

$$SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \theta^n_0) - SD^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \quad = \quad -\sum_{j=1}^{J^n} \frac{(\hat{\mu}^n_{j+} - \mu^n_{j+})^2}{\mu^n_{j+}} + O_p\Big(\sum_{j=1}^{J^n} (\mu^n_{j+})^{-1/2} \Big). \tag{18}$$

The proof of (18), particularly the computation of the order of the approximation error, requires some work: The distance function a_λ is differentiable only on $(0,\infty)\times(0,\infty)$. Hence we have to consider the case $X_{jk}=0$ separately. For reasons of brevity, we refer to **11**, Lemma 5.8. The proof particularly requires assumption (RC2) and (MD0), i.e., the cell expectations must be bounded away from zero. Let us now consider the leading term of the expansion, Pearson's χ^2 Statistic for the row sums,

$$\sum_{j=1}^{J^n}\frac{(\hat{\mu}^n_{j+}-\mu^n_{j+})^2}{\mu^n_{j+}}=\sum_{j=1}^{J^n}\frac{(X^n_{j+}-\mu^n_{j+})^2}{\mu^n_{j+}}=\sum_{j=1}^{J^n}a_1(X^n_{j+},\mu^n_{j+}),$$

more carefully. This statistic obviously has the same order as $SD^n_\lambda(\mu^n_{\cdot+},\theta^n_0)$, for instance $SD^n_1(\mu^n_{\cdot+},\theta^n_0)=\sum_{j=1}^{J^n}\sum_{k=1}^{K}a_1(X^n_{jk},\mu^n_{jk})=\sum_{j=1}^{J^n}\sum_{k=1}^{K}(X^n_{jk}-\mu^n_{jk})^2/\mu^n_{jk}$. Hence it represents the bias that arises since we estimate the marginal covariate distribution. It cannot be ignored but must be part of an approximation of the test statistic.

In conclusion, Eq. (17) and (18) give

$$Z^n_\lambda(\hat{\mu}^n_{\cdot+},\theta^n_0)=Z^n_\lambda(\mu^n_{\cdot+},\theta^n_0)-\sum_{j=1}^{J^n}a_1(X^n_{j+},\mu^n_{j+})+\sum_{j=1}^{J^n}O_p\Big((\mu^n_{j+})^{-1/2}\Big). \tag{19}$$

Inserting in (16) and writing explicitly $SD^n_\lambda-m^n_\lambda$ instead of Z^n_λ, we have the desired sum of independent variables which does not depend on estimators anymore:

$$\begin{aligned}
&SD^n_\lambda(\hat{\mu}^n_{\cdot+},\hat{\theta}^n)-m^n_\lambda(\hat{\mu}^n_{\cdot+},\hat{\theta}^n)\\
&=\ SD^n_\lambda(\mu^n_{\cdot+},\theta^n_0)-m^n_\lambda(\mu^n_{\cdot+},\theta^n_0)-\sum_{j=1}^{J^n}a_1(X^n_{j+},\mu^n_{j+})-c^n_\lambda(\mu^n_{\cdot+},\theta^n_0)I^n(\mu^n_{\cdot+},\theta^n_0)^{-1}U^n(\theta^n_0)\\
&\quad+O_p(1)+\sum_{j=1}^{J^n}O_p(1)\Big((\frac{\mu^n_{j+}}{\mu^n_{++}})^{1/2}+(\frac{1}{\mu^n_{j+}})^{1/2}\Big).
\end{aligned} \tag{20}$$

