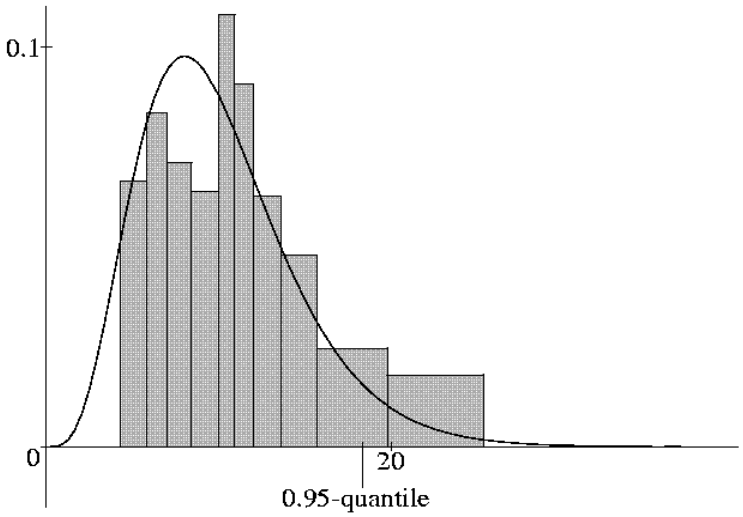
UNIVERSITÄT POTSDAM
Institut für Mathematik

Statistical Scaling of Categorical Data

Henning Läuter

Ayad Ramadan



Mathematische Statistik und
Wahrscheinlichkeitsrechnung

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

Statistical Scaling of Categorical Data

Henning Läuter and Ayad Ramadan

Institute of Mathematics, University of Potsdam
e-mail: laeuter@uni-potsdam.de
ayad_math@yahoo.com

Preprint 2010/01

Januar 2010

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik der Universität Potsdam

Adresse: Am Neuen Palais 10
14469 Potsdam

Telefon: +49-331-977 1500
Fax: +49-331-977 1578
E-mail: neisse@math.uni-potsdam.de

ISSN 1613-3307

Statistical Scaling of Categorical Data

Henning Läuter[1] **and Ayad Ramadan**[2]

[1] *University of Potsdam, Institute of Mathematics*
D-14469 Potsdam, Germany
E-mail: laeuter@uni-potsdam.de
[2] *University of Potsdam, Institute of Mathematics*
D-14469 Potsdam, Germany
E-mail: ayad_math@yahoo.com

Abstract

Estimation and testing of distributions in metric spaces are well known. R.A. Fisher, J. Neyman, W. Cochran and M. Bartlett achieved essential results on the statistical analysis of categorical data. In the last 40 years many other statisticians found important results in this field.
Often data sets contain categorical data, e.g. levels of factors or names. There does not exist any ordering or any distance between these categories. At each level there are measured some metric or categorical values. We introduce a new method of scaling based on statistical decisions. For this we define empirical probabilities for the original observations and find a class of distributions in a metric space where these empirical probabilities can be found as approximations for equivalently defined probabilities. With this method we identify probabilities connected with the categorical data and probabilities in metric spaces. Here we get a mapping from the levels of factors or names into points of a metric space. This mapping yields the scale for the categorical data.
From the statistical point of view we use multivariate statistical methods, we calculate maximum likelihood estimations and compare different approaches for scaling.

Key words: Multivariate scaling, discrimination, power of multivariate tests

1 Introduction

Estimation and testing for distributions of metric random variables are known since the end of the nineteenth century. R.A. Fisher and many other statisticians developed very efficient statistical methods for analyzing medical and biological data. These methods correspond to regression, multivariate analysis and in general to data analysis. Many procedures, e.g. the procedures of the analysis of variance belong to the basic methods in applied statistics.
Essential contributions about statistics of categorical data were developed first by R.A. Fisher, J. Neyman, W. Cochran and M. Bartlett. One finds very different strong results for analyzing categorical data since the 1960s. Mostly data structures from social, biological, medical and technical areas are analyzed. In biomedical applications categories as sex, race or social strata are considered, in technical problems one works with technical patterns or places. In social problems one uses verbal assessments or marks, in political or philosophical context one finds arrangements as ”liberal”, ”moderate” or ”conservative”.
In this paper we introduce a method of scaling based on statistical decisions, especially classification methods are used. We will concentrate on methods and examples with categorical data. But it will be clear that the proposed procedures can be used as a pretreatment in other data structures for generating such transformed data which conform with assumptions in standard software.
Multidimensional scaling is considered by several authors. In most of the cases they use similarities or dissimilarities and then they find scales for the categories (Everitt and Dunn (2001)).

