

Evaluating The Significance Level of Goodness-of-Fit Statistics For Large Discrete Data

Gerhard Osius *

Universität Bremen, Institut für Statistik, Fachbereich 3
Postfach 330440, D-28334 Bremen, Germany

Key Words:

Binomial data, Bootstrap, Edgeworth expansion, Multinomial data, Power-divergence statistic, Saddlepoint approximation, Sparse data.

Abstract

Goodness-of-fit tests for multinomial models with parameters to be estimated are usually based on Pearson's X^2 or the deviance (likelihood ratio) D , both belonging to the family of power-divergence statistics SD_λ . Our aim is to evaluate the significance level of SD_λ for large samples and models with large degrees of freedom, but not necessarily large expected counts in each cell, allowing sparse data too. We consider approximations based on limiting distributions, Edgeworth and saddlepoint expansions, and the parametric bootstrap. Computational details are given, and two larger studies serve as numerical examples.

Introduction

Two devices are in common use to judge the fit of a chosen model for discrete data: the more informal analysis of suitably defined residuals and formal goodness-of-fit tests based on summary statistics. Our focus will be on goodness-of-fit statistics, both from a theoretical and computational point of view. The most common goodness-of-fit statistics are Pearson's X^2 and the deviance D (i.e. the likelihood ratio statistic). Cressie and Read (1984) have embedded these and other familiar statistics in the family of power-divergence statistics SD_λ with a real parameter $\lambda \in \mathbb{R}$, which are defined as follows. For each cell $i = 1, \dots, I$ the deviation of the observed count

*The author thanks the referees for helpful comments.

Y_i from its fitted value (expected count) \hat{Y}_i is measured using a “distance” function

$$a_\lambda(y, \hat{y}) = \frac{2y}{\lambda(\lambda + 1)} \left[\left(\frac{y}{\hat{y}} \right)^\lambda - 1 \right] - \frac{2}{\lambda + 1}(y - \hat{y}) ,$$

where the second term is introduced to make a_λ non-negative. The cases $\lambda = 0, -1$ are defined by continuity as $\lambda \rightarrow 0, -1$. The function a_λ is not symmetrical, but the arguments are reversed in passing from λ to $-(\lambda + 1)$

$$a_{-(\lambda+1)}(y, \hat{y}) = a_\lambda(\hat{y}, y) .$$

Since we consider here only models with positive fitted values $\hat{y} > 0$ but allow zero observed counts $y = 0$, which are typically encountered in sparse data, we restrict ourselves to values $\lambda > -1$.

The power-divergence statistic is the sum of all deviations

$$SD_\lambda = \sum_i a_\lambda(Y_i, \hat{Y}_i) .$$

In almost all applications, the observed total Y_+ equals the fitted total \hat{Y}_+ (the index “+” indicates summation over the index it replaces) and hence the sum of all second terms in a_λ vanishes, thus reducing SD_λ to the original definition of Cressie and Read (1984).

For $\lambda = 1$ we get Pearson’s statistic $X^2 = SD_1$, with

$$a_1(y, \hat{y}) = (y - \hat{y})^2 / \hat{y} ,$$

the deviance $D = SD_0$ is obtained for $\lambda = 0$ with

$$a_0(y, \hat{y}) = 2 [y \log(y/\hat{y}) - (y - \hat{y})] ,$$

and $\lambda = -1/2$ yields Freeman-Tukey’s statistic $FT = SD_{-1/2}$ with

$$a_{-1/2}(y, \hat{y}) = 4(y^{1/2} - \hat{y}^{1/2})^2 .$$

Read and Cressie (1988) suggested the value $\lambda = 2/3$ as a compromise between the rival values 0 and 1, and we denote their statistic by $CR = SD_{2/3}$. Despite the richness of the family SD_λ only the traditional values $\lambda = 0, 1$ are widely used.

The classical goodness-of-fit test is based on an asymptotic χ^2 -distribution for SD_λ , which is appropriate for an increasing sample size $n := Y_+$, provided the number I of cells remains *constant* as $n \rightarrow \infty$. This so-called *fixed-cells asymptotic* implies that all fitted counts increase, i.e. $\hat{Y}_i \rightarrow \infty$.

Hence the approximation of the significance level $P\{SD_\lambda \geq sd_\lambda\}$ for the observed value sd_λ based on the asymptotic χ^2 -distribution can be expected to be sufficiently accurate, only if all fitted counts \hat{Y}_i are reasonably “large”. According to a popular rule of thumb, going back to Cochran (1952), $\hat{Y}_i \geq 5$ seems large enough, but we will not pursue this topic.

Although very important, the fixed-cells asymptotic are not generally reliable in practice, in particular not for sparse data, where sample size n is large, but many (or even all) fitted values \hat{Y}_i remain small, because the number of cells I is large, too. For such a situation an *increasing-cells asymptotic* is appropriate which only requires $I \rightarrow \infty$ as $n \rightarrow \infty$. Several authors have investigated the asymptotic behaviour of D and X^2 under increasing cells, specific sampling distributions for the observed counts (e.g. multinomial, binomial or Poisson) and particular types of models (e.g. log-linear), cf. McCullagh (1985a,b, 1986), Dale (1986) and Koehler (1986). For general $\lambda > -1$, the asymptotic normality of SD_λ was derived in Osius (1985) and generalized in Rojek (1989). The increasing-cells approach thus provides an alternative approximation of the significance level based on the normal distribution of SD_λ , provided that both n and I are large.

Although the limiting distribution of SD_λ differs for the two asymptotics, it would be desirable for practical purposes, to have approximations for the tail probabilities $P\{SD_\lambda \geq sd_\lambda\}$ which apply for large n , no matter whether the fitted counts \hat{Y}_i are small, moderate or large. Our aim is to propose approximate significance levels based on different approaches

- limiting distributions for SD_λ (chi-squared and normal),
- Edgeworth and saddlepoint expansions,
- the parametric bootstrap.

The computational aspects for these methods will be given detailed enough to enable implementation in common statistical software packages (e.g. GLIM) or programming languages (e.g. S-Plus). For illustration purposes the methods are applied to two sets of sparse data, namely a study on infant mortality and on cancer.

The basic ideas are presented here in an expository and informal way, giving references to detailed work. We choose the convenient setup of multinomial models for contingency tables with no attempts to achieve a maximum of generality.

Multinomial Models

To fix the setup, we suppose that the observed counts form a $J \times K$ contingency table $\mathbf{Y} = (Y_{jk})$ with $I = JK$ cells. Let the columns represent a cat-

egorical response with K levels (e.g. a binary response for $K = 2$), and the rows are interpreted as groups j which are usually characterized by an additional S -dimensional vector $\mathbf{x}_j = (x_1, \dots, x_S) \in \mathbb{R}^S$ of observed covariables. Thus the random variable Y_{jk} is the number of items in group j with response k , with an assumed positive expectation $\mu_{jk} = E(Y_{jk}) > 0$. Our primary interest focuses on the (conditional) probability $\pi_{jk} = \mu_{jk}/\mu_{j+} \in (0, 1)$ for response k in group j , for which we consider models of a generalized linear type

$$\pi_{jk} = \pi_{jk}(\boldsymbol{\theta}) := G_k(\mathbf{x}_j^T \boldsymbol{\theta}). \quad (1)$$

Here $\boldsymbol{\theta} = (\theta_1, \dots, \theta_S)$ is an unknown parameter vector, and G_k are known smooth functions with values in $(0, 1)$.