6.2 Limiting normality of the approximated statistic

We derive the limiting normal distribution of the approximated statistic from (20). Let the stochastic terms be denoted by $\Psi^n_{\lambda+}$, i.e.,

$$\begin{aligned}
\Psi^n_{\lambda+} &= SD^n_\lambda(\mu^n_{\cdot+},\theta^n_0)-\sum_{j=1}^{J^n}a_1(X^n_{j+},\mu^n_{j+})-c^n_\lambda(\mu^n_{\cdot+},\theta^n_0)(I^n(\mu^n_{\cdot+},\theta^n_0))^{-1}U^n(\theta^n_0)\\
&= \sum_{j=1}^{J^n}\Psi^n_{\lambda j}
\end{aligned}$$

with

$$\Psi^n_{\lambda j} = \sum_{k=1}^{K}a_\lambda(X^n_{jk},\mu^n_{j+}\pi^n_{jk}(\theta^n_0))-a_1(X^n_{j+},\mu^n_{j+})-c^n_\lambda(\mu^n_{\cdot+},\theta^n_0)(I^n(\mu^n_{\cdot+},\theta^n_0))^{-1}U^n_j(\theta^n_0).$$

Since $E(SD^n_\lambda)=m^n_\lambda$, $E(\sum_{j=1}^{J^n}a_1(X^n_{j+},\mu^n_{j+}))=J^n$ and $E(U^n)=0$, the expected value of $\Psi^n_{\lambda+}$ is $m^n_\lambda-J^n$. Note that we called this expectation m^{*n}_λ in Theorem 1, i.e., $m^{*n}_\lambda=$

$E(\Psi^n_{\lambda+}) = m^n_\lambda - J^n$. Hence, using this notation and adding J^n to both sides of (20), we have

$$\begin{aligned} & SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) - m^{*n}_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) \\ = \; & SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) - m^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) + J^n \\ = \; & \Psi^n_{\lambda+} - E(\Psi^n_{\lambda+}) + O_p(1) + \sum_{j=1}^{J^n} O_p(1)\Big((\frac{\mu^n_{j+}}{\mu^n_{++}})^{1/2} + (\frac{1}{\mu^n_{j+}})^{1/2}\Big). \end{aligned} \tag{21}$$

Let σ^{n2}_λ denote the variance of the approximation, i.e., $\sigma^n_\lambda = \sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) = (Var(\Psi^n_{\lambda+}))^{1/2}$. The explicit formula can be derived by straightforward calculus and is given in Theorem 1. The limiting normal distribution of $\Psi^n_{\lambda+}$,

$$\frac{\Psi^n_{\lambda+} - E(\Psi^n_{\lambda+})}{\sigma^n_\lambda} \quad \xrightarrow{\mathcal{L}} \quad N(0,1), \tag{22}$$

follows from the central limit theorem if Ljapounov's condition is satisfied, i.e.,

$$\sum_{j=1}^{J^n} E(|\Psi^n_{\lambda j} - E(\Psi^n_{\lambda j})|^{2+\delta})/(\sigma^n_\lambda)^{2+\delta} \quad \longrightarrow \quad 0 \quad (\delta > 0).$$

We verify the condition for $\delta = 2$ (cf. **11**, Sec. 6). Since we assume (VC), $J^n/\sigma^{n2}_\lambda = O(1)$, we only must show

$$\sum_{j=1}^{J^n} E\Big(\Big(\Psi^n_{\lambda j} - E(\Psi^n_{\lambda j})\Big)^4\Big) \quad = \quad o((J^n)^2).$$

Using obvious inequalities, this holds if

$$\sum_{j=1}^{J^n} E\Big(\Big(\sum_{k=1}^{K} a_\lambda(X^n_{jk}, \mu^n_{jk}) - a_1(X^n_{j+}, \mu^n_{j+}) - \Big(\sum_{k=1}^{K} E(a_\lambda(X^n_{jk}, \mu^n_{jk})) - 1\Big)\Big)^4\Big) = o((J^n)^2),$$

$$\|c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \cdot (I^n(\mu^n_{\cdot+}, \theta^n_0))^{-1}\|^4 \sum_{j=1}^{J^n} E\Big(\|U^n_j(\theta^n_0) - E(U^n_j(\theta^n_0))\|^4\Big) = o((J^n)^2).$$

The first statement follows from the fact that the fourth moments of a_λ are bounded, which is another auxiliary result about Poisson expectations, such as (8), and requires that all cell expectations are bounded away from zero (RC2,MD0) (see **11**, Sec. 4).

