TREND TESTS FOR THE EVALUATION OF DOSE-RESPONSE RELATIONSHIPS IN EPIDEMIOLOGICAL EXPOSURE STUDIES

L.A. HOTHORN[1], M. VAETH[2], T. HOTHORN[3]

[1]Unit of Bioinformatics, University of Hannover, Herrenhaeuser Str. 2, D-30419 Hannover, Germany
[2]Department of Biostatistics, University of Aarhus, Vennelyst Boulevard 6, DK-8000 Aarhus C, Denmark
[3]Department of Medical Informatics, Biometry and Epidemiology, University of Erlangen-Nuremberg, D-91054, Germany

Short title:
TREND TEST IN EPIDEMIOLOGY

Keywords:
Trend test, model selection, order restricted information criterion, binomial data

September, 2003

Correspondence to:
Ludwig A. Hothorn
Phone +495117625566
Fax +495117624966
e-mail: hothorn@bioinf.uni-hannover.de

This paper was worked out during a sabbatical stay of the first author at the Department of Biostatistics of the University of Aarhus under the financial support of the Volkswagen Stiftung grand no. II/78 956

TREND TESTS FOR THE EVALUATION OF DOSE-RESPONSE RELATIONSHIPS IN EPIDEMIOLOGICAL EXPOSURE STUDIES

1. INTRODUCTION

In epidemiological exposure studies frequently an unexposed group E_1 is compared with several exposed groups, an one-way design $[E_1, E_2,...,E_k]$ is assumed here. One important objective in epidemiology is causation. According to Hill (1971)[1] and Weed and Gorelic (1996)[2] the demonstration of a global dose-response relationship belongs to the causation criteria. However, care is needed because hidden bias may produce a dose-response when the factor is without effect Rosenbaum (2003)[3]. In many exposure studies this is the primary objective. The outcome of the study is the number of cases with the investigated disease, e.g. a selected tumor site, and the number without this disease (controls). Throughout this paper we assume a sufficient total number of observations, and therefore asymptotic tests can be used. Exposure studies are performed as cohort studies Takebayashi et al. (2003)[4] where the data represent a 2 by *k* contingency table, or as case control studies Barros-Dios et al. (2003)[5] where two multinomial distributions are compared.

Numerous papers exist on the analysis of dose-response relationships in epidemiological studies, particularly model-based (Royston et al. (1999)[6]) and test-based (Leuraud and Benichou (2001)[7]) approaches. The basic difficulty is that the shape of the dose-response is a-priori unknown- it is an outcome of the study - but the choice of the model or the test depends seriously on that shape. Therefore, inherently a broad class of models or tests should be used and a model selection problem arises. Model selection itself belongs to the delicate problems in statistics. But here, model selection is not the objective; it is only a tool for decision of correct trend. And the outcome of a significant trend test, e.g. a p-value is frequently not sufficient for epidemiologists. In the case of a significant trend they want information on the shape of the dose-response and/or a measure of the magnitude of the effect, e.g. relative risk or odds ratios. Hereby some level for false positive and false negative decision rate, and the correctness of the selected model, should be controlled.

The exposure in case control studies is frequently continuously. Categorization at pre-selected cut-off points to a small number of ordered categories is common. E.g. Pike et al. (1998)[8] estimated the relative risk of breast cancer on four categories of age of menopause and Baumgartner et al. (1998)[9] reported the odds ratio of sarcopenia on six categories of physical disability. Inappropriate chosen cut-off points reduce the power of the trend test seriously according to Greenland (1995)[10]. Some exposures are naturally grouped, e.g. 2 cups coffee/day and by the impreciseness of the definition of a ”cup” and of “coffee” (e.g. in terms of caffeine content), e.g. data between 1.5 and 2.5 cups/day are not really different even if curious pseudo-continuous caffeine contents were calculated into a continuous exposure (Weimann et al. (1997)[11]). If a single cut-off point would exists and would be a-priori known, the best approach is a two sample test “above vs. below the cut-off point” in both terms of power and interpretation (an odds ratio and its one-sided confidence interval can be estimated). For the trend test approach discussed here the worst case is continuous data and a nearly linear shape, while the best case is naturally grouped exposure with a single change point. The reality is almost in-between, i.e. the appropriateness of the

approach depends on the design and the data of the exposure study. For continuous exposure models a continuous covariate can be used. However, the choice of an appropriate model (e.g. linear, logistic, etc.) remains open whereas the model influences the inference. In this paper power-optimal trend tests for the comparison of *k* ordered binomial proportions are discussed. The concept of multiple contrasts is used because of its simplicity and the availability the distribution under the alternative. After a significant trend test, information is provided which contrast is the "best", and therefore which shape of the dose-response is likely. Alternatively, an information criterion-based approach for the likelihood ratio test under total order restriction according to Anraku (1999)[12] is investigated.

Some exposure studies are characterized by few observations with higher exposure, particularly in the highest exposure only some cases are available (fortunately from an ethical point of view). An example is the study on child cancer and magnetic fields from high voltage installations where sample size is 2 in the highest exposure group but 6457 in the unexposed group Olsen et al. (1993)[13]. However, the identification of a trend in such a highly unbalanced design is quite complicated: a significant trend may depend only on these cases, and the size and power of unbalanced designs differ seriously from those in balanced designs.

Design, randomization and definition of the dose are important tasks for evaluation. In preclinical dose-response studies and clinical dose finding studies the dose is a a-priori defined factor with fixed levels, the randomization of the animals/patients is according to the dose groups before dosing and the subject specific outcome is dichotomous: responder/non-responder. Examples are toxicological studies on carcinogenicity with the outcome: tumor/no tumor (Neuhäuser and Hothorn, 1997)[14] and a placebo controlled, double blinded trial to assess dose dependence of any adverse event (present/ absent) by cabergoline in hyperprolactinaemia Webster et al. (1992)[15]. In this randomized clinical trial the patients received either placebo or 0.125, 0.5, 0.75 and 1.0 mg cabergoline. Related are epidemiological cohort studies but the commonly used case control studies are different. To the cases (i.e. subjects with a disease) controls (i.e. subjects without this disease) are randomly matched according to selected criteria. Each subject is characterized by several covariates, while the exposure covariate is commonly continuously and of primary interest. Sometimes, this continuous covariate is post-hoc categorized and the analysis is directed on these virtual exposure levels.

The paper is organized as follows. In chapter 2 some real data examples are discussed under the view of design, categorization, data and interpretation. In chapter 3 is shown, that the likelihood ratio test of identical multinomials against the *L* condition (local odds ratios are non-negative) is asymptotically equivalent to the comparing of *k* independent binomial proportions against a simple ordered alternative. Chapter 4 describes order restricted tests for *k* binomial proportions, and Chapter 5 describes related approaches for identification the kind of alternative. Two different approaches are discussed: the identification of the best contrast and a model selection approach based on the order restricted information criterion. Chapter 6 discussed the more restricted alternative of one change point. In Chapter 7 the relation between the multiple contrast test and the score test in a logistic model is described to allow correction for additional categorical or continuous confounders In Chapter 8 some real data examples are analyzed and interpreted using to above described approaches.

2. EXAMPLES FOR EXPOSURE STUDIES

Some 2 by *k* tables from epidemiological exposure studies from several areas were selected as real data examples. The number of groups varies from 4 to 12, the spontaneous rate from 0 to 0.35, the degree of unbalanceness ($n_{highest} / n_{unexposed}$) from 2.1 to 0.0003 and the relative risk of the highest exposure versus unexposed ($RR_{highest,unexposed}$) from 1.57 to ∞.

A common public health problem is the question if routinely coffee consumption causes any disease. Weinmann et al. (1997)[11] investigated the primary cardiac arrest (PCA) depending on caffeine intake. The summary exposure data (crude, unadjusted) are given in Table 1.

Coffee intake in cups	j	n_{PCA}	$n_{no\ PCA}$	n_j	$\hat{\pi}_j$	RR_{j1}
<1/week	1	79	149	228	0.346	-
1/week …1.9/day	2	48	101	149	0.322	0.930
2…4.9/day	3	126	196	322	0.391	1.129
≥ 5 /day	4	86	72	158	0.544	1.571

Table 1: PCA and coffee consumption

The cut-points: 1/week, 2/day and 5/day were fixed a-priori (with is equivalent to 0-19.5, 19.6-274.6, 274.7-686.6 and 686.7-4120.1 mg caffeine/day) and hence the number of categories (four) was a-priori defined. In the paper no explicit discussion on the choice of the cut-points was found. Unexposed (nonuser) is defined to less than 1 cup/week and the upper cut-point may be due to the meta-analysis result of Greenland (1994)[16] which makes inference for 5 cups/day. The question arises if caffeine cause PCA in a dose-depended manner, and if the shape of the dose-response relationship – based on the pre-selected categories – is plausible. Costantino et al. (1995)[17] published a case-control study for respiratory cancer possibly caused by long-term exposure to coke oven emissions in the period 1966-1975. The definition of the categories seems to be rather formal.