An adequate *sampling model* for this situation is the *product-multinomial*, in which the vectors of row counts $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{jK})$ are independent for $j = 1, \dots, J$, each having a multinomial distribution $M_K(N_j, \boldsymbol{\pi}_j)$ of size $N_j = Y_{j+}$ with probability vector $\boldsymbol{\pi}_j = (\pi_{j1}, \dots, \pi_{jK})$. The product-multinomial model, assumed throughout, also arises from other popular sampling models for \mathbf{Y} (e.g. Poisson or single-multinomial) by conditioning on the observed row totals Y_{j+} , which contain no information about the probabilities $\boldsymbol{\pi}$ of interest (Haberman, 1974).

The log-likelihood function for the product-multinomial model is up to a constant given by

$$l(\boldsymbol{\theta}) = \sum_j \sum_k Y_{jk} \log \pi_{jk}(\boldsymbol{\theta}).$$

To fit the model, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ (or some asymptotically equivalent estimate) has to be determined – typically by an iterative procedure – to obtain the fitted probabilities $\hat{\pi}_{jk} = \pi_{jk}(\hat{\boldsymbol{\theta}})$ and fitted cell counts $\hat{Y}_{jk} = N_j \hat{\pi}_{jk}$, both being positive.

The power-divergence statistic may now be obtained by summing up first the deviations within each group and then over all groups

$$SD_\lambda(\hat{\boldsymbol{\theta}}) = \sum_j \sum_k a_\lambda(Y_{jk}, \hat{Y}_{jk}) = \sum_j A_\lambda(\mathbf{Y}_j, \hat{\mathbf{Y}}_j), \quad (2)$$

where A_λ serves as a “distance measure” between K -dimensional vectors

$$A_\lambda(\mathbf{y}, \hat{\mathbf{y}}) := \sum_k a_\lambda(y_k, \hat{y}_k). \quad (3)$$

In view of $a_\lambda(c\mathbf{y}, c\hat{\mathbf{y}}) = c a_\lambda(\mathbf{y}, \hat{\mathbf{y}})$ for any real $c > 0$, the statistic may also be written in terms of the observed and fitted frequencies $\hat{P}_{jk} := Y_{jk}/N_j$ and $\hat{\pi}_{jk}$ as

$$SD_\lambda(\hat{\boldsymbol{\theta}}) = \sum_j N_j A_\lambda(\hat{\mathbf{P}}_j, \hat{\boldsymbol{\pi}}_j).$$

Usually large values of the statistic $SD_\lambda(\hat{\boldsymbol{\theta}})$ indicate a lack of fit, at least if its expectation does not depend upon the parameter $\boldsymbol{\theta}$. Since the exact distribution of $SD_\lambda(\hat{\boldsymbol{\theta}})$ under the model is not tractable in practice (except in very simple situations), we have to rely on approximations for the significance level $P\{SD_\lambda(\hat{\boldsymbol{\theta}}) \geq sd_\lambda\}$ to perform a goodness-of-fit test.

Limiting Distributions

The most popular and simple approximation of the significance level is based on the limiting χ^2 -distribution of SD_λ (the argument $\hat{\boldsymbol{\theta}}$, being fixed here, is now omitted) with $df = J(K - 1) - S$ degrees of freedom (cf. Read and Cressie 1988)

$$P\{SD_\lambda \geq sd_\lambda\} \approx P\{\chi_{df}^2 \geq sd_\lambda\} , \quad (\text{CCA})$$

which will be referred to as the *classical χ^2 -approximation*. The accuracy of this approximation (derived for increasing-cells asymptotics) increases as *all* fitted counts \hat{y}_{jk} resp. all N_j tend to infinity, but may be extremely poor if a relevant fraction of fitted counts are small.

In contrast, the increasing-cells asymptotic requires the number J of groups to tend to infinity as $n \rightarrow \infty$, while the number K of responses categories and the number S of parameters remain *fixed*. No restrictions are imposed here on the fitted counts \hat{Y}_{jk} resp. the group sizes N_j , some or even all of which may be low. Only in the extreme case of individual groups (i.e. $N_j = 1$ for *all* j) the statistic SD_λ may have no diagnostic power for particular model-dependent values of λ (cf. McCullagh 1985a, Osius and Rojek 1992).

Under increasing-cells asymptotic, the power divergence statistic SD_λ has an asymptotic normal distribution (Rojek 1989) with expectation μ_λ and variance σ_λ^2 , given by (16) and (17) in the next section. This is not surprising, since SD_λ is an increasing sum of components $A_\lambda(\mathbf{Y}_j, \hat{\mathbf{Y}}_j)$, which are almost independent except for their dependence through the common estimate $\hat{\boldsymbol{\theta}}$ of fixed dimension. The asymptotic normality of the *standardized* power-divergence statistic

$$T_\lambda = (SD_\lambda - \mu_\lambda)/\sigma_\lambda \quad (4)$$

leads to the *normal approximation*

$$P\{T_\lambda \geq t_\lambda\} \approx P\{N(0, 1) \geq t_\lambda\} . \quad (\text{NA})$$

for the significance level of the observed value $t_\lambda = (sd_\lambda - \mu_\lambda)/\sigma_\lambda$. The accuracy of (NA) increases with J , its error being $O(J^{-1/2})$.

Before deriving the moments μ_λ and σ_λ^2 in general, we note a remarkable approximation, whose accuracy depends on the harmonic mean

$$HM = \left[(N_1^{-1} + \dots + N_J^{-1}) / J \right]^{-1}$$

of the group sizes. For large J we get (cf. Osius and Rojek 1992)

$$\sigma_\lambda^2 \approx 2J(K-1) = 2(df + S), \quad \text{if HM is large,} \quad (5)$$

$$\mu_\lambda \approx J(K-1) = df + S, \quad \text{if } HM/\sqrt{J} \text{ is large,} \quad (6)$$

In view of $df + S \approx df$ (for large df) these approximate moments agree with the moments of the classical χ^2 -limit, and lead to comparable results for (NA) and (CCA), provided HM/\sqrt{J} is large, too. The main disadvantage of the classical χ^2 -approximation for large J stems from the fact that the expectation of SD_λ can differ markedly from that of the limiting χ^2 -distribution (i.e. df) if HM/\sqrt{J} is not large enough, and this may result (see example 1 later) in completely misleading significance values based on (CCA).

For practical purposes, however, a smooth transition between the two approximations (CCA) and (NA) is available, which is due to one referee of the paper by Osius and Rojek (1992). The idea is to approximate the normal distribution $N(\mu, \sigma^2)$ by a scaled χ^2 -distribution $\beta \cdot \chi_\nu^2$ with the same expectation $\mu = \beta\nu$ and variance $\sigma^2 = 2\beta^2\nu$. Hence the *rescaled power-divergence statistic* SD_λ/β_λ with

$$\beta_\lambda = \sigma_\lambda^2 / 2\mu_\lambda \quad (7)$$

has an approximate χ^2 -distribution with a real-valued degree of freedom

$$\nu_\lambda = 2\mu_\lambda^2 / \sigma_\lambda^2, \quad (8)$$

provided ν_λ is large.

For fixed-cells asymptotic on the other hand, the approximations (5) and (6) yield $\beta_\lambda \approx 1$ and $\nu_\lambda - S \approx df$. Hence both approximations (CCA) and (NA) may be incorporated in the *rescaled χ^2 -approximation*

$$P\{T_\lambda \geq t_\lambda\} \approx P\{\chi_{df_\lambda}^2 \geq sd_\lambda / \beta_\lambda\} \quad (\text{RCA})$$

with $df_\lambda = \nu_\lambda$ or preferably $df_\lambda = \nu_\lambda - S$ degrees of freedom, provided both are large. The approximation (RCA) will give comparable results to (CCA) resp. (NA) for large J and large HM/\sqrt{J} .