Let us consider the second equation. Condition (LC1) immediately yields $I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1}$ $= O(n^{-1})$. In the last section we already used $c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) = O(J^n)$ for the proof of (15). This, together with $\mu^n_{++}/n = O(1)$, gives

$$\|c^n_\lambda(\mu^n_{\cdot+}, \theta^n_0) \cdot I^n(\mu^n_{\cdot+}, \theta^n_0)^{-1}\|^4 = O\Big((\frac{J^n}{n})^4\Big) = O\Big((\frac{J^n}{\mu^n_{++}})^4\Big).$$

Condition (RC2), (RC3) and $E(X^n_{jk} - \mu^n_{jk})^4 = O((\mu^n_{jk})^2)$ yield

$$E\Big(\|U^n_j(\theta^n_0) - E(U^n_j(\theta^n_0))\|^4\Big) = E\Big(\|\sum_{k=1}^{K}(X^n_{jk} - \mu^n_{jk}) D^T_\theta \log \pi^n_{jk}(\theta^n_0)\|^4\Big) = O\Big(\sum_{k=1}^{K}(\mu^n_{jk})^2\Big).$$

Hence we have the desired result,

$$\begin{aligned}
&\|c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) \cdot (I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1}\|^4 \sum_{j=1}^{J^n} E\Big(\|U_j^n(\theta_0^n) - E(U_j^n(\theta_0^n))\|^4\Big) \\
&= (\frac{J^n}{\mu_{++}^n})^4 \sum_{j=1}^{J^n}\sum_{k=1}^{K} (\mu_{jk}^n)^2 \cdot O(1) \\
&= O(J^n)^2 \cdot \frac{(J^n)^2 \sum_{j=1}^{J^n}\sum_{k=1}^{K}(\mu_{jk}^n)^2}{(\mu_{++}^n)^4} \\
&= O((J^n)^2) \cdot o(1).
\end{aligned}$$

The last equality holds due to the conditions concerning the marginal distribution (MD1, MD2) which require $J^n/\mu_{++}^n \to 0$ (see end of Sec. 5) and since $\sum_{j=1}^{J^n}\sum_{k=1}^{K}(\mu_{jk}^n)^2 < (\mu_{++}^n)^2$. Hence Ljapounov's condition is verified.

6.3 Consistency of the variance estimation

We finally need the consistency of $\sigma_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)$ as an estimator for $\sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n)$, i.e.,

$$\frac{\sigma_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - \sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n)}{\sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n)} = o_p(1). \tag{23}$$

By assumption (VC), the variance σ_λ^{n2} increases with order J^n. Hence, it suffices to show

$$\sigma_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - \sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) = o_p(J^n). \tag{24}$$

Write v_λ^{n2} for the variance of the difference between SD_λ^n and Pearsons's χ^2 Statistic for the row sums,

$$v_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) = Var\Big(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n)\Big).$$

Then, by definition,

$$\sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) = v_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) - c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) I^n(\mu_{\cdot+}^n, \theta_0^n)^{-1} c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)^T.$$

The main part of the proof of (24) is to show

$$c_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) + o_p(J^n), \tag{25}$$

$$v_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = v_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) + o_p(J^n), \tag{26}$$

$$I^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) = I^n(\mu_{\cdot+}^n, \theta_0^n) + o_p(n). \tag{27}$$

The proofs of these statements are very technical and, to some extent, similar to those carried out in Sec. 6.1, thus considering the transitions from $\hat{\theta}^n$ to θ_0^n and from $\hat{\mu}_{\cdot+}^n$ to $\mu_{\cdot+}^n$ separately. We omit details and refer to **11**, Sec. 6.2.