Exposure /mg/m^3-mo	j	n_{cancer}	$n_{no\ cancer}$	n_j	$\hat{\pi}_j$	RR_{j1}
Unexposed	1	72	9426	9498	0.008	-
1-199	2	33	2314	2347	0.014	1.855
200-399	3	26	1181	1207	0.022	2.842
400-599	4	17	550	567	0.030	3.955
>599	5	22	465	487	0.045	5.960

Table 2: Respiratory cancer and PAH

This is an example for an unbalanced design, i.e. only 487 cases in the high exposure group compared with 9498 unexposed cases. A more extreme unbalanced designs was used by Lausen et al. (2002)[18] for the association between all major types of child cancer and magnetic fields from high voltage installations based on continuous exposure data by Olsen et al. (1993)[13]. The original continuous exposure data were categorized using a sieve parameter of 0.05 μTesla.

Exposure μTesla.	j	n_{cancer}	$n_{no\ cancer}$	n_j	$\hat{\pi}_j$	RR_{j1}
0-0.05	1	1698	4759	6457	0.263	-
0.051-0.101	2	0	9	9	0	0.000
0.101-0.15	3	2	3	5	0.4	1.525
0.151-.20	4	1	3	4	0.25	0.953
0.201-0.25	5	1	3	4	0.25	0.953
0.251-0.30	6	0	4	4	0	0.000
0.301-0.35	7	0	2	2	0	0.000
0.351-0.85	8	1	0	2	0.5	1.906
0.851-1.6	9	2	0	2	1	3.812
>1.61	10	2	0	2	1	3.812

Table 3: Child cancer and magnetic fields

Xiong and El Barmi (2002)[19] reported the absence or presence of hypoglycemia in the relation the mean daily insulin levels. The definition of the categories seems to be rather formal.

Insulin level	j	$n_{present}$	n_{absent}	n_j	$\hat{\pi}_j$	RR_{j1}
<.25	1	4	40	44	0.091	
0.251-0.49	2	21	74	95	0.221	2.432
0.5-0.74	3	28	59	87	0.322	3.540
0.75-0.99	4	15	26	41	0.366	4.024
>1	5	12	46	58	0.207	2.276

Table 4: Percentage hypoglycemia and daily insulin level

Slama et al.(2003)[20] investigated the percentage rate of spontaneous abortion between weeks 5 and 20 of pregnancy according to the age of the males in a random cross-sectional population of 1151 French women who had been pregnant between 1985 and 2000 (the strata of 20-24 years old females is given in Table 5). The definition of the categories seems to be rather formal in a five years pattern.

Males age	j	$n_{abortion}$	n_{normal}	n_j	$\hat{\pi}_j$	RR_{j1}
<25	1	33	226	259	0.127	-
25-29	2	37	321	358	0.103	0.811
30-34	3	3	61	64	0.047	0.368
35-39	4	7	5	12	0.583	4.579

Table 5: Abortion rate and male age

In a case-control study of Norwegian nickel-refinery workers, Grimsrud et al. (2002)[21] examined the dose-related associations between smoking adjusted lung cancer rates and cumulative exposure to different forms of nickel (in Table 6 the sulfidic form is presented). The definition of the categories is ordinal.

Sulfidic nickel exposure	j	$n_{lung\ cancer}$	$n_{no\ cancer}$	n_j	$\hat{\pi}_j$	RR_{j1}
Unexposed	1	57	10	67	0.149	-
Low	2	93	27	120	0.225	1.51

Low-medium	3	95	48	143	0.336	2.25
Medium	4	92	42	134	0.313	2.10
Medium-high	5	94	40	134	0.299	2.00
High	6	94	46	140	0.329	2.20

Table 6: Lung cancer and cumulative exposure to sulfidic nickel

Numerous papers have been published on the impact of radiation exposure of Japanese A-bomb survivors on several health effects. Otake et al (1996)[22] reported the incidence of severe mental retardation of prenatal exposed A-bomb survivors (Table 7) and Otake et al (1996)[23] reported the incidence of severe cataracts depending on DS86 eye organ dose (Table 8). The definition of the categories seems to be rather formal in both examples.

Radiation Dose /Gy	$n_{\text{with mr}}$	$n_{\text{without mr}}$	n_j	$\hat{\pi}_j$	RR_{j1}
<0.005	9	1060	1069	0.008	
0.005-0.095	3	209	212	0.014	1.7
0.095-0.495	2	213	215	0.009	1.1
0.495-0.995	4	39	43	0.093	11.0
>0.995	12	14	26	0.462	54.8

Table 7: Severe mental retardation and organ specific radiation dose

Radiation Dose /Gy	$n_{\text{with cat}}$	$n_{\text{without cat}}$	n_j	$\hat{\pi}_j$
<0.005	0	292	292	0.000
0.005-0.494	12	436	448	0.027
0.495-0.994	38	393	431	0.088
0.995-1.994	136	226	362	0.376
1.995-2.994	81	48	129	0.628
2.995-3.994	27	11	38	0.711
>3.995	29	13	42	0.690

Table 8: Severe cataracts depending on DS86 eye organ dose

3. THE ANALYSIS OF CONTINGENCY TABLES UNDER INEQUALITY CONSTRAINTS

Here we consider inequality constraints defined in terms of local odds ratios in a $2\times k$ contingency table. Let L denote the condition that all local odds ratios are non-negative and I the condition of independence then we prove that "The likelihood ratio test of identical multinomials against L (and the maximum likelihood fit) is apparently equivalent to the one for comparing k independent binomial proportions against a simple ordered alternative" (Agresti and Coull, 2002, page 53)[24].
Consider a $2\times k$ contingency table of counts $y_{ij}, i=1,2,\ j=1,\ldots,k$. We shall consider three models for this table:

Model 1: The counts y are realizations of independent Poisson variates with means λ_{ij}. Log-linear parameters $\boldsymbol{\theta}=(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta})$ are introduced by
$\ln(\lambda_{ij})=\alpha+\beta_i+\gamma_j+\delta_{ij}$

satisfying $\sum_i \beta_i = \sum_j \gamma_i = \sum_i \delta_{ij} = \sum_j \delta_{ij} = 0$.

The likelihood function for the parameters in Model 1 is denoted $L_1(\boldsymbol{\theta})$

Model 2: The counts *y* are realizations of *k* independent binomial(n_i, p_i). This model is obtained from Model 1 as the conditional distribution of *y* given $y_{\cdot j} = n_j$ and we have the following relationship between the parameters of Model 2 and Model 1

$$\ln\left(\frac{p_j}{1-p_j}\right) = \beta_2 - \beta_1 + \delta_{2j} - \delta_{1j} = 2\left(\beta_2 + \delta_{2j}\right)$$

The likelihood function for the parameters in Model 2 is denoted $L_2(\boldsymbol{p})$. Since Model 2 can be obtained as a conditional distribution in model 1 we have

$$L_1(\boldsymbol{\theta}) = L_2(\boldsymbol{p}) L_{\bar{2}}(\boldsymbol{\theta})$$

where $L_{\bar{2}}(\boldsymbol{\theta})$ is the likelihood function based on *c* independent Poisson distributions for the margins $y_{\cdot j}$.

Model 3: The counts *y* are realizations of two independent multinomial($m_i, \boldsymbol{\pi}_i$).This model is obtained from Model 1 as the conditional distribution of *y* given $y_{i\cdot} = m_i$ and we have the following relationship between the local odds ratios in Model 3 and the parameters in Model 1

$$\frac{\pi_{2j+1}\pi_{1j}}{\pi_{2j}\pi_{1j+1}} = \exp\left(\delta_{2j+1} - \delta_{2j} - \delta_{1j+1} + \delta_{1j}\right) = \exp\left(2\left(\delta_{2j+1} - \delta_{2j}\right)\right)$$

The likelihood function for the parameters in Model 3 is denoted $L_3(\boldsymbol{\pi})$. Since Model 3 can be obtained as a conditional distribution in model 1 we have

$$L_1(\boldsymbol{\theta}) = L_3(\boldsymbol{\pi}) L_{\bar{3}}(\boldsymbol{\theta})$$

where $L_{\bar{3}}(\boldsymbol{\theta})$ is the likelihood function based on two independent Poisson distributions for the margins $y_{i\cdot}$.

From the relationships between the different parameterizations it follows immediately that the condition *I* of independence in the Poisson model, $\forall i, j : \delta_{ij} = 0$, is identical to the hypothesis $p_1 = \cdots = p_k$ of homogeneity in the binomial model, and the hypothesis of identical multinomial probabilities, $\forall\, j : \pi_{1j} = \pi_{2j}$, in Model 3. Moreover, it is also well-known that the parameter estimates and the fitted values under this hypothesis are identical in the three models.