Expectation and Variance of the Statistics

We now derive the asymptotic expectation μ_λ and variance σ_λ^2 of the power-divergence statistic $SD_\lambda(\boldsymbol{\theta})$ for large J , which are needed for the normal and rescaled χ^2 -approximation. The derivation is only sketched here to give the general idea, referring the interested reader to Rojek (1989) or Osius and Rojek (1992) for details. We first look at the sum $SD_\lambda(\boldsymbol{\theta})$ with the true parameter $\boldsymbol{\theta}$ instead of its estimate, and denote its expectation and variance (under the model) by

$$\mu_\lambda(\boldsymbol{\theta}) := E_\theta\{SD_\lambda(\boldsymbol{\theta})\}, \quad (9)$$

$$v_\lambda^2(\boldsymbol{\theta}) := \text{var}_\theta\{SD_\lambda(\boldsymbol{\theta})\}. \quad (10)$$

A first order expansion of the centered sum $Z_\lambda(\boldsymbol{\theta}) = SD_\lambda(\boldsymbol{\theta}) - \mu_\lambda(\boldsymbol{\theta})$ gives

$$Z_\lambda(\hat{\boldsymbol{\theta}}) = Z_\lambda(\boldsymbol{\theta}) + \mathbf{D}Z_\lambda(\boldsymbol{\theta}) \cdot (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + O_p(1),$$

where \mathbf{D} is the differential operator. Using the score vector $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{D}l(\boldsymbol{\theta})$ and the information matrix $\mathbf{I}(\boldsymbol{\theta}) := \text{cov}_\theta\{\mathbf{U}(\boldsymbol{\theta})\}$, we get the familiar expansion

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathbf{I}^{-1}(\boldsymbol{\theta}) \cdot \mathbf{U}(\boldsymbol{\theta}) + O_p(n^{-1}).$$

The derivative of $Z_\lambda(\boldsymbol{\theta})$ may be written as $\mathbf{D}Z_\lambda(\boldsymbol{\theta}) = -\mathbf{c}_\lambda(\boldsymbol{\theta}) + O_p(n^{1/2})$ with the covariance vector

$$\mathbf{c}_\lambda(\boldsymbol{\theta}) := \text{cov}_\theta\{SD_\lambda(\boldsymbol{\theta}), \mathbf{U}(\boldsymbol{\theta})\}. \quad (11)$$

This leads to the fundamental representation

$$Z_\lambda(\hat{\boldsymbol{\theta}}) = \psi_\lambda(\boldsymbol{\theta}) + O_p(1). \quad (12)$$

with

$$\psi_\lambda(\boldsymbol{\theta}) := Z_\lambda(\boldsymbol{\theta}) - \mathbf{c}_\lambda^T(\boldsymbol{\theta}) \cdot \mathbf{I}^{-1}(\boldsymbol{\theta}) \cdot \mathbf{U}(\boldsymbol{\theta}) \quad (13)$$

Simple calculations give

$$\begin{aligned} E_\theta\{\psi_\lambda(\boldsymbol{\theta})\} &= 0, \\ \sigma_\lambda^2(\boldsymbol{\theta}) &:= \text{var}_\theta\{\psi_\lambda(\boldsymbol{\theta})\} = v_\lambda^2(\boldsymbol{\theta}) - Q_\lambda(\boldsymbol{\theta}), \end{aligned} \quad (14)$$

with the quadratic form

$$Q_\lambda(\boldsymbol{\theta}) := \mathbf{c}_\lambda^T(\boldsymbol{\theta}) \cdot \mathbf{I}^{-1}(\boldsymbol{\theta}) \cdot \mathbf{c}_\lambda(\boldsymbol{\theta}). \quad (15)$$

The variance $\sigma_\lambda^2(\boldsymbol{\theta})$ typically increases with J and, in fact, the following *variance condition*

$$J/\sigma_\lambda^2(\boldsymbol{\theta}) \quad \text{is bounded as } J \rightarrow \infty \quad (\text{VC})$$

is assumed in Osius and Rojek (1992) to derive the asymptotic normality of T_λ . The asymptotic expectation and variance of SD_λ are now taken from (9) and (14) evaluated at the estimate $\hat{\boldsymbol{\theta}}$

$$\mu_\lambda := \mu_\lambda(\hat{\boldsymbol{\theta}}) \quad (16)$$

$$\sigma_\lambda^2 := \sigma_\lambda^2(\hat{\boldsymbol{\theta}}) . \quad (17)$$

Using (VC) and the consistency of $\hat{\boldsymbol{\theta}}$, we get from (12) the following important representation of the standardized power-divergence statistic

$$T_\lambda = \psi_\lambda(\boldsymbol{\theta})/\sigma_\lambda(\boldsymbol{\theta}) + O_p(J^{-1/2}) . \quad (18)$$

We now turn to the computation of $\mu_\lambda(\boldsymbol{\theta})$, $v_\lambda^2(\boldsymbol{\theta})$ and $\mathbf{c}_\lambda(\boldsymbol{\theta})$. For Pearson's statistic X^2 (i.e. $\lambda = 1$) the expectation (9) is constant

$$\mu_1(\boldsymbol{\theta}) = J(K - 1) .$$

The variance (10) is given by (cf. McCullagh and Nelder 1989, p.169)

$$v_1^2(\boldsymbol{\theta}) = 2J(K - 1) + \sum_j \frac{1}{N_j} \left[\sum_k \frac{1}{\pi_{jk}(\boldsymbol{\theta})} - K^2 - 2(K - 1) \right] ,$$

and the covariance vector (11) does not depend on the group sizes N_j :

$$c_{1s}(\boldsymbol{\theta}) = \sum_j \sum_k \frac{1}{\pi_{jk}(\boldsymbol{\theta})} \cdot \frac{\partial}{\partial \theta_s} \pi_{jk}(\boldsymbol{\theta}) , \quad s = 1, \dots, S .$$

Note in particular, how small expected counts $N_j \pi_{jk}(\boldsymbol{\theta})$ can blow up the variance $v_1^2(\boldsymbol{\theta})$.

Unfortunately, explicit expressions like the ones given above are not available for general λ . Nevertheless, the moments $\mu_\lambda(\boldsymbol{\theta})$, $v_\lambda^2(\boldsymbol{\theta})$ and $\mathbf{c}_\lambda(\boldsymbol{\theta})$ can be computed for any λ and $\boldsymbol{\theta}$ in a straight forward way. Introducing for a multinomial $M_K(N, \boldsymbol{\pi})$ random vector \mathbf{Y} the notations

$$e_\lambda(N, \boldsymbol{\pi}) := E_\pi \{ A_\lambda(\mathbf{Y}, N\boldsymbol{\pi}) \} \quad (19)$$

$$v_\lambda^2(N, \boldsymbol{\pi}) := \text{var}_\pi \{ A_\lambda(\mathbf{Y}, N\boldsymbol{\pi}) \} \quad (20)$$

$$c_{\lambda k}(N, \boldsymbol{\pi}) := \text{cov}_\pi \{ Y_k, A_\lambda(\mathbf{Y}, N\boldsymbol{\pi}) \} \quad \text{for } k = 1, \dots, K, \quad (21)$$

the expectation (9) may be written as

$$\mu_\lambda(\boldsymbol{\theta}) = E_\theta \left\{ \sum_j A_\lambda(\mathbf{Y}_j, N_j \boldsymbol{\pi}_j(\boldsymbol{\theta})) \right\} = \sum_j e_\lambda(N_j, \boldsymbol{\pi}_j(\boldsymbol{\theta})) \quad (22)$$

And similarly the independence of all rows \mathbf{Y}_j yields

$$v_\lambda^2(\boldsymbol{\theta}) = \sum_j v_\lambda^2(N_j, \boldsymbol{\pi}_j(\boldsymbol{\theta})) , \quad (23)$$

$$\mathbf{c}_\lambda(\boldsymbol{\theta}) = \left(\sum_j \sum_k \mathbf{c}_{\lambda k}(N_j, \boldsymbol{\pi}_j(\boldsymbol{\theta})) \cdot \frac{\partial}{\partial \theta_s} \log \pi_{jk}(\boldsymbol{\theta}) \right)_{s=1, \dots, S} . \quad (24)$$

The moments (19) to (21) can be computed using the definition of an expectation

$$E\{f(\mathbf{Y})\} = \sum_{\mathbf{y}} P\{\mathbf{Y} = \mathbf{y}\} \cdot f(\mathbf{y}) , \quad (25)$$

where the sum extends over all outcomes of the underlying multinomial distribution. The computational burden increases heavily with the number K of response categories, but is acceptable for moderate K and in particular for the binomial case $K = 2$, later to be treated in detail.