2 Basic model and estimation

2.1 Basic model

We consider a parametric family of multivariate multinomial distributions. The parameters are partly global which have an influence on each component of the random vectors and partly such ones which correspond only to one component.
Let

$$\Theta = \mathcal{M} \times \mathcal{T}$$

be the parameter space. For given probabilities $p_1, ..., p_k$ with

$$p_i \geq 0, \quad \forall i, \quad \sum_{i=1}^{k} p_i \, = \, 1$$

and a positive integer n we denote by $\mathsf{Mult}(n, p_1, ..., p_k)$ a k-nomial distribution.

Definition 1. For given L and $n_1, ..., n_L$ the class of distributions of the vector

$$W = (W_1, ..., W_L)$$

for independent $W_l, \quad l = 1, ..., L$ and

$$W_l \sim \mathsf{Mult}(n_l, p_1(\mu, t_l), ..., p_k(\mu, t_l)), \quad l = 1, ..., L$$

for $(\mu, t_l) \in \Theta \quad \forall \quad l$ and probabilities $p_1(\mu, t_l), ..., p_k(\mu, t_l)$ with

$$p_i(\mu, t_l) \geq 0, \quad \forall i, \quad \sum_{i=1}^{k} p_i(\mu, t_l) \, = \, 1$$

is called the basic model.

Voinov and Nikulin (1993) considered multivariate multinomial distributions for identically distributed W_l, here we use a more general model. We observe realizations w of W with

$$w \, = \, (h_{11}, ..., h_{1k}, h_{21}, ..., h_{Lk}).$$

Here all frequencies h_{li} are nonnegative, $(h_{l1}, ..., h_{lk})$ is a realization of W_l with

$$\sum_{j=1}^{k} h_{lj} \, = \, n_l, \quad \mathsf{P}(h_{l1}, ..., h_{lk}) \, = \, \frac{n_l}{h_{l1}! \cdot ... \cdot h_{lk}!} p_1(\mu, t_l)^{h_{l1}} \cdot ... \cdot p_k(\mu, t_l)^{h_{lk}}. \tag{1}$$

Such observation w can be represented in form of a table

Frequencies	h_{11}	h_{21}	h_{31}	$\cdots$	h_{L1}
	h_{12}	h_{22}	h_{32}	$\cdots$	h_{L2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	h_{1k}	h_{2k}	h_{3k}	$\cdots$	h_{Lk}
marginal sums	$h_{1+} = n_1$	$h_{2+} = n_2$	$h_{3+} = n_3$	$\cdots$	$h_{L+} = n_L$

2.2 Estimation problem

The L and $n_1, ..., n_L$ are given in the basic model. The parameters $\mu, t_1, ..., t_L$ determine the distribution of W. Therefore the estimation problem can be formulated. Given a realization w of W we have to find an estimation of $(\mu, t_1, ..., t_L)$ or only for $(t_1, ..., t_L)$.

2.3 Typical example

We consider the classification of observations as a typical example. Here we consider realizations of p-dimensional random variables with the possible distributions $\mathsf{P}_{\vartheta_1}, ..., \mathsf{P}_{\vartheta_k}$. These continuous or discrete distributions characterize classes $K_1, ..., K_k$ and we assume that densities $f_{\vartheta_1}, ..., f_{\vartheta_k}$ w.r.t. a σ-finite measure are given. Let $\pi_1, ..., \pi_k$ be a priori probabilities for the classes. Let Y be a variable with the density

$$f(y) \, = \, \sum_{j=1}^{k} \pi_j f_{\vartheta_j}(y).$$

Then the conditional probability for $y \in K_i$ given y is a realization Y is determined by

$$\mathsf{P}(y \in K_i \mid Y = y) = \frac{\pi_i f_{\vartheta_i}(y)}{\sum_{j=1}^{k} \pi_j f_{\vartheta_j}(y)} =: \widetilde{p}_i(y). \tag{2}$$