The consistency of the variance estimator now follows from (25), (26), $c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) = O(J^n)$ (compare Sec. 6.1) and from

$$I^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)^{-1} = I^n(\mu_{\cdot+}^n, \theta_0^n)^{-1} + o_p(n^{-1}).$$

The last statement holds by Eq. (27), combined with assumption (LC1), $n^{-1}I^n(\mu_{\cdot+}^n, \theta_0^n) \longrightarrow I_\infty$ positive definite. This gives (24),

$$\begin{aligned}
\sigma_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - \sigma_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) &= v_\lambda^{n2}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - v_\lambda^{n2}(\mu_{\cdot+}^n, \theta_0^n) \\
&\quad -c_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) I^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)^{-1} c_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n)^T \\
&\quad +c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) I^n(\mu_{\cdot+}^n, \theta_0^n)^{-1} c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)^T \\
&= o_p(J^n) + o_p\Big(\frac{(J^n)^2}{n}\Big) \\
&= o_p(J^n).
\end{aligned}$$

We further used that J^n is smaller than n.

6.4 Summary

Our limit theorem, Theorem 1, now follows from the results given in Sec. 6.1 through 6.3. There we used all conditions given in Sec. 5, except of both conditions concerning the marginal distribution,

$$\text{(MD1)} \ \sum_{j=1}^{J^n} \Big(\frac{\mu_{j+}^n}{\mu_{++}^n}\Big)^{1/2} = o((J^n)^{1/2}) \quad \text{and} \quad \text{(MD2)} \ \sum_{j=1}^{J^n} (\mu_{j+}^n)^{-1/2} = o((J^n)^{1/2}).$$

These we did not yet need in full strength. We will, however, need them now for our final conclusions.

In Section 6.1 we derived an approximation of the goodness-of-fit statistic (see (21)),

$$\begin{aligned}
&SD_\lambda^n(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) - m_\lambda^{*n}(\hat{\mu}_{\cdot+}^n, \hat{\theta}^n) \\
&= \Psi_{\lambda+}^n - E(\Psi_{\lambda+}^n) + O_p(1) + \sum_{j=1}^{J^n} O_p(1)\Big((\frac{\mu_{j+}^n}{\mu_{++}^n})^{1/2} + (\frac{1}{\mu_{j+}^n})^{1/2}\Big)
\end{aligned}$$

with

$$\Psi_{\lambda+}^n = SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - \sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n) - c_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)(I^n(\mu_{\cdot+}^n, \theta_0^n))^{-1} U^n(\theta_0^n),$$

$$E(\Psi_{\lambda+}^n) = E(SD_\lambda^n(\mu_{\cdot+}^n, \theta_0^n)) - E\Big(\sum_{j=1}^{J^n} a_1(X_{j+}^n, \mu_{j+}^n)\Big) = m_\lambda^n(\mu_{\cdot+}^n, \theta_0^n) - J^n = m_\lambda^{*n}(\mu_{\cdot+}^n, \theta_0^n).$$

Note that the centering term involves not only the expectation of SD_λ^n, namely m_λ^n, but also the number of rows J^n which is the expectation of Pearson's χ^2 Statistic for the row sums.

The limiting normal distribution of the standardized approximation was shown in Section 6.2, Eq. (22),

$$\frac{\Psi^n_{\lambda+} - E(\Psi^n_{\lambda+})}{\sigma^n_\lambda} \quad \xrightarrow{\mathcal{L}} \quad N(0,1)$$

where σ^n_λ denotes the standard error of $\Psi^n_{\lambda_+}$, $\sigma^n_\lambda = (Var(\Psi^n_{\lambda+}))^{1/2}$. For the test statistic we need an estimator of σ^n_λ and take $\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)$. Summing up, we have

$$\begin{aligned}
&\frac{SD^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n) - m^{*n}_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} \\
&= \frac{\Psi^n_{\lambda+} - E(\Psi^n_{\lambda+})}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} + \frac{O_p(\sum_{j=1}^{J^n}(\mu^n_{j+}/\mu^n_{++})^{1/2}) + O_p(\sum_{j=1}^{J^n}(\mu^n_{j+})^{-1/2}) + O_p(1)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} \qquad (28)
\end{aligned}$$

The asymptotic normality of $\Psi^n_{\lambda_+}$ stated above, combined with the consistency of the variance estimation,

$$\frac{\sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} \quad \xrightarrow{P} \quad 1,$$