Edgeworth and Saddlepoint Expansions

Let us now turn to a different approach to approximate the tail probabilities of T_λ , which is based on the first four cumulants only and not on a particular limiting distribution (like χ^2 or normal) for the power-divergence statistic. For notational convenience, the parameter λ – being fixed in the general discussion to follow – is omitted as an index. The third and fourth standardized cumulant ρ_3 and ρ_4 of T are measures of *skewness* and *kurtosis* of T (and of SD).

A direct Edgeworth expansion leads to the *Edgeworth approximation* (cf. Barndorff-Nielsen and Cox 1989, Sec. 4.2)

$$P\{T \geq t\} \approx 1 - \Phi(t) + \varphi(t) \left[\frac{1}{6} \rho_3 H_2(t) + \frac{1}{24} \rho_4 H_3(t) + \frac{1}{72} \rho_3^2 H_5(t) \right] \quad (\text{EA})$$

Here Φ and φ denote the distribution and density function of the standard normal distribution $N(0, 1)$, and $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$ and $H_5(x) = x^5 - 10x^3 + 15x$ are Hermite polynomials.

For fixed t , the Edgeworth expansion is of order $O(J^{-3/2})$ and typically quite accurate around the center $t = 0$ but less reliable in the tails, where $H_3(t)$ and $H_5(t)$ will be appreciable and may even cause negative values for (EA), cf. example 1 later. This deficiency can be tempered by using a tilted Edgeworth or saddlepoint expansion, which is basically an Edgeworth expansion for a modified distribution centered around the argument

of interest, see Barndorff-Nielsen and Cox (1989), Sec. 4.3, for details on the following discussion. More precisely, if $M(s)$ resp. $K(s) = \log M(s)$ are the moment resp. cumulant generating function of the standardized variable $Z(\hat{\boldsymbol{\theta}})/\sigma$, we first seek a root $\hat{\tau}$ of the equation

$$K'(\hat{\tau}) = t \quad (26)$$

in order to obtain

$$\begin{aligned} r &= \operatorname{sgn}(\hat{\tau}) \cdot [2\hat{\tau} \cdot K'(\hat{\tau}) - 2K(\hat{\tau})]^{1/2} \\ v &= \hat{\tau} \cdot [K''(\hat{\tau})]^{1/2}. \end{aligned} \quad (27)$$

The *saddlepoint approximation* of the significance level is now given by

$$P\{T \geq t\} \approx 1 - \Phi(r) - \varphi(r) \left[\frac{1}{r} - \frac{1}{v} \right] \quad (\text{SA})$$

with error $O(J^{-1})$ uniformly in the observed value t of T . The computation of $\hat{\tau}$, r and v requires knowledge of the cumulant generation function $K(s)$, which in turn characterizes the unknown distribution. However, Taylor expansions are available

$$\begin{aligned} K(s) &\approx \frac{1}{2}s^2 + \frac{1}{6}\rho_3s^3 + \frac{1}{24}\rho_4s^4, \\ K'(s) &\approx s + \frac{1}{2}\rho_3s^2 + \frac{1}{6}\rho_4s^3, \\ K''(s) &\approx 1 + \rho_3s + \frac{1}{2}\rho_4s^2, \end{aligned} \quad (28)$$

which can be used to approximate $\hat{\tau}$, r and v .

The Edgeworth and saddlepoint approximations and (28) depend on the cumulants ρ_3 and ρ_4 of T , which are difficult to obtain. In view of (18) these cumulants can be approximated with an error $O(J^{-1/2})$ by the corresponding *standardized* cumulants $\rho_3(\boldsymbol{\theta})$ and $\rho_4(\boldsymbol{\theta})$ of $\psi(\boldsymbol{\theta})$, which are given in the next section, cf. (30), (33) and (34). Evaluating the cumulants at the estimate $\hat{\boldsymbol{\theta}}$ leads to the approximation

$$\rho_k \approx \rho_k(\hat{\boldsymbol{\theta}}) \quad \text{for } k = 3, 4 \quad (29)$$

with error $O_p(J^{-1/2})$. Using these approximations in the Edgeworth expansion increases the error in (EA) to the order $O_p(J^{-1})$. The additional error in (SA) due to (28) and (29) are not so easily tractable.

Skewness and Kurtosis of the Statistics

In this section we derive the cumulants $\kappa_{\lambda 3}(\boldsymbol{\theta})$ and $\kappa_{\lambda 4}(\boldsymbol{\theta})$ of $\psi_{\lambda}(\boldsymbol{\theta})$ – with λ reintroduced as an index – from which the skewness and kurtosis are easily obtained as the standardized cumulants

$$\rho_k(\boldsymbol{\theta}) = \kappa_{\lambda k}(\boldsymbol{\theta}) / \sigma_{\lambda}^k(\boldsymbol{\theta}) \quad \text{for } k = 3, 4 \quad (30)$$

The random variable $\psi_{\lambda}(\boldsymbol{\theta})$ may be written as a sum

$$\psi_{\lambda}(\boldsymbol{\theta}) = \sum_j \psi_{\lambda j}(\boldsymbol{\theta})$$

with a contribution $\psi_{\lambda j}(\boldsymbol{\theta}) := Z_{\lambda j}(\boldsymbol{\theta}) + V_{\lambda j}(\boldsymbol{\theta})$ from each group j , having expectation 0, consisting of the sum of the centered derivation

$$Z_{\lambda j}(\boldsymbol{\theta}) := A_{\lambda}(\mathbf{Y}_j, N_j \boldsymbol{\pi}_j(\boldsymbol{\theta})) - e_{\lambda}(N_j, \boldsymbol{\pi}_j(\boldsymbol{\theta}))$$

and a linear combination $V_{\lambda j}(\boldsymbol{\theta}) := \mathbf{d}_{\lambda j}^T(\boldsymbol{\theta}) \cdot \mathbf{Y}_j$ given by the vector

$$\mathbf{d}_{\lambda j}^T(\boldsymbol{\theta}) := -\mathbf{c}_{\lambda}^T(\boldsymbol{\theta}) \cdot \mathbf{I}^{-1}(\boldsymbol{\theta}) \cdot \mathbf{D}^T \log \boldsymbol{\pi}_j(\boldsymbol{\theta}) ,$$

which is orthogonal to $\boldsymbol{\pi}_j$. Since the groups are independent under product-multinomial sampling, the m -th cumulant $\kappa_{\lambda m}(\boldsymbol{\theta})$ of $\psi_{\lambda}(\boldsymbol{\theta})$ can be obtained as a sum

$$\kappa_{\lambda m}(\boldsymbol{\theta}) = \sum_j \kappa_m(\psi_{\lambda j}(\boldsymbol{\theta})) .$$