Assuming at first that Y is a discrete random variable. For independent n observations we find n_l-times the value y_l. We denote the frequencies $h_{l1}, ..., h_{lk}$ of the classes $K_1, ..., K_k$ in the n_l observations and see that $(h_{l1}, ..., h_{lk})^t$ is a realization of a multinomial distribution $\mathsf{Mult}(n_l, \widetilde{p}_{l1}, ..., \widetilde{p}_{lk})$ with $\widetilde{p}_{li} = \widetilde{p}_i(y_l)$. From (2) we see that $\widetilde{p}_{li}$ depend on

$$\mu = (\vartheta_1, ..., \vartheta_k, \pi_1, ..., \pi_k)$$

and $(y_1, ..., y_L)$, i.e.

$$\widetilde{p}_{li} =: p_i(\mu, y_l).$$

Hence with (2) we have the representation

$$\widetilde{p}_{li} = p_i(\mu, y_l) = \frac{\pi_i f_{\vartheta_i}(y_l)}{\sum_{j=1}^{k} \pi_j f_{\vartheta_j}(y_l)}. \tag{3}$$

For given $n_1, ..., n_L$ the data can be represented in a $k \times L-$table :

l	1	2	3	$\cdots$	L
class 1	h_{11}	h_{21}	h_{31}	$\cdots$	h_{L1}
class 2	h_{12}	h_{22}	h_{32}	$\cdots$	h_{L2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
class k	h_{1k}	h_{2k}	h_{3k}	$\cdots$	h_{Lk}

Remember that $n_1, ..., n_L$ with

$$n_l = \sum_{i=1}^{k} h_{ti}, \quad l = 1, ..., L$$

. So we can interpret this table as a stratified two-dimensional table. The columns are realizations of independent variables

$$W_l \sim \mathsf{Mult}(n_l, \widetilde{p}_{l1}, ..., \widetilde{p}_{lk}), \quad l = 1, ..., L.$$

We remark that (2) holds also for continuous distributions and therefore the stratified tables are obtained for all considered distributions.

2.3.1 Meaning of the parameters

In subsection 2.3 we considered a typical situation. The parameter μ determines the distribution, the parameters $y_1, ..., y_L$ determine the locations where we observe. Hence we call μ the distributional parameter. The parameter $(y_1, ..., y_L)$ is called the location. At the points $y_1, ..., y_L$ we have the same probabilities or empirical frequencies as at the objects or categories. Hence one calls the parameters $y_1, ..., y_L$ the scale parameters. These scale parameters lie in a metric space and therefore statistical analyses on the categories can be done with these new scale parameters.

2.3.2 Numerical example

We consider a one-way classification problem as it is known e.g. in analysis of variance. At $L = 8$ levels of a factor we observe some electrical resistances given in the table.

One-way classification

Factor A								
Levels	1	2	3	4	5	6	7	8
resistances	64.8	65.7	63.9	65.2	60.2	64.5	61.9	62.5
	63.2	63.1	62.5	63.2	58.9	62.9	60.5	62.1
	61.7	62.8	59.9	62.1	58.3	60.1	59.2	60.1
	65.2	66.3	64.9	65.9	60.6	66.0	60.1	62.6

We suppose that the resistances depend on the levels, otherwise the levels would have the same value on a scale. The levels are characterized by geometric forms, there is no natural ordering or distance between them. Our aim is to find an appropriate scale for the levels and this scale should be dependent on the measured resistances.
On each level resistances are observed, some are small, some others are moderate or high. In this example we say that values in in $I_1 := (64.0, 100]$ are high, in $I_2 := (61.0, 64.0]$ are moderate, in $I_3 := (0, 61.0]$ are small. These are classes $1, 2, 3$. The frequencies for these classes are given in the table.

Frequencies for the classes and levels

Levels	1	2	3	4	5	6	7	8
class 1	2	2	1	2	0	2	0	0
class 2	2	2	2	2	0	1	1	3
class 3	0	0	1	0	4	1	3	1

For each level l we have $n_l = 4$ observations. We denote by $q(j|l)$ the conditional probability for observing values in an interval I_j at level l. Then we can assume that the frequencies at level l are realizations of $\mathsf{Mult}(n_l, q(1|l),\, q(2|l)\, q(3|l))$.
Modelization means that we postulate that there is a space (here now $\mathbb{R}^1$) with distributions $\mathsf{P}_{\vartheta_1}, ..., \mathsf{P}_{\vartheta_3}$ and corresponding densities $f_{\vartheta_1}, f_{\vartheta_2}, f_{\vartheta_3}$ such that with appropriate a priori probabilities the conditional probabilities (3) are near to $q(i|l)$. This assumption says that the probabilities at special points of the metric space are the same as the empirical probabilities coming from the original observations. The estimates of the distributions yield the scaling for levels.