(Sec. 6.3, Eq. (23)) gives or the leading term in (28)

$$\frac{\Psi^n_{\lambda+} - E(\Psi^n_{\lambda+})}{\sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)} \cdot \frac{\sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)} \quad \xrightarrow{\mathcal{L}} \quad N(0,1).$$

Hence, in order to establish Theorem 1, i.e., the asymptotic normality of the statistic in the left-hand side of (28), it remains to show that the approximation error in the right-hand side disappears in the limit. Rewriting gives

$$\frac{O_p(\sum_{j=1}^{J^n}(\mu^n_{j+}/\mu^n_{++})^{1/2}) + O_p(\sum_{j=1}^{J^n}(\mu^n_{j+})^{-1/2}) + O_p(1)}{(J^n)^{1/2}} \cdot \frac{(J^n)^{1/2}}{\sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)} \cdot \frac{\sigma^n_\lambda(\mu^n_{\cdot+}, \theta^n_0)}{\sigma^n_\lambda(\hat{\mu}^n_{\cdot+}, \hat{\theta}^n)}.$$

The two last terms are clearly (stochastically) bounded due to our variance condition (VC), i.e., $J^n/\sigma^{n2}_\lambda(\mu^n_{\cdot+}, \theta^n_0) = O(1)$, and the consistency of the variance estimation. The first term is crucial: it involves the approximation errors due to the estimation of the unknown covariate distribution (compare (9) and (19)) and must tend to zero so that our stated normality holds. We formulated this requirement separately for the two sums of the ratio and called the resulting two conditions (MD1) and (MD2). This completes the proof.

7 Application and discussion

We illustrate our method by a real application with published data. A suitable data set is provided by Karn and Penrose **13**, who reported on a study on infant mortality. They are from records of U.C.H. Obstetric Hospital for the years 1935 – 46 and contain information on 13 730 infants (7037 male, 6693 female, no twins) and their mothers. We are interested here only in parts of the data (Table 1 from Karn and Penrose) which relate non-survival at 28 days (including stillbirth), regarded as a response, to the following variables:

- birth weight W, recorded in 25 classes: 1.0 (0.5) 13.5 lb.,
- gestation time T, recorded in 41 classes: 155 (5) 355 days,
- gender G of infant, recorded as a factor: 1 = male, 2 = female.

Karn and Penrose fitted a linear logistic model (to the survival rate) separately for males and females, using the model

$$1 + W + W^2 + T + T^2 + W.T$$

with $S = 6$ parameters (for the symbolic notation see McCullagh and Nelder **15**, Sec. 3.4). We investigate the fit of this model only for the *female* infants in more detail. These data have also been analyzed by Osius **14** for binomial sampling.

Focusing on the values $\lambda = 1$ and $\lambda = 0$, which give the traditional Pearson and Likelihood Ratio Statistic (deviance), we also examine the intermediate value $\lambda = 2/3$ suggested by Cressie and Read **2**. The three statistics SD_λ are given in Table 2a. They obviously differ considerably and the corresponding "classical" p–levels based on an asymptotic χ^2 distribution with 339 degrees of freedom vary from 68.81% (deviance) to 0.99% (Pearson). Since the data are sparse, the χ^2 distribution may not be reliable. Therefore we computed the asymptotic expectation and variance of SD_λ and the corresponding standardized statistic T_λ (cf. Table 2a). The p–levels based on the asymptotic normal distribution of T_λ are 2.66% ($\lambda = 0$), 1.59% ($\lambda = 2/3$) and 30.65% ($\lambda = 1$). Although for each statistic the p–levels based on the different asymptotic distribution differ dramatically, it is not obvious which one is more reliable. This provoked us to investigate the assumptions (MD1) and (MD2). We computed the sample means $M_1 = 19.40$, $M_{1/2} = 9.7368$ and $M_{-1/2} = 3.1316$. This gives $M_1/M_{1/2} = 1.99$ which is not really large but not too small either. The ratio $M_{-1/2}/J$, however, computes to 0.0091 which is far from being large, in contrast to (MD2). To find out whether the normal p–levels are reliable in this situation, we computed a p–level via parametric bootstrap. On the basis of the estimated expected values $\hat{\mu}_{jk}$ we generated 10000 Poisson resamples of the study. The results are given in Table 2b. Although the bootstrapped p–levels for T_λ are much closer to those based on the normal than those based on the χ^2 approximation, the difference is still too large to be satisfactory. The reason becomes evident by considering the simulated moments of T_λ in Table 2b. The expectation (which should be 0) is not sufficiently small (except for $\lambda = 2/3$) and the variance is much smaller than 1. This leads to smaller bootstrapped p–levels in comparison to the normal approximation. Looking at the third and fourth standardized cumulants of T_λ, only the deviance has fairly small bootstrapped values (which should be close to zero for a normal distribution). Pearson's statistic, however, shows considerable skewness and kurtosis. This and the large variance is explained by the fact that very small expected cell counts (in the denominator) make a large contribution to the statistic if the observed value is 1. We obtained similar results for the uncommon values