Introducing for a multinomial $M_K(N, \boldsymbol{\pi})$ vector \mathbf{Y} and a vector $\mathbf{d} \in \mathbb{R}^K$ orthogonal to $\boldsymbol{\pi}$ the notation

$$\mu_{\lambda m}(N, \boldsymbol{\pi}, \mathbf{d}) := E_{\pi} \{ [A_{\lambda}(\mathbf{Y}, N\boldsymbol{\pi}) - e_{\lambda}(N, \boldsymbol{\pi}) + \mathbf{d}^T \mathbf{Y}]^m \} , \quad (31)$$

we can express the (central) moments of the random variable $\psi_{\lambda j}(\boldsymbol{\theta})$ as

$$E_{\theta} \{ \psi_{\lambda j}^m(\boldsymbol{\theta}) \} = \mu_{\lambda m}(N_j, \boldsymbol{\pi}_j, \mathbf{d}_{\lambda j}(\boldsymbol{\theta})) . \quad (32)$$

The cumulants are finally obtained from the central moments, using the relations $\kappa_3 = \mu_3$ and $\kappa_4 = \mu_4 - 3\mu_2^2$, as follows

$$\kappa_{\lambda 3}(\boldsymbol{\theta}) = \sum_j \mu_{\lambda 3}(N_j, \boldsymbol{\pi}_j, \mathbf{d}_{\lambda j}(\boldsymbol{\theta})) , \quad (33)$$

$$\kappa_{\lambda 4}(\boldsymbol{\theta}) = \sum_j \left[\mu_{\lambda 4}(N_j, \boldsymbol{\pi}_j, \mathbf{d}_{\lambda j}(\boldsymbol{\theta})) - 3\mu_{\lambda 2}^2(N_j, \boldsymbol{\pi}_j, \mathbf{d}_{\lambda j}(\boldsymbol{\theta})) \right] . \quad (34)$$

For Pearson's statistic X^2 , explicit expressions in terms of multinomial moments are available for the moments (31), which are given in section 8 for the binomial case $K = 2$. But for general λ , these moments can only be evaluated using the definition (25).

Bootstrapping the Statistics

A radically different approach to the asymptotic expansions discussed so far is based on a simulation technique, the bootstrap method introduced by Efron (1979) – more recent accounts are given in the book Hall (1992) and the conference proceedings LePage and Billard (1992) and Jckel, Rothe and Sendler (1992). The increasing access of statisticians to powerful and fast computers has established simulation techniques as an important tool to investigate the accuracy of approximations derived from asymptotic results. A simulation study to examine the adequacy of the proposed approximations can only cover particular choices for: (a) the model, given by K and functions G_k , (b) the number J of groups, (c) the sizes N_j and covariates \mathbf{x}_j , and (d) the parameter $\boldsymbol{\theta}$. However, in any practical application of a goodness-of-fit test, (a) to (c) are given and a consistent estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + O_p(n^{-1/2})$ is available for the unknown parameter. This suggests the following simulation technique known as the *parametric bootstrap*.

A contingency table $\mathbf{Y}^* = (Y_{jk}^*)$ is called a *resample* of the original counts $\mathbf{Y} = (Y_{jk})$, if its distribution has the same parametric form as the distribution of \mathbf{Y} , with the unknown parameter $\boldsymbol{\theta}$ replaced by its estimate $\hat{\boldsymbol{\theta}}$ based on \mathbf{Y} , i.e. all rows \mathbf{Y}_j^* are independent with a $M_K(N_j, \boldsymbol{\pi}_j(\hat{\boldsymbol{\theta}}))$ distribution. All terms derived from the resample \mathbf{Y}^* are marked by a star, e.g. $\hat{\boldsymbol{\theta}}^*$ is the estimate and $T^* = (SD^* - \mu^*)/\sigma^*$ the standardized power-divergence for Y^* . The distribution of T is now estimated by the distribution of its resample T^* , which leads to the following *bootstrap estimate* for the significance level

$$p := P\{T \geq t\} \approx P\{T^* \geq t\} . \quad (\text{BE})$$

Since T is (asymptotically) pivotal (i.e. its limit distribution does not depend on the parameter $\boldsymbol{\theta}$), the accuracy of this estimate can be obtained as follows (see Hall and Titterton 1989 or Hall 1992, sec. 3.1). The difference between the simple Edgeworth expansions (using the H_2 term only) for both sides in (BE) has the same order as

$$\rho_3 - \rho_3^* = O_p(J^{-1/2}n^{-1/2}) = AM^{-1/2} \cdot O_p(J^{-1}) , \quad (35)$$

since $\rho_3 = O(J^{-1/2})$. Hence the error in (BE) is for a *bounded* arithmetic mean AM of the group sizes not larger than the error $O_p(J^{-1})$ in (EA), but may be considerably less for *unbounded* AM . It must be emphasized, that the order of the error is not maintained if we bootstrap SD (which is not pivotal) instead of T (Hall and Wilson 1991). Note however, that T is still pivotal under *fixed-cells* asymptotics, having a standardized χ^2 limit.

The *bootstrap approximation* for p is obtained from a sufficiently large number R of independent resamples $\mathbf{Y}_1^*, \dots, \mathbf{Y}_R^*$ as the observed frequency

$$\hat{p} = \#\{r|T_r^* \geq t\}/R. \tag{BA}$$

and has variance $\hat{p}(1 - \hat{p})/R$. The number R of resamples can be chosen according to the accuracy desired for \hat{p} . For less extreme tail probabilities, e.g. $p \geq 1\%$, about $R = 10\,000$ resamples should be sufficient.

The generation of a resample \mathbf{Y}^* requires – independently for each group – the generation of a $M_K(N, \boldsymbol{\pi})$ random vector for given N and $\boldsymbol{\pi}$, which can be taken as the sum of N independent $M_K(1, \boldsymbol{\pi})$ random vectors. And drawing a random vector \mathbf{Z} from a $M_K(1, \boldsymbol{\pi})$ distribution merely requires a uniform random number U from the unit interval $(0,1)$. The unique category $k = 1, \dots, K$ for which $Z_k = 1$ being given by the condition $\gamma_{k-1} < U \leq \gamma_k$, where $\gamma_k = \pi_1 + \dots + \pi_k$ are the cumulative probabilities.

The bootstrap method is time-consuming, because for each resample \mathbf{Y}^* the estimate $\hat{\boldsymbol{\theta}}^*$ has to be computed and this typically requires an iterative procedure. In this case it is reasonable to reduce the accuracy of the fitting procedure and stop the iteration if the deviance D (being among the statistics of interest) is suitably stable. However, a more radical one-step procedure (starting with the original estimate $\hat{\boldsymbol{\theta}}$) can not be generally recommended, because from our experience the power-divergence statistics thus obtained are not stable enough in sparse data.

The bootstrap estimate of the expectation μ , variance σ^2 and higher cumulants ρ_m of T are the corresponding cumulants of T^* . The error here is $O_p(n^{-1/2}) = AM^{-1/2} O_p(J^{-1/2})$ which is, for *bounded* AM , of the same order $O_p(J^{-1/2})$ as the error in the approximations (16), (17) and (29) based on (18), but may again be considerably less for *unbounded* AM .