2.4 Estimation of the parameters

2.4.1 Maximum likelihood estimation

We observe $w = (h_{11}, ..., h_{1k}, h_{21}, ..., h_{Lk})$ as a realization of W. With (1) the likelihood function is given by

$$L(\mu, t_1, ..., t_L \mid w) = \prod_{l=1}^{L} \frac{n_l}{h_{l1}! \cdot ... \cdot h_{lk}!} p_1(\mu, t_l)^{h_{l1}} \cdot ... \cdot p_k(\mu, t_l)^{h_{lk}} \tag{4}$$

and therefore the log likelihood is determined up to factors by

$$\tilde{l}(\mu, t_1, ..., t_L \mid w) = \sum_{l=1}^{L} \sum_{i=1}^{k} h_{li} \ln p_i(\mu, t_l). \tag{5}$$

Definition 2. Any value $(\hat{\mu}, \hat{t}_1, ..., \hat{t}_L)$ maximizing $\tilde{l}$, is called **maximum likelihood estimate** for $(\mu, t_1, ..., t_L)$.

2.4.2 Least squares estimate

h_{li}/h_{lj} are well motivated estimates for p_{li}/p_{lj}. Therefore it is a possibility to define estimates for $(\mu, t_1, ..., t_L)$ on the basis of these p_{li}/p_{lj}.

Definition 3.

$$(\hat{\pi}_1, ..., \hat{\pi}_k, \hat{\vartheta}_1, ..., \hat{\vartheta}_k, \hat{y}_1, ..., \hat{y}_L)$$

is called **least squares estimate** for

$$(\pi_1, ..., \pi_k, \vartheta_1, ..., \vartheta_k, y_1, ..., y_L),$$

if

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{L} \Big(\frac{\hat{\pi}_i f_{\hat{\vartheta}_i}(\hat{y}_l)}{\hat{\pi}_j f_{\hat{\vartheta}_j}(\hat{y}_l)} - \frac{n_{li}}{n_{lj}} \Big)^2 = \min_{\pi, \vartheta, a} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{L} \Big(\frac{\pi_i f_{\vartheta_i}(y_l)}{\pi_j f_{\vartheta_j}(y_l)} - \frac{n_{li}}{n_{lj}} \Big)^2. \tag{6}$$

3 Test-based estimation of scaling parameters

The partial parameters $y_1, ..., y_L$ describe the scaling parameters which can be determined separately. M.G. Kendall and A. Stuart (1967) and lateron H. Ahrens and J. Läuter (1981) introduced a method for scaling which bases on a test statistic and the $y_1, ..., y_L$ are determined in such a way that the power of a connected test is maximal.

3.1 Multivariate analysis of variance

We consider k classes with the distributions $\mathsf{N}_p(\nu_1, \Sigma), ..., \mathsf{N}_p(\nu_k, \Sigma)$. We observe

$$v_{11}, ..., v_{1n_1}, ..., v_{k1}, ..., v_{kn_k},$$

where v_{ij} is a realization of the distribution $\mathsf{N}_p(\nu_i, \Sigma)$. For testing the hypothesis

$$\mathcal{H}: \quad \nu_1 = ... = \nu_k \quad \text{against} \quad \mathcal{K}: \quad \nu_i \neq \nu_j \quad \text{for at least one pair} \quad (i,j)$$

often Hotelling's T_0^2 is used. This statistic has the form

$$T_0^2 = \frac{n-k-p+1}{(k-1)(n-k)p} \sum_{t=1}^{k} n_t (v_{t\cdot} - v_{\cdot\cdot})^t S^{-1} (v_{t\cdot} - v_{\cdot\cdot}),$$

where

$$n = \sum_{i=1}^{k} n_k, \quad v_{t\cdot} = \frac{1}{n_t} \sum_{j=1}^{n_t} v_{tj}, \quad v_{\cdot\cdot} = \frac{1}{n} \sum_{t=1}^{k} n_t v_{t\cdot},$$

$$S := \frac{1}{n-k} \sum_{t=1}^{k} \sum_{s=1}^{n_t} \Big(v_{ts} - v_{t\cdot}\Big)\Big(v_{ts} - v_{t\cdot}\Big)^t.$$