In Model 2 we now consider the hypothesis given by the order restriction $p_1 \le p_2 \le \cdots \le p_k$ with at least one strict inequality. From the relationships between the parameterizations it follows that the condition alternatively may be expressed as

$$\delta_{21} \le \delta_{22} \le \cdots \le \delta_{2k}$$

or as the condition *L* of non-negative local odds ratios in model 3:

$$\frac{\pi_{2\,j+1}\pi_{1j}}{\pi_{2j}\pi_{1\,j+1}} \geq 0, \quad j = 1,\ldots,k$$

To prove that "The LR test of identical multinomials against L (and the ML fit) is apparently equivalent to the one for comparing k independent binomial proportions against a simple ordered alternative" we need the following lemmas.

Lemma 1. The maximum of the likelihood function $L_{\bar{2}}(\boldsymbol{\theta})$ is obtained for $\hat{\lambda}_{\cdot j} = y_{\cdot j}$ and the maximum value is

$$M_2 = \prod_{j=1}^{k} \frac{y_{\cdot j}^{y_{\cdot j}}}{y_{\cdot j}!} \exp\left(-y_{\cdot j}\right)$$

Moreover, $\quad \forall \boldsymbol{\beta}, \boldsymbol{\delta} \;\; \exists \alpha, \boldsymbol{\gamma}: \quad L_{\bar{2}}(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = M_2$.

Proof: It is well-known that the maximum likelihood estimate of the mean of a Poisson distribution based on a single observation is $\hat{\lambda}_{\cdot j} = y_{\cdot j}$ and the maximum value is simply obtained by inserting these values. The final claim follows by choosing α and $\boldsymbol{\gamma}$ such that

$$\alpha + \gamma_j = \ln(y_{\cdot j}) - \ln\left(\sum_i \exp\left(\beta_i + \delta_{ij}\right)\right)$$

Lemma 2. The maximum of the likelihood function $L_{\bar{3}}(\boldsymbol{\theta})$ is obtained for $\hat{\lambda}_{i\cdot} = y_{i\cdot}$ and the maximum value is

$$M_3 = \prod_{i=1}^{2} \frac{y_{i\cdot}^{y_{i\cdot}}}{y_{i\cdot}!} \exp\left(-y_{i\cdot}\right)$$

Moreover, $\quad \forall \boldsymbol{\gamma}, \boldsymbol{\delta} \;\; \exists \alpha, \boldsymbol{\beta}: \quad L_{\bar{3}}(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = M_3$.

Proof: The proof is identical to that of Lemma 1.

We are now ready to show

Proposition: The likelihood ratio test of identical multinomials against L (and the fitted values) is equivalent to the one for comparing c independent binomial proportions against a simple ordered alternative.

Proof: Consider Model 2 and let $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ denote the maximum likelihood estimates under the order restriction $p_1 \leq p_2 \leq \cdots \leq p_k$ and let $\hat{\beta}_i$ and $\hat{\delta}_{ij}$ denote the corresponding values in the parameterization of Model 1. We then have

$$\max_L L_1(\boldsymbol{\theta}) = \max_L \left(L_2(\boldsymbol{p}) L_{\bar{2}}(\boldsymbol{\theta})\right) \leq \max_L L_2(\boldsymbol{p}) \max_L L_{\bar{2}}(\boldsymbol{\theta}) = L_2(\hat{\boldsymbol{p}}) M_2,$$

since, by Lemma 1, we can always select the parameters α and $\boldsymbol{\gamma}$ such that the final factor is maximal. Also

$$\begin{aligned} L_2(\hat{\boldsymbol{p}}) M_2 &= L_2(\hat{\boldsymbol{p}}) \max_L L_{\bar{2}}(\alpha, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\delta}}) = \max_L \left(L_2(\hat{\boldsymbol{p}}) L_{\bar{2}}(\alpha, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\delta}})\right) \\ &= \max_L \left(L_1(\alpha, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \hat{\boldsymbol{\delta}})\right) \leq \max_L L_1(\boldsymbol{\theta}), \end{aligned}$$

where the second equality follows from Lemma 1. Consequently,
$\max_L L_1(\boldsymbol{\theta}) = L_2(\hat{\boldsymbol{p}})M_2$
In a similar fashion we may show that
$\max_L L_1(\boldsymbol{\theta}) = L_3(\hat{\boldsymbol{\pi}})M_3$
where $\hat{\boldsymbol{\pi}}$ denote the maximum likelihood estimate under condition *L*.
Let LR_2 and LR_3 denote the likelihood ratio tests of *I* against *L* in Model 2 and Model 3, respectively, then

$$LR_2 = \frac{\max_I L_2(\boldsymbol{p})}{\max_L L_2(\boldsymbol{p})} = \frac{\max_I L_1(\boldsymbol{\theta})}{\max_L L_1(\boldsymbol{\theta})} = \frac{\max_I L_3(\boldsymbol{\pi})}{\max_L L_3(\boldsymbol{\pi})} = LR_3$$

and the fitted values are also identical since the maximum likelihood estimates, expressed in terms of the $\boldsymbol{\theta}$-parameterization, coincide.

4. GLOBAL TESTS ON DOSE-RESPONSE RELATIONSHIP FOR PROPORTIONS

In chapter 3 was shown that both case-control and cohort exposure studies can be analyzed by comparing *k* independent binomial proportions against a simple ordered alternative if asymptotic tests based on large sample sizes can be used. The number of diseased and the number of healthy persons for each exposure group E_i are organized in the following 2 by *k* table, where index 1 belongs to group without exposure:

	E_1	….	E_k	Total
With disease	n_{11}		n_{k1}	$n_{.1}$
Without disease	n_{10}		n_{k0}	$n_{.0}$
Sample size	$n_{1.}$		$n_{k.}$	$n_{..}$

Table 9: Principle of 2 by *k* tables for exposure studies

The estimator for the proportions per exposure group is $p_j = n_{j1}/n_{j.} \quad j = 1,...,k$, the total is $p = n_{.1}/n..$ and the expected values for the proportions are denoted as π_j. The hypotheses system for a total order is:

$H_0: \ \pi_1 = \pi_2 = ... = \pi_k$

$H_1: \ \pi_1 \le \pi_2 \le ... \le \pi_k$ with at least one strict inequality

Extensive references on order restricted tests exists, including for 2 by *k* contingency tables, e.g. recently Leuraud and Benichou (2001)[7], Agresti and Coull (2002)[24]. Peddada et al (2001)[25], Bretz and Hothorn (2003)[26] and Salanti and Ulm (2003)[27]. However, no uniform most powerful trend test exists for all possible shapes of the alternative. The possible shapes can be seen as equality-inequality-pattern of H_1. This can be simplified explained for an extreme convex shape, e.g. {0, 0, 0, π}. Clearly the so-called Helmert contrast is most powerful because optimal pooling all lower doses and comparing with the high dose: $C_{Helmert} = p_4 - (p_1 + p_2 + p_3)/3$. On the other hand, for an extreme concave shape, e.g. {0, π, π, π} the Helmert contrast looses strongly power.

But the shape of the exposure-response relationship is a-priori unknown. Irrespective from numerous recent proposals the likelihood ratio test (Robertson et al. ,1988)[28] represents an appropriate solution for the above situation, although it is not uniformly most powerful but almost powerful. This test is numerically complicated, particularly its distribution under the alternative, needed for power/sample size calculations.

The multiple contrast test (Bretz and Hothorn, 2002)[29] approximate its false negative rate and is simpler. For k-1 exposure groups 2^{k-1}-1 different shapes exist, and for each shape a contrast with minimum false negative rate exists. The idea is to select the best contrast, sensitive for a certain shape. To select the best contrast is simply done by a maximum test. A single contrast test is a linear combination of the proportions

$$t_{SC} = \frac{\sum_{j=1}^{k} c_j p_j}{\sqrt{p(1-p)\sum_{j=1}^{k} c_j^2 / n_{j.}}} \text{ (with } \sum_j c_j = 0\text{)}$$

and is asymptotically normal distributed. Different variance estimators can be used, but to keep the problem simple only the total pooled estimator is used here. In epidemiology the total sample size $n_{..} = \sum_{j=1}^{k} n_{j.}$ is commonly large and therefore asymptotic test versions are used throughout here. The contrast coefficients c_j are specific for each contrast test, e.g. for the Helmert contrast [c_j = -1, j = 1, ... , k-1 and c_k = k]. A multiple contrast test is the maximum of s pre-defined single contrast tests $t_{MC} = \max_s(t_{SC}^i)$, $i = 1,...,s$. This maximum statistics is asymptotically s-variate normal distributed with a non-product-moment correlation matrix. The correlation between two contrast tests with contrast coefficients a_j and b_j is $\rho_{a,b} = \frac{\sum_{j=1}^{k} a_j b_j / n_{j.}}{\sqrt{(\sum_{j=1}^{k} a_j^2 / n_{j.})(\sum_{j=1}^{k} b_j^2 / n_{j.})}}$. For the balanced designs with k= 3 and 4 this approach based on q= 3 respective 7 contrasts can be simply demonstrated in Table 10.