Binomial Models (Binary Data)

In many situations the response of interest may be classified in two categories, usually termed “success” and “failure”. The previous presentation for general K will now be specialized to $K = 2$, leading to some simplifications. For notational simplicity, the row counts $\mathbf{Y}_j = (Y_{j1}, Y_{j2})$ are identified with the first component (success), i.e. $Y_j := Y_{j1}$, the “failures” being determined by $Y_{j2} = N_j - Y_j$. The product-multinomial sampling model now states that all Y_j are independent, each having a binomial $B(N_j, \pi_j)$ distribution, where $(\pi_j, 1 - \pi_j)$ corresponds to our previous probability vector $\boldsymbol{\pi}_j$. And the model (1) reduces to a generalized linear model

$$\pi_j = \pi_j(\boldsymbol{\theta}) = G(\mathbf{x}_j^T \boldsymbol{\theta}) \quad \text{resp.} \quad \eta_j := g(\pi_j) = \mathbf{x}_j^T \boldsymbol{\theta} \tag{36}$$

with link function $g = G^{-1}$ (cf. McCullagh and Nelder (1989)). The “distance measure” $A_\lambda(p, \pi)$ for $0 \leq p \leq 1$ and $0 < \pi < 1$ is given by

$$A_\lambda(p, \pi) = a_\lambda(p, \pi) + a_\lambda(1 - p, 1 - \pi). \tag{37}$$

Fixing π , the function $A_\lambda(-, \pi)$ is convex with a minimum $A_\lambda(\pi, \pi) = 0$, leading to bounds

$$\begin{aligned} A_\lambda(p, \pi) &\leq A_\lambda(0, \pi) && \text{for } p \leq \pi \\ A_\lambda(p, \pi) &\leq A_\lambda(1, \pi) && \text{for } p \geq \pi \end{aligned} \quad (38)$$

And the power-divergence statistic is

$$SD_\lambda(\hat{\boldsymbol{\theta}}) = \sum_j N_j A_\lambda(\hat{P}_j, \hat{\pi}_j), \quad (39)$$

with $\hat{P}_j := Y_j/N_j$ as the observed proportion of success in group j . The expectation (25) required to compute the moments reduces to

$$E\{f(Y)\} = \sum_{y=0}^N b(y) \cdot f(y), \quad (40)$$

where Y is $B(N, \pi)$ -distributed with probabilities $b(y) = P\{Y = y\}$. Unless N is small, it is advisable to compute the sum by starting with the “middle term” $y_0 := \text{Int}(N\pi)$ where $b(y)$ attains its maximum, and then working downwards $y_0 - 1, \dots$, as well as upwards $y_0 + 1, \dots$ using the recurrence relation between successive values of $b(y)$. For large N (> 100 , say) it may not be necessary to complete the sum down to $y = 0$ or up to $y = N$ because the individual contributions become too small to affect the internal representation of the sum in the computer. Since (38) provides bounds for the functions $f(y)$ of interest, an appreciable lower and upper part of the sum may be omitted for large N .

The vector \mathbf{c}_λ from (21) is determined by the value

$$c_\lambda(N, \pi) = \text{cov}\{Y, A_\lambda(Y, N\pi)\} \quad (41)$$

for a $B(N, \pi)$ -variable Y via $c_{\lambda 1} = c_\lambda$ and $c_{\lambda 2} = -c_\lambda$. And the components of covariance vector (24) reduce to

$$c_{\lambda s}(\boldsymbol{\theta}) = \sum_j x_{js} \cdot c_\lambda(N_j, \pi_j) \cdot G'(\eta_j)/\pi_j(1 - \pi_j), \quad s = 1, \dots, S. \quad (42)$$

In order to evaluate the cumulants $\kappa_{\lambda m}(\boldsymbol{\theta})$ of $\psi_\lambda(\boldsymbol{\theta})$ for $m = 3, 4$ we introduce for a $B(N, \pi)$ -variable Y and $d \in \mathbb{R}$ the following moments corresponding to (31)

$$\mu_{\lambda m}(N, \pi, d) = E_\pi\{[A_\lambda(Y, N\pi) - e_\lambda(N, \pi) + d(Y - N\pi)]^m\} \quad (43)$$

Using this notation, equation (32) remains valid with the vector $\mathbf{d}_{\lambda j}(\boldsymbol{\theta})$ replaced by the scalar

$$d_{\lambda j}(\boldsymbol{\theta}) := \mathbf{c}_\lambda^T(\boldsymbol{\theta}) \cdot \mathbf{I}^{-1}(\boldsymbol{\theta}) \cdot \mathbf{b}_j(\boldsymbol{\theta}),$$

where the vector $\mathbf{b}_j(\boldsymbol{\theta}) \in \mathbb{R}^S$ is defined by

$$b_{js}(\boldsymbol{\theta}) := x_{js} \cdot G'(\eta_j) / \pi_j(1 - \pi_j) .$$

For Pearson's statistic X^2 we get the familiar quadratic "distance"

$$A_1(p, \pi) = \frac{(p - \pi)^2}{\pi(1 - \pi)} ,$$

from which the moments (19-20) and (41) are easily obtained as

$$\begin{aligned} e_1(N, \pi) &= 1 , \\ v_1^2(N, \pi) &= 2 + \frac{1}{N} \left[\frac{1}{\pi(1 - \pi)} - 6 \right] , \\ c_1(N, \pi) &= 1 - 2\pi . \end{aligned}$$

Furthermore, the moments (43) can be computed for $\lambda = 1$ using the first eight central moments μ_m of the *standardized* binomial $B(N, \pi)/s$, with $s^2 = N\pi(1 - \pi)$, obtained from a recurrence formulae (cf. Johnson and Kotz 1969, chap. 3):

$$\begin{aligned} \mu_3 &= (1 - 2\pi)/s, \\ \mu_4 &= 3(1 - 2/N) + 1/s^2, \\ \mu_5 &= \mu_3 \cdot [2(5 - 6/N) + 1/s^2], \\ \mu_6 &= 5(3 - 26/N + 24/N^2) + 5(5 - 6/N)/s^2 + 1/s^4, \\ \mu_7 &= \mu_3 \cdot [3(35 - 154/N + 120/N^2) + 4(14 - 15/N)/s + 1/s^4], \\ \mu_8 &= 7[(15 - 340/N + 1044/N^2 - 720/N^3) \\ &\quad + 2(35 - 154/N + 120/N^2)/s^2 + (17 - 18/N)s^4] + 1/s^6, \end{aligned}$$

thus giving the moments

$$\begin{aligned} \mu_{12}(N, \pi, d) &= \mu_4 - 1 + 2ds\mu_3 + d^2s^2, \\ \mu_{13}(N, \pi, d) &= (\mu_6 - 3\mu_4 + 2) + 3ds(\mu_5 - 2\mu_3) \\ &\quad + 3d^2s^2(\mu_4 - 1) + d^3s^3\mu_3, \\ \mu_{14}(N, \pi, d) &= (\mu_8 - 4\mu_6 + 6\mu_4 - 3) + 4ds(\mu_7 - 3\mu_5 + 3\mu_3) \\ &\quad + 6d^2s^2(\mu_6 - 2\mu_4 + 1) + 4d^3s^3(\mu_5 - \mu_3) + d^4s^4\mu_4. \end{aligned}$$

Examples

Let us now look at two examples for binary data, one study on infant mortality and one on cancer. Both examples will serve us merely to illustrate the different approximations for the significance level of power-divergence statistics, and no attempt is made here to discuss the studies in detail or to judge the fit using alternative methods.

Example 1: A Study on Infant Mortality

In a larger study on infant mortality Karn and Penrose (1951-52) analyzed data assembled from records of U.C.H. Obstetric Hospital for the years 1935-46, containing 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 relates 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 of models see McCullagh and Nelder 1989, sec. 3.4). We investigate the fit of this model only for the *female* infants in more detail.