Table 2: Goodness–of–fit results for Karn and Penrose data (females)

a) Statistics with asymptotic moments and p–levels

Distance (λ)	sum SD_λ	d.f.	χ^2 p–level	expected value m^*_λ	variance σ^2_λ	statistic T_λ	normal p–level
Deviance (0)	325.74	339	68.81 %	284.11	463.98	1.93	2.66 %
Cressie-Read (2/3)	344.64	339	40.47 %	276.12	1018.08	2.15	1.59 %
Pearson (1)	402.59	339	0.99 %	345.00	12963.90	0.51	30.65 %

b) Results of parametric bootstrap (10000 resamples): p–levels, moments and standardized cumulants of T_λ

Distance (λ)	p–level of T_λ	expectation	variance	stand. 3rd cumulant	stand. 4th cumulant
Deviance (0)	0.66 %	-0.34	0.77	0.19	0.04
Cressie-Read (2/3)	2.05 %	0.09	0.87	1.24	10.01
Pearson (1)	9.85 %	-0.13	0.56	9.60	255.73

c) Results of parametric bootstrap (10000 resamples): p–levels and moments of SD_λ

Distance (λ)	p–level of SD_λ	expectation	variance
Deviance (0)	0.02 %	250.3	397.1
Cressie-Read (2/3)	0.33 %	250.8	756.3
Pearson (1)	2.02 %	291.4	5143.3

$\lambda = 1.5, 2$ and 2.5, and this effect increased with λ. Summing up, the bootstrapped p–levels of the deviance ($p = 0.66\%$) and the Cressie–Read Statistic ($p = 2.05\%$) do not indicate a very satisfactory fit (which still may be acceptable taking the very large sample size into account). Pearson's statistic ($p = 9.85\%$), and other statistics with $\lambda > 1$, in view of their large variance, do not seem to be powerful enough to detect this.

We note that it is important to bootstrap the statistic T_λ (which is pivotal by Theorem 1) rather than the sum SD_λ. Bootstrapping SD_λ would, in fact, give much smaller but less reliable p–levels (cf. Table 2c). One can, however, never be sure whether T_λ is pivotal in the situation of a particular (sparse) data set. Especially in the example of the Karn & Penrose data some caution is needed, as indicated by our above results. The alternative to bootstrapping T_λ would be to bootstrap SD_λ. The sum SD_λ, however, may be regarded pivotal only on the basis of its asymptotic χ^2 distribution. For the Karn & Penrose data this is even harder to believe. In fact, our bootstrapped moments of SD_λ given in Table 2c show

a considerable deviation from the χ^2 expectation (d.f. = 339) and variance (2× d.f. = 678) for all three statistics. Hence the standardized version T_λ of SD_λ seems to be more reliable for bootstrapping, although its bootstrapped expectations respectively variances given in Table 2b are not close to 0 respectively 1 for all values of λ.

In conclusion, the above data appear to be too sparse (27% of all row sums are 1) even for our normal approximation. The value of $M_{-1/2}$ in comparison to J is small but should be large according to (MD2). However, bootstrapping T_λ (instead of SD_λ) provides more adequate p–levels and indicates that our asymptotic moments need further corrections when data are very sparse.