Design	Alternative	Contrast c_i
k=3	$\pi_1 < \pi_2 = \pi_3$	{-2 1 1}
	$\pi_1 = \pi_2 < \pi_3$	{-1 -1 2}
	$\pi_1 < \pi_2 < \pi_3$	{-1 0 1}
k=4	$\pi_1 < \pi_2 = \pi_3 = \pi_4$	{-3 1 1 1}
	$\pi_1 = \pi_2 < \pi_3 = \pi_4$	{-1 -1 1 1}
	$\pi_1 = \pi_2 = \pi_3 < \pi_4$	{-1 -1 -1 3}
	$\pi_1 < \pi_2 < \pi_3 < \pi_4$	{-3 -1 1 3}
	$\pi_1 = \pi_2 < \pi_3 < \pi_4$	{-1 -1 0 2}
	$\pi_1 < \pi_2 = \pi_3 < \pi_4$	{-1 0 0 1}
	$\pi_1 < \pi_2 < \pi_3 = \pi_4$	{-2 0 1 1}

Table10: Contrast coefficients for the balanced design with three exposure groups

Size and power of the multiple contrast test was compared with the likelihood ratio test for normal distributed variable (Bretz and Hothorn, 2003)[26] and with the Cochran-Armitage test (Hothorn and Bretz, 2000)[30] for binomial proportions. The power in the unbalanced design was characterized recently (Bretz and Hothorn, 2002)[29]. The above multiple contrast test is defined for differences in proportions, but can be re-formulated for the relative risk, commonly used in epidemiology. The estimators for the relative risk (RR) versus unexposed (j=1) are: $RR_{j1} = \frac{n_{j1}/n_{j.}}{n_{11}/n_{1.}} \quad j = 2,...,k$ The single contrast tests can be formulated for relative risk, e.g. for the reverse Helmert contrast:

$$t_{revHelmert} = \frac{-kp_1 + \sum_{j=2}^{k} p_j}{\sqrt{p(1-p)(\sum_{j=2}^{k} 1/n_j + k^2/n_1)}}$$

$$t^{RR}_{revHelmert} = \frac{-k\frac{n_{11}}{n_{1.}} + \sum_{j=2}^{k}\frac{n_{j1}}{n_{j.}}}{\sqrt{p(1-p)(\sum_{j=2}^{k} 1/n_j + k^2/n_1)}} = \frac{-k + \sum_{j=2}^{k}\frac{n_{j1}}{n_{j.}}\frac{n_{1.}}{n_{11}}}{\frac{n_{1.}}{n_{11}}\sqrt{p(1-p)(\sum_{j=2}^{k} 1/n_j + k^2/n_1)}} = \frac{-k + \sum_{j=2}^{k} RR_{j1}}{\frac{n_{1.}}{n_{11}}\sqrt{p(1-p)(\sum_{j=2}^{k} 1/n_j + k^2/n_1)}}$$

For general contrasts hold true $t^{RR}_{SingleContrast} = \frac{c_1 + \sum_{j=2}^{k} c_j RR_{j1}}{\frac{n_{1.}}{n_{11}}\sqrt{p(1-p)(\sum_{j=1}^{k} c_j^2/n_j)}}$

5. IDENTIFICATION OF THE DOSE-RESPONSE SHAPE

The trend tests decide between the null-hypothesis and the alternative hypotheses globally only, e.g. based on the asymptotical distribution of the test statistics under the null-hypothesis. I.e. either a trend exists or not. But the alternative is not unique. E.g. for a design with three exposure and the unexposed groups the following 2^3-1 hypotheses are possible:

$H_1^1: \pi_1 < \pi_2 = \pi_3 = \pi_4$, $H_1^2: \pi_1 = \pi_2 < \pi_3 = \pi_4$, $H_1^3: \pi_1 = \pi_2 = \pi_3 < \pi_4$,

$H_1^4: \pi_1 < \pi_2 < \pi_3 < \pi_4$, $H_1^5: \pi_1 = \pi_2 < \pi_3 < \pi_4$, $H_1^6: \pi_1 < \pi_2 = \pi_3 < \pi_4$,

$H_1^7: \pi_1 < \pi_2 < \pi_3 = \pi_4$

But the global trend tests give no answer which particular alternative exists. Two different approaches can be used for answering this question, important for epidemiologists: i) the best contrast approach, and ii) model selection approach based on the information criterion for order restriction according to Anraku (1999)[12] . We are interested in an identification of one of the possible 2k-1-1 elementary alternatives, i.e. a classification into $H_1^1,...,H_1^r$. Consequently, the misclassification rate, i.e. the proportion of erroneous identified elementary alternatives is used as a performance measure later on.

5.1 THE BEST CONTRAST APPROACH

The global test decision of the multiple contrast based on the maximum of all included single contrasts $t_{MC} = \max_s(t_{SC}^i), i = 1,...,s$, whereby each single contrast is power optimal for a particular type of alternative (see Table 10 in chapter 4). Therefore, this maximum contrast can be used as an estimator for the dose-response shape (naïve approach). Two versions of this approach can be used: i) classification after a significant trend test only, ii) classification independent of the outcome of the trend test. E.g. for a design with two exposure groups three alternatives are possible: $H_1^1: \pi_1 = \pi_2 < \pi_3$, $H_1^2: \pi_1 < \pi_2 = \pi_3$, $H_1^3: \pi_1 < \pi_2 < \pi_3$. The classification rate for a convex (or concave) shape will be relatively high, because misclassification is given by the linear shape while from a concave (convex) shape this misclassification will be very small. But the classification rate for a linear shape will be smaller because of the two competing alternatives H_1^1, H_1^2. According to the asymmetric testing problem in 2 by *k* tables these misclassification rate are not symmetrically. Generally, the classification rates will increase with decreasing numbers of inequalities in the alternative (see simulation results).

Reformulating this setup, we observe *k* binomial random variables $x_1,...,x_k$ where $x_j = \sum_{i=1}^{n} I(j_i = j)v_i$ and $n_j = \sum_{i=1}^{n} I(j_i = j)$, I denotes the indicator function ($x_i = n_{i1}$ in Table 9). Furthermore, we assume that the parameters $\pi(j)$ are ordered $\pi_1 \le ... \le \pi_k$. A possible dose-response is described by a contrast vector $c = (c_1,...,c_k)$. When s such contrast vectors are given, the problem is to estimate of the underlying dose-response relationship.
A simple estimator is a function $\psi: (x_1,...,x_k) \to \{1,...,s\}$. A simple estimator can be derived from the associated contrast test. Let $T(x_1,...,x_k) = (T_1(x_1,...,x_k),...,T_s(x_1,...,x_k))$ denote the vector of appropriately standardized test statistics for each of the s contrasts. Then $\psi_1 = \psi(x_1,...,x_k) = \arg\max_{l \in \{1,...,s\}} |T_l|$. This approach is denoted the naïve best contrast approach (class_naive).

We are now interested in the variability of the simple estimator ψ_1: how likely is each of the s possible values under the observed data? This question can be addressed via the parametric bootstrap. We draw repeated realizations from *k* binomial distributions with sample sizes n_j and the estimated success parameter $\hat{\pi}(j) = x_j / n_j$ for j=1,…,*k*.

- Draw B bootstrap samples $x_1^{*b},...,x_k^{*b}, x_j^{*b} \sim B(n_j, \hat{\pi}(j)), b = 1,...,B$
- Compute $\psi_1^b = \psi_1(x_1^{*b},...,x_k^{*b})$
- And compute the relative frequency of each possible value from 1,…,s

This is a measure for the variance of the estimator. Under special circumstances, an improved estimator can be computed by majority voting over $\psi_1^b: \psi_2 = \arg\max_{l \in \{1,...,s\}} \sum_b I(\psi_1^b = l)$. This approach is denoted the parametric bootstrap best contrast approach (class_boot).

5.1.1 SIMULATION STUDY FOR THE DESIGN WITH THREE GROUPS

The simulation study is structured in two parts: i) for the general unbalanced design with an unexposed and two exposure groups for both change point alternative and total order alternative (Chapter 5.1.1), ii) for the general unbalanced design for more than two exposure groups for the total alternative (Chapter 5.1.2), while the change point alternative is discussed in Chapter 6.2.