The classification of all (female) infants according to W and T yields $J = 345$ groups with a wide range of sizes from 1 to 265 (see table 1) which is typical for larger studies with several observed variables. The relevant information on the power-divergence statistics D , X^2 , FT and CR is summarized in table 2. The four statistics differ dramatically, and so do the conclusions based on the classical χ^2 -approximation with $df = 339$. However, this approximation is not justified here, due to considerable sparseness of the data. The large fraction of very low group sizes N_j leads to a low harmonic mean $HM = 2.4$ (and a moderate arithmetic mean $AM = 19.4$). This causes extremely small fitted counts for many groups, the overall (female) mortality rate was only 4.1%. Consequently, the asymptotic expectation and variance of these statistics are quite distinct from those of the limiting χ^2 -distribution.

Passing to the approximation based on the normal- or the rescaled χ^2 -approximation with $df = \nu$ gives comparable results (being quite distinct from the classical χ^2 -approximation). The reduced $df = \nu - S$ yields only for X^2 a considerable decrease of the significance level, but the reduction may not be justified here, since ν is not large. Taking skewness and kurtosis into account, which are not negligible for X^2 and CR , the saddlepoint approximation may be more trustworthy than the Edgeworth expansion,

Size	1	2 – 3	4 – 5	6 – 9	10 – 19	20 +
Percent	27	24	10	10	10	19

Table 1: Distribution of the $J=345$ group sizes N_j for all female infants in example 1.

statistic		FT	D	CR	X^2
parameter λ		$-1/2$	0	$2/3$	1
power-divergence	SD	442.8	325.7	344.6	402.6
rescaled:	SD/β	550.2	523.3	204.4	22.2
	$df = \nu$	506.3	455.7	171.6	19.0
standardized:	T	1.38	2.24	1.77	0.51
<i>cumulants of SD</i>					
expectation	μ	407.5	283.6	289.4	345.0
variance	σ^2	655.9	353.1	975.9	12526.8
skewness	ρ_3	0.1515	0.1669	2.1663	21.5315
kurtosis	ρ_4	0.0339	0.0439	21.6770	779.4321
<i>P-level approximations</i>					
classical χ^2 :	$df = 339$	0.01%	68.81%	40.47%	0.99%
rescaled χ^2 :	$df = \nu$	8.68%	1.53%	4.42%	27.58%
	$df = \nu - 6$	6.09%	0.92%	2.17%	5.28%
normal		8.41%	1.25%	3.84%	30.34%
Edgeworth		8.72%	1.61%	5.75%	(< 0)
saddlepoint		8.73%	1.61%	8.37%	8.73%
bootstrap		5.40%	0.77%	2.38%	7.54%
± S.E.		± 0.23%	± 0.09%	± 0.13%	± 0.26%
<i>bootstrapped cumulants of T</i>					
expectation		- 0.2352	- 0.2636	- 0.2232	- 0.1065
variance		0.9673	0.9614	0.8518	0.5457
skewness	ρ_3	0.1712	0.1991	1.1392	9.2451
kurtosis	ρ_4	- 0.0324	0.0137	5.7147	171.0157

Table 2: Power-divergence statistics, cumulants and significance levels for all female infants in example 1 (study on infant mortality). The bootstrap results are based on $R = 10\ 000$ resamples.

which leads to a “negative probability” for X^2 . Looking finally at the bootstrap estimate for the significance level (based on $R = 10000$ resamples), we find a remarkable agreement with the rescaled χ^2 -approximation based on $df = \nu - S$ for D , FT and CR , but less for X^2 (where the latter approximation is questionable, since ν is not large).

Concerning the final decision of the goodness-of-fit test based on the deviance D , the model is rejected on the common 5% level no matter which approximation is used (except inadequate the classical χ^2). But only the bootstrap and the rescaled χ^2 approximation with reduced df also reject the model on the 1% level (which may be appropriate in view of the large sample size). However, the test based on X^2 , does not reject the model on the 5% level for any approximation, except the inadequate classical χ^2 .

Passing to the corresponding analysis of fit for the *males* we only note, that the bootstrapped significance levels ranged from 21% (for X^2) to 35% (for FT) and again were very near the rescaled χ^2 levels (based on the reduced df), except for X^2 , thus not rejecting the model. For X^2 , the Edgeworth and even the saddlepoint approximations gave “negative” P-levels, due to an extremely high kurtosis. But for the remaining statistics these approximation were consistent with the bootstrapped value.

Turning to the cumulants of T , we observe for the females (cf. table 1), that the bootstrap approximation to the expectation is clearly below the nominal value of 0, and this also holds for the males. The bootstrapped variance is also well below the nominal value of 1 for CR and X^2 among the females, but not among the males.

Example 2:

Ille-et-Vilaine study on Oesophageal Cancer

The data are taken from the Ille-et-Vilaine study on oesophageal cancer as given in Appendix I of Breslow and Day (1980). This is a retrospective case-control study with 975 individuals (200 cases and 775 controls), classified according to the 3 covariables: AGE (6 classes), alcohol (ALC) and tobacco (TOB) consumption (4 classes each). Of the 96 possible combinations only $J = 88$ different groups contained at least one individual at risk, with group sizes from 1 to 60 (see also table 3). Although the sampling scheme of this study is not of the product-binomial type assumed here, we nevertheless apply the approximations for illustration and numerical comparison with significance levels for conditional tests suggested by McCullagh (1985ab).

We only look at the model 1+AGE+ALC+TOB with $S = 12$ parameters, containing the main effects of the covariables viewed as factors. The relevant

Size	1	2 – 3	4 – 5	6 – 9	10 – 19	20 +
Percent	14	17	14	18	22	16

Table 3: Distribution of the $J=88$ group sizes N_j for example 2.

statistic		FT	D	CR	X^2
parameter λ		$-1/2$	0	$2/3$	1
power-divergence	SD	112.4	82.3	79.4	86.6
rescaled:	SD/β	95.7	104.9	76.0	32.8
	$df = \nu$	98.5	107.2	76.8	33.3
standardized:	T	-0.20	-0.16	-0.06	-0.07
<i>cumulants of SD</i>					
expectation	μ	115.6	84.2	80.2	88.0
variance	σ^2	271.4	132.2	167.7	464.4
skewness	ρ_3	0.3121	0.2749	0.5298	1.8451
kurtosis	ρ_4	0.1626	0.1192	0.9143	9.9206
<i>P-level approximations</i>					
classical χ^2 :	$df = 76$	0.43%	28.98%	37.21%	19.13%
rescaled χ^2 :	$df = \nu$	55.98%	54.51%	50.39%	49.42%
	$df = \nu - 12$	23.28%	23.34%	16.08%	5.37%
normal		57.78%	56.27%	52.53%	52.67%
Edgeworth		55.83%	54.51%	49.16%	41.90%
saddlepoint		55.84%	54.51%	49.18%	42.65%
bootstrap		24.19%	20.52%	17.54%	16.92%
± S.E.		± 0.43%	± 0.44%	± 0.38%	± 0.37%
<i>bootstrapped cumulants of T</i>					
expectation		- 0.8204	- 0.9233	- 0.7905	- 0.4634
variance		0.8771	0.8788	0.7081	0.5904
skewness	ρ_3	0.2705	0.2551	0.3426	0.6669
kurtosis	ρ_4	0.0680	0.0771	0.5401	2.4653

Table 4: Power-divergence statistics, cumulants and significance levels for example 2 (Ille-et-Vilaine study on oesophageal cancer). The bootstrap results are based on $R = 10\ 000$ resamples.

information on the four power-divergence statistics D , X^2 , FT and CR is summarized in table 4.