With

$$H = \sum_{t=1}^{k} n_t \Big(v_{t\cdot} - v_{\cdot\cdot}\Big)\Big(v_{t\cdot} - v_{\cdot\cdot}\Big)^t$$

we find

$$T_0^2 = \frac{n-k-p+1}{(k-1)(n-k)p}\, tr(HS^{-1}).$$

H. Läuter (2007) showed that this method can be applied to our scaling problem. One defines the observations v_{ij} in an appropriate way. The categories should be denoted by $C_1 z, ..., C_L z$ and the components of the vector z are values describing the categorical variables to be scaled. We observe $m_{tl} =$ number of cases where category l leads to class t. With

$$D_t := \Big(\frac{m_{t1}}{n_t} - \frac{km_{\cdot 1}}{n}\Big) C_1 + ... + \Big(\frac{m_{tL}}{n_t} - \frac{km_{\cdot L}}{n}\Big) C_L$$

and

$$F_{tl} := C_l - \frac{1}{n_t}\Big(m_{t1} C_1 + ... + m_{tL} C_L\Big)$$

we get

$$H := \sum_{t=1}^{k} n_t D_t\, z\, z^t D_t^t, \quad S := \frac{1}{n-k} \sum_{t=1}^{k} \sum_{l=1}^{L} m_{tl} F_{tl}\, z\, z^t F_{tl}^t.$$

Hence we have

$$T_0^2 = \frac{n-k-p+1}{(k-1)(n-k)p}\, z^t \Big[\sum_{i=1}^{k} n_i D_i^t S^{-1} D_i\Big] z. \tag{7}$$

For a good decision in the analysis of variance it is necessary that the observed value of the test statistic is large. Then it is natural to look for such z-values which maximizes T_0^2.

Definition 4. If z^* maximizes T_0^2 then the vectors $C_1 z^*, ..., C_L z^*$ are called **test-maximal scaling** of the categories.

The calculation of these z^* is rather difficult. With methods from optimal experimental design one finds solutions numerically. In special cases explicit solutions are given.

3.1.1 Once again the example

We considered an example in subsection 2.3.2. There we had 8 levels and at each level 4 resistances were measured. The connection between the resistances and the levels is expressed by scaled values. We obtain with the test-maximal values the following table.

Scaled values

Factor A								
Levels	5	7	8	3	6	1	2	4
scaled values	-0.9007	-0.5999	0.0017	0.0670	0.1324	0.4331	0.4331	0.4331

Here we see that the levels $1, 2, 4$ have the same response, also $3, 6$ are similar. The response under levels 5 and $1, 2, 4$ are very different.

4 Comparison of the different scaling

We characterize the different estimated scaled parameters. Especially for two-way classification problems we discuss and compare the maximum-likelihood and the test-maximal scaling. Here it will be shown that the bias of the maximum-likelihood scaling is smaller than for the test-maximal scaling.
At first one sees from (7) that the norm of z^* is not restricted. Therefore the test-maximal scaling cannot be a consistent estimate for some distributional parameters. One has some degrees of freedom in choosing the direction and norm of z^*. Maximum likelihood scale estimates and least squares estimates are consistent under some conditions.

Theorem 1. *For $k = 2$ no consistent estimate exists for $(\mu, t_1, ..., t_L)$.*

Theorem 2. *Let be $\mathsf{P}_{\vartheta_i} = \mathsf{N}(\vartheta_i, 1)$. For $p = 1$, $k \geq 3$, $L \geq 2$ and $0 = \vartheta_1 < \vartheta_2$ the maximum likelihood estimate for $(\mu, t_1, ..., t_L)$ is consistent.*

The result from Theorem 2 holds also for other distributions, but at least properties as monotone likelihood ratio are needed.

Acknowledgement The authors are very grateful to Prof. H. Liero for her helpful comments and suggestions and also to Dr. M. Läuter.

References

Ahrens, H. and J. Läuter (1981). *Mehrdimensionale Varianzanalyse*. Akademie-Verlag.

Everitt, B. and G. Dunn (2001). *Applied Multivariate Data Analysis*. Hodder Education London.

Kendall, M. and A. Stuart (1967). *The Advanced Theory of Statistics*. Griffin&Comp. London.

Läuter, H. (2007). Modeling and scaling of categorical data. Technical report, Univ. Linz.

Voinov, V. and M. Nikulin (1993). *Unbiased Estimators and Their Applications*. Mathematics and its Applications. Kluwer Academic Publishers.