The parametric bootstrap approach was investigated by a simulation study for the two scenarios i) the change point problem with the elementary hypotheses $H_1^1: \pi_1 = \pi_2 < \pi_3$, $H_1^2: \pi_1 < \pi_2 = \pi_3$ and the total order problem with the elementary hypotheses. $H_1^1: \pi_1 = \pi_2 < \pi_3$, $H_1^2: \pi_1 < \pi_2 = \pi_3$, $H_1^3: \pi_1 < \pi_2 < \pi_3$. The comparison between the estimation functions $\psi_1 \, and \, \psi_2$ for the both scenarios and order restricted designs with higher dimensions is discussed in chapter 5.1.2. Table 11 shows the model classification rates for the two scenarios (bold … correct model classification, italic … estimates under the null-hypothesis, underlined … misclassification). Because of the asymmetric test statistics the expected value shapes were investigated for different spontaneous rates π_1 (i.e. the rate of the unexposed group) (balanced sample size of 100). For both scenarios the identification of the convex model is better than for the concave model in the direction of smaller π_1 while for $\bar{\pi} \approx 0.5$ both rates are similar. Both rates are larger for the simpler change point scenario because of the simpler alternative. For the total order scenario the classification rate of the linear model is substantial smaller. The classification rate for a convex (or concave) shape will be relatively high, because the most likely misclassification is by the linear shape and the misclassification from the concave (convex) shape is very small. The classification rate for a linear shape will be smaller because of the two nearly equally competing alternatives convex and concave. According to the asymmetric testing problem in 2 by *k* tables these misclassification rate are not symmetrically. The first row represents the null-hypothesis. The power estimates is the estimation for size, and the model classification estimates are irregular because the power must be larger than 0.5 per definition.

Table 12 represents the influence of different non-centrality parameters Δ for the convex model $\{\pi_1, \pi_1, \pi_1 + \Delta\}$. While for both scenarios the power estimates are similar the increase of the classification rates for the total order scenarios is weaker than for the change point scenario.

	Max(H^1,H^2)			Max(H^1,H^2,H^3)			
π_i	power	H^1	H^2	power	H^1	H^2	H^3
0.3/0.3/0.3	*.050*	*.514*	*.486*	*.048*	*.442*	*.333*	*.224*
0.1/0.1/0.3	.991	**.987**	.004	.989	**.924**	.000	.076
0.1/0.3/0.3	.991	.030	**.961**	.991	.003	**.827**	.170
0.1/0.2/0.3				.969	.175	.196	**.629**

0.2/0.2/0.4	.959	**.936**	.023	.960	**.834**	.004	.124
0.2/0.4/0.4	.966	.040	**.926**	.957	.007	**.796**	.155
0.2/0.3/0.4				.910	.249	.214	**.537**
0.3/0.3/0.5	.940	**.906**	.034	.939	**.799**	.006	.135
0.3/0.5/0.5	.926	.044	**.882**	.930	.010	**.770**	.150
0.3/0.4/0.5				.861	.228	.236	**.536**
0.4/0.4/0.6	.923	**.887**	.036	.926	**.771**	.013	.143
0.4/0.6/0.6	.924	.039	**.885**	.932	.009	**.777**	.147
0.4/0.5/0.6				.872	.238	.236	**.527**

Table 11: Classification rates for the best contrast approach (class_boot) for several spontaneous rates π_1 .(n_i=100)

	Max(H^1,H^2)			Max(H^1,H^2,H^3)			
π_i	power	H^1	H^2	power	H^1	H^2	H^3
0.3/0.3/0.35	.177	**.739**	.261	.182	**.629**	.154	.209
0.3/0.3/0.4	.472	**.861**	.139	.477	**.732**	.067	.201
0.3/0.3/0.45	.764	**.918**	.082	.764	**.791**	.031	.179
0.3/0.3/0.5	.940	**.906**	.034	.941	**.842**	.004	.154
0.3/0.3/0.55	.992	**.986**	.014	.990	**.866**	.003	.131
0.3/0.3/0.6	.999	**.998**	.002	.999	**.915**	.000	.086

Table 12: Classification rates for the best contrast approach (class_boot) for several non-centrality parameters Δ (n_i=100)

As expected, with decreasing sample size the classification rates decrease. For the total order scenario the classification rate of the concave shape becomes larger than these of the convex shape for smaller sample sizes. The rates for the linear shape decrease more than for the convex shape. Clearly estimations with a power < 0.5 are not valid, because the model identification based on a preliminary significant multiple contrast test. From Tables 11-13 a high correlation between power and classification rates could be concluded. However, this relationship is quite complicated, e.g. with higher dimensions *k* up to 10 the classification rates decreases while the power increases (see chapter 5.1.2). In epidemiology frequently unbalanced designs occur, particularly with large sample sizes in the unexposed group and smaller sample sizes in the exposed groups. For a selected sample size pattern in Table 4 the classification rates are given. The classification rate decreases seriously for a convex shape and small sample size in the highest exposure group. For the linear shape even misclassification occurs. Detailed investigations on unbalancedness can be found in Chapter 6.3.

		Max(H^1,H^2)			Max(H^1,H^2,H_3)			
π_i	n_i	power	H^1	H^2	power	H^1	H^2	H^3
0.3/0.3/0.5	150	.991	**.983**	.017	.987	**.873**	.003	.124
0.3/0.5/0.5	150	.984	.020	**.980**	.988	.002	**.858**	.140
0.3/0.4/0.5	150				.970	.188	.178	**.634**
0.3/0.3/0.5	125	.975	**.954**	.021	.973	**.835**	.005	.133
0.3/0.5/0.5	125	.968	.028	**.939**	.976	.003	**.850**	.146
0.3/0.4/0.5	125				.929	.204	.211	**.585**
0.3/0.3/0.5	100	.940	**.906**	.034	.939	**.799**	.006	.135

0.3/0.5/0.5	100	.926	.044	**.882**	.930	.010	**.770**	.150
0.3/0.4/0.5	100				.861	.228	.236	**.536**
0.3/0.3/0.5	75	.869	**.824**	.044	.858	**.694**	.017	.147
0.3/0.5/0.5	75	.852	.048	**.804**	.864	.015	**.788**	.188
0.3/0.4/0.5	75				.767	.273	.246	**.481**
0.3/0.3/0.5	50	.699	**.645**	.054	.713	**.555**	.026	.131
0.3/0.5/0.5	50	.683	.061	**.622**	.699	.040	**.746**	.214
0.3/0.4/0.5	50				.625	.300	.290	**.410**
0.3/0.3/0.5	25	*.450*	*.384*	*.065*	*.486*	*.349*	*.042*	*.094*
0.3/0.5/0.5	25	*.427*	*.073*	*.354*	*.467*	*.084*	*.678*	*.239*
0.3/0.4/0.5	25				*.404*	*.309*	*.330*	*.360*

Table 13: Classification rates for the best contrast approach (class_boot) for several sample sizes

		Max(H^1,H^2)			Max(H^1,H^2,H^3)			
π_i	n_i	power	H^1	H^2	Power	H^1	H^2	H^3
0.3/0.3/0.5	100/100/100	.975	**.954**	.021	.973	**.835**	.005	.133
0.3/0.5/0.5	100/100/100	.968	.028	**.939**	.976	.003	**.850**	.146
0.3/0.4/0.5	100/100/100				.929	.204	.211	**.585**
0.3/0.3/0.5	200/80/20	0.506	**0.455**	0.050	.529	**.523**	.079	.358
0.3/0.5/0.5	200/80/20	0.927	0.026	**0.901**	.928	.000	**.950**	.050
0.3/0.4/0.5	200/80/20				.667	.050	.634	**.315**

Table 14: Classification rates for the best contrast approach (class_boot) for an unbalanced design

5.1.2 SIMULATION STUDY FOR TOTAL ORDER

The two approaches will be compared by a simulation study for the total order restriction according to the dimension *k* (Table 15), the sample size n (Table 16), the spontaneous rate p_0 (Table 17), several shapes (Table 18). The approaches are only defined for significant global tests, i.e. with power > 0.50. Therefore, non-significant simulation results are marked italic.

Class_naive	.913	.823	.773	.657	.575	.524	.476
Class_boot	.971	.901	.801	.749	.673	.608	.561
Power	.827	.853	.883	.889	.885	.892	.903
k	3	4	5	6	7	8	9

Table 15: Classification rates and power for several dimensions *k* (n_j=100, alternative: 0.01/0.01/…/0.01/0.07)

As expected the classification rates decrease with increasing dimension *k* of the design for a selected shape, although the power increases slightly simultaneously (see Table 15).

Class_naive	.739	.793	.799	.823	.817	.856
Class_boot	.817	.912	.901	.901	.875	.903
Power	.394	.614	.787	.853	.926	.952
n_j	25	50	75	100	125	150

Table 16: Classification rates and power for several balanced sample sizes n_j (Alternative: 0.01/0.01/0.01/0.07)

With decreasing sample sizes the classification rates decrease to some extent, but the decrease is much smaller than the related power decrease (see Table 16).