The classical χ^2 -approximation gives quite different P-levels (the one for FT being extremely low), and again appears unreliable in view of the sparseness of the data (cf. table 3). The normal and rescaled χ^2 -approximation with $df = \nu$ gives comparable results for all four statistics, which are more than halved in passing to the reduced $df = \nu - S$, which again may be questionable (for X^2), since ν is not large. For each statistic, the Edgeworth and saddlepoint expansion give almost identical values differing noticeably from the normal approximation only for X^2 (which could be expected from the higher cumulants). The significance level based on the bootstrap with $R = 10\,000$ resamples show moderate variation across the four statistics and are again very close to those from the rescaled χ^2 -approximation with reduced $df = \nu - S$ (except for X^2), but are considerably smaller than those based on the other approximations. One possible explanation is that the approximation (16) for the expectation of SD is not accurate enough here, since the bootstrapped expectation of T is clearly below 0 (cf. table 4).

McCullagh (1985a,1985b) computed the *conditional* cumulants of X^2 given $\hat{\theta} = \theta$ as $\kappa_1 = 77.38$, $\kappa_2 = 401.1$ and $\rho_3 = 1.98$. This leads to a conditional significance level of 23.1% (based on a simple Edgeworth expansion, using only the H_2 term), which is quite different from our bootstrapped value.

Hence the model is not rejected by any of the four statistics, no matter which approximation is used (except for FT and the inadequate classical χ^2) in agreement with other considerations by Breslow and Day (1980, Sec.6.5) confirming a satisfactory fit.

Discussion

The following remarks – reflecting our present theoretical knowledge and (limited) practical experience with power-divergence statistics – should only be taken as rough guidelines for applications. It should also be borne in mind that we have assumed a *large sample size* (at least a few hundred, say) and *models with large degrees of freedom* $J - S$ (at least 50, say), and hence a large number J of multinomial groups throughout, with no attempt to pursue the accuracy of the approximations if these assumptions are violated.

1. Although a goodness-of-fit test is usually based on a *single* power-divergence statistic, like the deviance D or Pearson's X^2 , the calculation of several statistics SD_λ , at least D and X^2 but preferably also FT and CR , turns out to be a useful diagnostic tool to detect low expected counts.

2. The classical χ^2 -approximation for SD_λ with $df = J - S$ is reliable only if all expected counts are not too small, which generally leads to comparable values for the statistics D , X^2 , FT and CR , provided the specified model is correct.
3. If considerable variation is observed between D and X^2 (and FT , CR), leading to markedly different P-levels for the classical χ^2 -approximation, then the normal- or rescaled χ^2 -approximation are preferable, which typically give similar results. The latter is preferable, using the reduced $df = \nu - S$ for large ν , because it provides a smooth transition between the classical χ^2 - and the normal approximation (and was very close to the bootstrapped value in our examples, except for X^2).
4. Calculation of the third and fourth moment together with the Edgeworth or preferably the saddlepoint approximation is highly desirable to detect skewness and kurtosis. Particularly for sparse data the examples revealed an appreciable skewness and kurtosis of X^2 (but not for D and FT), leading to an insufficient accuracy of the saddlepoint approximation (SA), which may partly be due to extreme discreteness in the upper tail of X^2 . In this situation the deviance D and FT may be favoured for a goodness-of-fit test, since they have less skewness and kurtosis (leading to comparable results for *all* approximations in the examples). Pearson's X^2 , however, has other advantages: its expectation is constant, and explicit expression for all relevant moments simplifies the computation.
5. The parametric bootstrap is intuitively appealing and at present seems to be the most accurate way to approximate the significance level of any T_λ . Furthermore, the bootstrap remains valid also under fixed-cells asymptotics, thus eliminating the need to choose between the two asymptotics in practice. However, the computing time is now multiplied by the number R of resamples, which should be about 10 000 to obtain a reliable significance level within the range of interest (down to 1%).
6. The additional time to compute the moments and cumulants necessary for all approximations took (in our examples) roughly twice the time as to fit the model. All computations (including the bootstrap) can be performed even on personal computers taking from a few minutes to several hours, depending on the computer's CPU, the sample size and the dimension of the model.
7. Bootstrapping the cumulants of SD_λ in the examples suggests, that the expectation μ_λ tends to be underestimated by (16). An ad-hoc

correction, motivated by (6), is to replace μ_λ by $\mu_\lambda - S$. In our examples, this brings all approximations based on μ_λ much closer to the bootstrapped value, but further investigation concerning the approximation of μ_λ are needed.

In conclusion, for large (and possibly sparse) data and large degrees of freedom, the rescaled χ^2 - and saddlepoint approximations (RCA) and (SA) provide a substantial improvement over the often misleading classical χ^2 -approximation for the favorite statistics D , X^2 as well for FT and CR , and bootstrapping is even preferable, but much more computer intensive. The unique effort to implement these methods on a computer is not very large (as compared to the total time spent on a single larger study) and enables a state-of-the-art evaluation of significance levels for the power-divergence statistics in all future applications.

References

- Barndorff-Nielsen, O.E. and Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics* London: Chapman and Hall.
- Breslow, N.E. & Day, N.E. (1980). *Statistical methods in cancer research, Volume I: The analysis of case-control studies*. International Agency for Research on Cancer, Lyon.
- Cochran, W.G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics* 23, 315.
- Cressie, N.A.C., & Read, T.R.C. (1984), Multinomial Goodness-of-Fit tests. *Journal of the Royal Statistical Society, Ser. B*, 46, 440-464.
- Dale, J.R. (1986), Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Product Multinomials. *Journal of the Royal Statistical Society, Ser. B*, 48, 48-59.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1-26.
- Haberman, S.J. (1974). *The analysis of frequency data*. The University of Chicago Press, Chicago and London.
- Hall, P. (1992). *The bootstrap and Edgeworth expansions*. New York: Springer.
- Hall, P. & Titterton, D.M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *Journal of the Royal Statistical Society, Ser. B*, 51, 459-467.
- Hall, P. & Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* 47, 757-762.
- Jckel, K.-H., Rothe, G. & Sendler, W. (Eds.) (1992). *Bootstrapping and related techniques*. New York: Springer.
- Johnson, N.L. & Kotz, S. (1969). *Distributions in statistics: Discrete distributions*. Houghton Mifflin Company, Boston.

- Karn, M.N. & Penrose, L.S. (1951-52). Birth weight and gestation time in relation to maternal age, parity and infant survival. *Annals of Eugenetics* (London) 16, 147-164.
- Koehler, K.J (1986), Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables. *Journal of the American Statistical Association* 81, 483-493.
- LePage, R. & Billard, L. (Eds.) (1992). *Exploring the limits of bootstrap*. New York: Wiley.
- McCullagh, P. (1985a), On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential Family Models. *International Statistical Review* 53, 61-67.
- McCullagh, P. (1985b), Sparse Data and Conditional Tests. *Bulletin of the International Statistical Institute, Proceedings of the 45th Session of ISI* (Amsterdam), Invited Paper 28.3,1-10.
- McCullagh, P. (1986), The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data. *Journal of the American Statistical Association* 81, 104-107.
- McCullagh, P. & Nelder, J.A. (1989), *Generalized Linear Models* (2nd. Ed.). London: Chapman and Hall.
- Osius, G. (1985), Goodness-of-Fit Tests for Binary Data With (Possible) Small Expectations but Large Degrees of Freedom. *Statistics & Decision*, Suppl. No. 2, 213-224.
- Osius, G. & Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association* 87, 1145-1152.
- Read, T.R.C., & Cressie, N.A.C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer.
- Rojek, D. (1989), *Asymptotik für Anpassungstests in Produkt Multinomial Modellen bei wachsendem Freiheitsgrad*. Unpublished Ph.D. Thesis, University of Bremen, Germany.