8 Conclusions

In conclusion, it may be said that through the consideration of an "increasing cells" approach a limiting normal distribution of SD_λ for the important case of Poisson sampling could be derived. Particularly in view of the fact that χ^2 tests are often used in situations not covered by standard asymptotic results, this result helps to gain insight about how the distribution gets shifted when our asymptotic approach allowing small expectations applies. In particular, we provided a basis for investigations of *conditional* Poisson models. In this context case-control studies, i.e., column-multinomial sampling, are probably most interesting: Since the distribution of the covariates is not given there and only associations shall be modeled, one has to deal with an asymptotically infinite number of nuisance parameters, too. Besides extensions to other distribution models, it is certainly desirable to derive more accurate tests based on higher order approximations such as Edgeworth and saddle-point approximations. Instead of relying on asymptotic distributional results, a goodness-of-fit test may also be derived from its bootstrapped distribution. Since "in several respects the bootstrap does a better job of estimating the distribution of a pivotal statistic than it does for a non pivotal statistic" (Hall **16**, Sec. 3.1) one should bootstrap the statistic T_λ (which is asymptotically pivotal by Theorem 1) rather than the sum SD_λ. Hence the asymptotic normality of T_λ turns out to be important even if a test is based on the bootstrap. Such an approach was, for example, carried out by Osius **14** for row-multinomial tables and would be advisable in order to improve the tests derived.

References

1. N.A.C. Cressie and T.R.C. Read (1984) Multinomial goodness-of-fit tests, *J. Roy. Statist. Soc. Ser. B*, **46**, 3, 440 – 464.

2. N.A.C. Cressie and T.R.C. Read (1988) *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.

3. J.B.S. Haldane (1940) The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small, *Biometrika*, 31, 346 – 355.

4. C. Morris (1975) Central limit theorems for multinomial sums, *Ann. Statist.*, **3**, 1, 165 – 188.

5. P. McCullagh (1986) The conditional distribution of goodness-of-fit statistics for discrete data, *J. Americ. Statist. Assoc.*, **81**, 393, 104 – 107.

6. G. Osius (1985) Goodness-of-fit tests for binary data with (possible) small expectations but large degrees of freedom, *Statist. Decisions*, **2**, 213 – 224.

7. D. Rojek (1989) *Asymptotik für Anpassungstests in Produkt-Multinomialmodellen bei wachsendem Freiheitsgrad*, Ph.D. thesis, University of Bremen, Germany.

8. G. Osius and D. Rojek (1992) Normal goodness-of-fit tests for multinomial models with large degrees of freedom, *J. Americ. Statist. Assoc.*, **87**, 420, 1145 – 1152.

9. K.J. Koehler (1986) Goodness-of-fit tests for log-linear models in sparse contingency tables, *J. Americ. Statist. Assoc.*, **81**, 483 – 489.

10. J.R. Dale (1986) Asymptotic normality of goodness-of-fit statistics for sparse multinomials, *J. Roy. Statist. Soc. Ser. B*, **48**, 1, 48 – 59.

11. U.U. Müller (1997) *Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Poisson and Case Control Data*, Ph.D. thesis, University of Bremen, Germany.
[www http://www.math.uni-bremen.de/∼uschi/]

12. P. Billingsley (1995) *Probability and Measure* (3rd Ed.), John Wiley & Sons, Chichester.

13. M.N. Karn and L.S. Penrose (1951 – 52) Birth weight and gestation time in relation to maternal age, parity and infant survival. *Annals of Eugenetics* (London), **16**, 147 – 164.

14. G. Osius (1994) Evaluating the significance level of goodness-of-fit statistics for large discrete data, *Computational Statistics* (ed. P. Dirschedl and R. Ostermann), 393 – 417, Physica-Verlag, Heidelberg.

15. P. McCullagh and J.A. Nelder (1989) *Generalized Linear Models* (2nd Ed.), Chapman and Hall, London.

16. P. Hall (1992) *The Bootstrap and Edgeworth Expansion*, Springer, New York.