Model of the alternative	Power	Class_naiv	Classs_boot
M1 .01/.07/.07/.07	.742	M1: 0.507	M1: 0.522
M2 .01/0.04/0.07/0.07	.753	M3: 0.311	M3: 0.395
M3 .01/.01/.07/.07	.915	M3: 0.764	M3: 0.787
M4 .01/.04/.04/.07	.671	M4. 0.397	M4: 0.378
M5 .01/.01/.04/.07	.840	M4: 0.482	M5: 0.375
M6 .01/.01/.01/.07	.853	M6: 0.823	M6: 0.901
M7 .01/0.03/0.05/.07	.712	M7: 0.263	M7: 0.331

Table 17 Best model, classification rates and power (n_j=100, k=3)

From Table 16 can be shown, that alternatives with only one inequality are detected precisely (more details see Chapter 6), while the incorrect classification for alternatives with more inequalities increases. With higher dimensions *k* this incorrect classification increases (Table 18) for the alternatives with more than one inequality.

Class_naive	0.480	0.397	0.364	0.327
Power	0.695	0.671	0.634	0.585
k	3	4	5	6

Table 18: Classification rates and power for dimensions *k* for model M4 (n_j=100)

Class_naïve	.823	.655	*.510*	*.478*	*.435*	*.404*	*.422*	*.405*
Class_boot	.901	.709	*.551*	*.528*	*.522*	*.496*	*.512*	*.489*
Power	.853	.697	*.386*	*.310*	*.222*	*.185*	*.183*	*.175*
p_1	.01	.06	.11	.16	.20	.30	.40	.50

Table 19: Classification rates and power for several spontaneous rates π_1 (Alternative: $\{\pi_1, \pi_1, \pi_1, \pi_1, \pi_1 + \Delta = 0.06\}$, n_j=100)

Analogously to the power decrease, a serious decrease of classification rate with increasing spontaneous rate p_1 can be seen in Table 19.

Class_naïve	*.671*	.745	.823	.849	.918	.925	.953
Class_boot	*.822*	.841	.901	.908	.948	.942	.963
Power	*.325*	.659	.853	.961	.985	.996	.998
non-centrality Δ	0.03	0.05	0.07	0.09	0.11	0.13	0.15

Table 20: Classification rates and power for several non-centralities Δ (Alternative: 0.01/0.01/0.01/Δ, n_j=100)

The increase of classification rates with increasing non-centrality Δ is much steeper than the related power increase (see Table 20)

From these simulation can be concluded,

- The model classification is possible for one-inequality alternatives with rather high classification rate, particularly for the parametric bootstrap approach.

- For alternatives with several inequalities acceptable classification rates are only given for the design k=3.
- The classification rates increase with smaller dimension *k*, higher sample sizes n_i, smaller spontaneous rate π_1, higher non-centralities Δ and for more concave shapes but not in a direct relationship to power .

5.2 INFORMATION CRITERION FOR ORDER RESTRICTION

Anraku (1999)[12] published a model selection approach based on the information criterion for order restriction for normal distributed variables. The Akaike (1974)[31] criterion for the unrestricted maximum likelihood estimator $\hat{\theta}$: $AIC(\hat{\theta}) = l(\hat{\theta}) - p$ (with $l(\hat{\theta})$ log-likelihood, p dimension of θ) was modified for order-restricted maximum likelihood estimators:

$$ORIC(\tilde{\theta}) = l(\tilde{\theta}) - penalty(k, n_i) .$$

The penalty term is calculated for each model using the level probabilities under order restriction according to Robertson et al (1988)[28]. The models are the null-hypothesis and all elementary alternative hypotheses as described above. In a simulation study for normal variables the correct model classification rates were about 50% for convex, concave, step and linear shapes. For the unbalanced design with an unexposed and two exposure groups in the binomial case the ORIC is as follows (Anraku, 2003)[32] .
The four models are:

$M_0\{H_0 : \pi_1 = \pi_2 = \pi_3\}$, $M_1\{H_{M_1} : \pi_1 = \pi_2 < \pi_3\}$, $M_2\{H_{M_2} : \pi_1 < \pi_2 = \pi_3\}$, $M_3\{H_{M_3} : \pi_1 < \pi_2 < \pi_3\}$

The likelihood is $L(\pi) = \frac{n_1!}{x_1!(n_1 - x_1)!}\pi_1^{x_1}(1-\pi_1)^{n_1-x_1}\frac{n_2!}{x_2!(n_2 - x_2)!}\pi_2^{x_2}(1-\pi_2)^{n_2-x_2}\frac{n_3!}{x_3!(n_3 - x_3)!}\pi_3^{x_3}(1-\pi_3)^{n_3-x_3}$

With the expected values π_j, the crude estimators

$\hat{\pi}_1 = x_1/n_1, \hat{\pi}_2 = x_2/n_2, \hat{\pi}_3 = x_3/n_3$ $(n_{11} = x_1, n_{12} = x_2, n_{13} = x_3)$.

The $\tilde{\pi}_j$ are the maximum likelihood estimates under the simple order restriction: $\tilde{\pi}_j = \min_{t:t\geq j}\max_{s:s\leq i}\frac{\sum_{j=s}^{t} w_j\pi_j}{\sum_{s=j}^{t} w_j}$. The likelihood for the null-model M_0 is:

$$L(\hat{\pi}_{H_0}) = \prod_{j=1}^{3}\frac{n_j!}{x_j!(n_j - x_j)!}\hat{p}_{H_0}^{x_1+x_2+x_3}(1-\hat{p}_{H_0})^{n_1+n_2+n_3-(x_1+x_2+x_3)} \text{ where}$$

$$\hat{p}_{H_0} = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3} = \frac{w_1\pi_1 + w_2\pi_2 + w_3\pi_3}{w_1 + w_2 + w_3} \text{ provided } w_j = n_j$$

The likelihood for the null-model M_1 is:

$$L(\hat{\pi}_{M_1}) = \prod_{j=1}^{3}\frac{n_j!}{x_j!(n_j - x_j)!}\tilde{\pi}_{(12)}^{x_1+x_2}(1-\tilde{\pi}_{(12)})^{n_1+n_2-(x_1+x_2)}\tilde{\pi}_3^{x_3}(1-\tilde{\pi}_3)^{n_3-x_3}$$

where $\hat{\pi}_{(12)} = \frac{x_1 + x_2}{n_1 + n_2}, \hat{\pi}_3 = x_3/n_3$ $if : \hat{\pi}_{(12)} < \hat{\pi}_3 \Rightarrow \tilde{\pi}_{12} = \tilde{\pi}_{(12)}, \tilde{\pi}_3 = \hat{\pi}_3$

$$if: \hat{\pi}_{(12)} \geq \hat{\pi}_3 \Rightarrow \tilde{\pi}_{12} = \tilde{\pi}_3 = \frac{w_{(12)}\hat{\pi}_{(12)} + w_3\hat{\pi}_3}{w_{(12)} + w_3} = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3}$$

The likelihood for the null-model M_2 is:

$$L(\hat{\pi}_{M_2}) = \prod_{j=1}^{3} \frac{n_j!}{x_j!(n_j - x_j)!} \tilde{\pi}_{(23)}^{x_2+x_3} (1-\tilde{\pi}_{(23)})^{n_2+n_3-(x_2+x_3)} \tilde{\pi}_1^{x_1} (1-\tilde{\pi}_1)^{n_1-x_1}$$

where $\hat{\pi}_{(23)} = \frac{x_2 + x_3}{n_2 + n_3}$ $\hat{\pi}_1 = x_1 / n_1$ $if: \hat{\pi}_1 < \hat{\pi}_{(23)} \Rightarrow \tilde{\pi}_1 = \hat{\pi}_1, \tilde{\pi}_{(23)} = \hat{\pi}_{(23)}$

$$if: \hat{\pi}_1 \geq \hat{\pi}_{(23)} \Rightarrow \tilde{\pi}_1 = \tilde{\pi}_{(23)} = \frac{w_1\hat{\pi}_1 + w_{(23)}\hat{\pi}_{(23)}}{w_1 + w_{(23)}} = \frac{x_1 + x_2 + x_3}{n_1 + n_2 + n_3}$$

The likelihood for the null-model M_3 is:

$$L(\hat{\pi}_{M_3}) = \prod_{j=1}^{3} \frac{n_j!}{x_j!(n_j - x_j)!} \pi_1^{x_1} (1-\pi_1)^{n_1-x_1} \pi_2^{x_2} (1-\pi_2)^{n_2-x_2} \pi_3^{x_3} (1-\pi_3)^{n_3-x_3}$$

The model-specific ORIC are:

$$ORIC(M_r) = \log L(\tilde{\pi}_M) - penalty(M_r).$$

Where the penalty terms are $penalty(M_r) = \sum_{i=1} iP\{i, j, w(M_r)\}$

With $w(H) = n_{1.} + n_{2.} + n_{3.}$ $w(M_1) = n_{1.} + n_{2.}, n_{3.}$ $w(M_2) = n_{1.}, n_{2.} + n_{3.}$ $w(H) = n_{1.}, n_{2.}, n_{3.}$

Because $P\{1,1,w(H)\} = 1$ $ORIC(H) = l(\tilde{\pi}_H) - 1$

Because $P\{1,2,w(M_1)\} = \frac{1}{2}$ $P\{2,2,w(M_1)\} = \frac{1}{2}$ $ORIC(M_1) = l(\tilde{\pi}_{M_1}) - \frac{3}{2}$

Because $P\{1,2,w(M_2)\} = \frac{1}{2}$ $P\{2,2,w(M_2)\} = \frac{1}{2}$ $ORIC(M_2) = l(\tilde{\pi}_{M_2}) - \frac{3}{2}$

Because $P\{1,3,w(M_3)\} = \frac{1}{4} - \frac{1}{2\pi}\sin^{-1}\rho$, $P\{2,3,w(M_3)\} = \frac{1}{2}$, $P\{3,3,w(M_3)\} = \frac{1}{4} + \frac{1}{2\pi}\arcsin\rho$,

and $\rho_{12} = -\sqrt{\frac{n_1 n_3}{(n_1 + n_2)(n_2 + n_3)}}$ $ORIC(M_3) = l(\hat{\pi}_{M_3}) - (2 + \frac{1}{\pi}\arcsin\rho_{12})$

Recently Zhao and Peng (2002)[33] published a modification of Anraku`s approach which converges better for small sample sizes. Because here large sample sizes are assumed throughout, this will be not investigated. Also recently Xiong and El Barmi (2002)[34] proposed a modified AIC criterion for the change point problem including a test between different model-specific AICs.

5.2.1 SIMULATION STUDY

For the unbalanced design with 2 exposure groups in a simulation study the correct classification rates of both approaches were compared for the change-point alternative $H_1^1: \pi_0 = \pi_1 < \pi_2$, $H_1^2: \pi_0 < \pi_1 = \pi_2$ and the simple order alternative (described in chapter 5.2).

	ORIC(M_0,M_1,M_2)			ORIC(M_0,M_1,M_2,M_3)			
π_i	M_0	M_1	M_2	M_0	M_1	M_2	M_3
0.3/0.3/0.3	*.758*	*.112*	*.129*	*.532*	*.080*	*.093*	*.295*
0.1/0.1/0.3	.001	**.979**	.021	.000	**.588**	.410	.002
0.1/0.3/0.3	.001	.020	**.980**	.001	.408	**.588**	.408
0.1/0.2/0.3				.002	.109	.174	**.715**
0.2/0.2/0.4	.002	**.958**	.041	.005	**.578**	.004	.414
0.2/0.4/0.4	.005	.029	**.967**	.003	.008	**.535**	.436
0.2/0.3/0.4				.0095	.184	.232	**.575**
0.3/0.3/0.5	.006	**.940**	.054	.004	**.582**	.014	.401
0.3/0.5/0.5	.004	.053	**.943**	.007	.011	**.566**	.416
0.3/0.4/0.5				.018	.221	.225	**.536**
0.4/0.4/0.6	.009	**.940**	.052	.008	**.557**	.014	.422
0.4/0.6/0.6	.009	.053	**.940**	.004	.015	**.555**	.427
0.4/0.5/0.6				.021	.221	.223	**.536**

Table 21: Model selection rate for the ORIC approach for several π_1 (n_i=100).

	ORIC(M_0,M_1,M_2)			ORIC(M_0,M_1,M_2,M_3)			
π_i	M_0	M_1	M_2	M_0	M_1	M_2	M_3
0.3/0.3/0.35	.465	**.362**	.174	.386	**.254**	.132	.229
0.3/0.3/0.4	.179	**.686**	.134	.156	**.458**	.102	.285
0.3/0.3/0.45	.042	**.867**	.092	.039	**.565**	.049	.348
0.3/0.3/0.5	.008	**.947**	.045	.0045	**.588**	.0125	.395
0.3/0.3/0.55	.0005	**.978**	.022	.000	**.593**	.0025	.405
0.3/0.3/0.6	.000	**.991**	.009	.000	**.595**	.000	.405

Table 22: Model selection rate for the ORIC approach for several non-centrality parameters Δ

		ORIC(M_0,M_1,M_2)			ORIC(M_0,M_1,M_2,M_3)			
π_i	n_i	M_0	M_1	M_2	M_0	M_1	M_2	M_3
0.3/0.3/0.5	150	.0010	**.975**	.025	.0000	**.596**	.001	.403
0.3/0.5/0.5	150	.0005	.021	**.979**	.0000	.0025	**.601**	.400
0.3/0.4/0.5	150				.004	.153	.175	**.669**
0.3/0.3/0.5	125	.001	**.972**	.028	.000	**.565**	.007	.428
0.3/0.5/0.5	125	.002	.026	**.972**	.0015	.005	**.596**	.398
0.3/0.4/0.5	125				.006	.193	.206	**.596**
0.3/0.3/0.5	100	.006	**.940**	.054	.004	**.582**	.014	.401
0.3/0.5/0.5	100	.004	.053	**.943**	.007	.011	**.566**	.416
0.3/0.4/0.5	100				.018	.221	.225	**.536**
0.3/0.3/0.5	75	.024	**.908**	.069	.012	**.574**	.025	.39
0.3/0.5/0.5	75	.025	.072	**.903**	.021	.023	**.525**	.432
0.3/0.4/0.5	75				.046	.242	.252	**.461**
0.3/0.3/0.5	50	.067	**.828**	.105	.051	**.517**	.062	.375
0.3/0.5/0.5	50	.064	.100	**.837**	.052	.055	**.574**	.320
0.3/0.4/0.5	50				.080	.284	.324	**.314**
0.3/0.3/0.5	25	.196	**.644**	.160	.152	**.433**	.100	.316
0.3/0.5/0.5	25	.199	.153	**.649**	.182	.082	**.382**	.355
0.3/0.4/0.5	25				.200	.254	.264	**.283**

Table 23: Model selection rate for the ORIC approach for several sample sizes n_j

π_i	n_i	ORIC(M_0,M_1,M_2)			ORIC(M_0,M_1,M_2,M_3)			
		M_0	M_1	M_2	M_0	M_1	M_2	M_3
0.3/0.3/0.5	100/100/100	.006	.940	.054	0.004	0.582	0.014	0.401
0.3/0.5/0.5	100/100/100	.004	.053	.943	0.007	0.011	0.566	0.416
0.3/0.4/0.5	100/100/100				0.018	0.221	0.225	0.536
0.3/0.3/0.5	133/133/34	.092	.842	.067	.075	.553	.030	.343
0.3/0.5/0.5	133/133/34	.005	.010	.977	.0035	.006	.630	.361
0.3/0.4/0.5	133/133/34				.0480	.145	.362	.446
0.3/0.3/0.5	180/102/18	.195	.681	.124	.156	.483	.078	.285
0.3/0.5/0.5	180/102/18	.006	.024	.971	.0035	.004	.609	.384
0.3/0.4/0.5	180/102/18				.082	.140	.406	.374
0.3/0.3/0.5	92/134/74	.103	.950	.038	.014	.600	.014	.374
0.3/0.5/0.5	92/134/74	.009	.036	.956	.008	.008	.620	.366
0.3/0.4/0.5	92/134/74				.036	.216	.258	.491
0.3/0.3/0.5	34/133/133	.005	.972	.024	.003	.613	.004	.381
0.3/0.5/0.5	34/133/133	.078	.076	.847	.055	.036	.587	.322
0.3/0.4/0.5	34/133/133				.061	.321	.175	.444

Table 24: Model selection rate for the ORIC approach for an unbalanced design

For the unbalanced design with three groups in a simulation study the correct classification rates of both approaches were compared for the two scenarios ORIC(M_0,M_1,M_2) and ORIC(M_0,M_1,M_2,M_3) . From the first row in Table 21 where no differences between the proportions is investigated, the principle difference to the best contrast approach can be seen. The ORIC approach does not control α , i.e. under the null-hypothesis only in 76 respective 53% cases M_0 will be selected, not 95%. The false selection rates are for the recessive and dominant model similar, while for the additive model much larger. The difference of selection rates between the two scenarios is substantial, while the difference between the three models is small. The influence of different non-centrality parameters Δ for the convex alternative $[\pi_0, \pi_0, \pi_0 + \Delta]$ is quite different for the both scenarios. For ORIC(M_0,M_1,M_2) is a strong increase of the selection rate, but for ORIC(M_0,M_1,M_2,M_3) there is a threshold and the false selection is concentrated in the additive model (see Table 22). With increasing sample size the model selection rates increases (Table 23), whereas for the linear alternative a much stronger increase can be observed. Four patterns of unbalanced sample sizes were used in Table 24, each adjusted to a total sample size of 300. When the informative group in the convex alternative has the lowest sample size a serious decrease in correct model selection occur. For these patterns the model selection rate of the linear alternative decreases substantially and even misclassifications are possible.

Alone from the simulation of the null-hypothesis the differences between both approaches becomes clear: the best contrast approach estimates α as size (power under H_0) and for this about 5% each model reveals the same classification rate. The ORIC approach identifies in about 76% the null-model and the alternative models with about 11 and 13%. Both approaches differ substantially: the parametric bootstrap is a testing approach with conditional characterization of the alternative, ORIC approach is a model estimation approach without control of the false positive error rate α. A direct comparison is not possible.

Convex and concave shapes will be classified similar by the ORIC-approach, but different by the best contrast approach reflecting the impact of power (impact of the different variance estimators). Analogously the classification rates increases with decreasing spontaneous rates p_0 and the sample size, particularly in the best contrast approach. For the typically unbalanced designs in epidemiology a serious decrease in classification rates occurs only if small sample sizes occur in informative groups.

Linear, convex and concave shapes will be classified similar by the ORIC-approach. These effects are analogously to the normal distributed experiment in Anraku's (1999)[12] simulation study.

6. ESTIMATION A CHANGE POINT
6.1 TESTING AND CLASSIFICATION APPROACH

A special case of order restricted inference is the consideration of step shapes only, the most special case is the identification of exactly one step – denoted as change-point problem according to Hirotsu and Marumo (2002)[34] Pastor-Barriuso et al.(2003)[35]. For example in a diabetes study with the relationship between 2-hour plasma glucose and mortality the following questions were formulated [35]: exists a certain glucose level at which the mortality risk increases markedly? Can this change-point estimated? What is the shape of the risk relationship above, below and around this change point? This chapter proposes a simple solution for the first and second question, and demonstrates the difficulties in answering the last question.

In chapter 4 approaches considering a trend as any monotone increasing pattern of the dose-response. Here we discuss a simpler approach: the assumption that a dose-response can be characterize by a lower part, an upper part and an abrupt change between both, only. This seems to be rather naïve. But we will demonstrate that this change point problem is a substantial member of the all-pattern trend problem. Moreover, in some epidemiological problems exactly this question arises, like in the diabetes example, and some practical data can be appropriately analyzed. Proposals in the literature are directed on the proof of the existence of such a change point only. But epidemiologist want not simply to know that such a change exist, they want furthermore information where this change is located. Here we demonstrate the estimation of the change point q characterized by its correct classification rate by means of multiple contrast tests, i.e. in a testing framework.
The hypotheses system for a change from q to q+1 is:

$$H_0 : \pi_1 = \pi_2 = ... = \pi_k$$

$$H_1 : \pi_1 = ... = \pi_q < \pi_{q+1} = ... = \pi_k \qquad q \in (1,...,k-1)$$

In epidemiology the exposure groups are not too frequently characterized by an exact amount of exposure, commonly only a range is available, e.g. 1-2 cups coffee per day, or 50-100 µg/L arsenic in drinking water. If exact doses are available the change point can be estimated using the hypothesis of sigmoidicity according to Hirotsu and Marumo (2002)[34]. The above hypotheses system can be tested by multiple step contrasts. Exactly (k-1) step contrasts are appropriate for testing the above hypothesis:

$$
\begin{array}{lllll}
(-k, & 1, & 1, & \ldots & 1) \\
(-(k-1), & -(k-1), & 2, & \ldots. & 2) \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
(-1, & -1, & -1, & \ldots & k)
\end{array}
$$

This principle was introduced as cumulative chi-square method Hirotsu (1982)[36]. For the simple design with an unexposed and three exposure groups exactly three possible change points q exists and for each change point exactly one contrast is optimal:

q	c_1	c_2	c_3	c_4
1	(–3	1	1	1)
2	(–2	–2	2	2)
3	(–1	–1	–1	3)

Optimal means simply the maximum test statistics because the t_{SC}^{i} itself are normal distributed and hence standardized, and t_{MC} is q-variate normal distributed. The contrast coefficients c_{ij} are defined for the general unbalanced design to Hirotsu et al. (1992)[37]:

$$
c_{qj} = \begin{cases} -n_j / \sum_{l=1}^{j} n_l & if \quad j = 1,\ldots,q \\ n_j / \sum_{l=j}^{k} n_l & if \quad j = q+1,\ldots,k \end{cases}
$$

These step contrasts reveal a nice property to transform the k-sample problem into an unbalanced two-sample problem, which can be used later for estimation the unadjusted relative risk (or odds ratio) "above/below" the change point Moreover, the step contrasts belong to a broader class of multiple contrasts. Isotonic contrasts approximate quite well the power of the likelihood ratio test for total ordered hypothesis, while the bivariate up/down-proposals (Neuhauser and Hothorn (1997)[38], Stewart and Ruberg (2000)[39]) use only the two extreme contrasts. This can be simply explained for a balanced design with four groups.

Type of contrasts	Alternative	Contrast c_i	No. of contrasts
Isotonic	$\pi_1 < \pi_2 = \pi_3 = \pi_4$	{-3 1 1 1}	2^{k-1}-1
	$\pi_1 = \pi_2 < \pi_3 = \pi_4$	{-1 -1 1 1}	
	$\pi_1 = \pi_2 = \pi_3 < \pi_4$	{-1 -1 -1 3}	
	$\pi_1 < \pi_2 < \pi_3 < \pi_4$	{-3 -1 1 3}	
	$\pi_1 = \pi_2 < \pi_3 < \pi_4$	{-1 -1 0 2}	
	$\pi_1 < \pi_2 = \pi_3 < \pi_4$	{-1 0 0 1}	
	$\pi_1 < \pi_2 < \pi_3 = \pi_4$	{-2 0 1 1}	
Step	$\pi_1 < \pi_2 = \pi_3 = \pi_4$	{-3 1 1 1}	k-1
	$\pi_1 = \pi_2 < \pi_3 = \pi_4$	{-1 -1 1 1}	
	$\pi_1 = \pi_2 = \pi_3 < \pi_4$	{-1 -1 -1 3}	
Up/down	$\pi_1 < \pi_2 = \pi_3 = \pi_4$	*{-3 1 1 1}*	2
	$\pi_1 = \pi_2 = \pi_3 < \pi_4$	*{-1 -1 -1 3}*	

Under this view multiple step contrasts represent a good compromise even for trend testing in general, much less dependence of the power on the shape compared with single linear contrast (frequently used) or up/down, still not the full information compared with isotonic contrasts, but with *k* instead of 2^{k-1}-1 contrasts.

For normal populations Worsley (1979)[40] proposed a likelihood ratio test taking the maximum over quadratic forms of step-contrasts. However, the distribution under the null-hypothesis is rather complicated. A related approach is the reduced monotonic regression approach by Schell and Singh (1987)[41] which allows multiple change points. A Bayesian version for binary data and multiple prognostic factors was recently published by Holmes and Heard (2003)[42]. If at least one –anyone- contrast is significant, a global decision against the null-hypothesis is possible. But this is only a global decision; no information on the change point q is available directly.

6.2 SIMULATIONS

The two approaches will be compared by a simulation study for the total order restriction according to the dimension *k* (Table 25, the sample size n (Table 26), the spontaneous rate π_0 (Table 27), several shapes (Table 28).

Class_naive	.985	.984	.966	.961	.957	.960
Class_boot	.992	.987	.977	.971	.971	.964
Power	.828	.845	.861	.899	.889	.894
k	3	4	5	6	7	8

Table 25: Classification rates and power for several dimensions *k* (n_j=100, alternative: 0.01/0.01/…/0.01/0.07)

As expected the classification rates decrease with increasing dimension k, although the power increases slightly simultaneously (see Table 25). The decrease is much weaker compared with that of the total order alternative.

Class_naive	*.808*	*.964*	*.972*	*.984*	*.986*
Class_boot	*.809*	*.973*	*.978*	*.987*	*.989*
Power	*.393*	*.618*	*.742*	*.845*	*.903*
n_j	*25*	*50*	*75*	*100*	*125*

Table 26: Classification rates and power for several balanced sample sizes n_j (Alternative: 0.01/0.01/0.01/0.07)

With decreasing sample sizes the classification rates decrease only slightly as long the power is > 0.50, and the decrease is much smaller than the power decrease (see Table 26).

Class_naïve	.984	*.884*	*.807*	*,759*	*,746*	*,681*	*,675*
Class_boot	.987	*.903*	*.817*	*.767*	*.766*	*.679*	*.682*
Power	.845	*.488*	*.373*	*.312*	*.266*	*.226*	*.210*
p_1	.01	.06	.11	.16	.20	.30	.40

Table 27: Classification rates and power for several spontaneous rates p_1 (Alternative: $\{\pi_1, \pi_1, \pi_1, \pi_1, \pi_1 + \Delta = 0.06\}$, n_j=100)

A serious decrease of classification rate with increasing spontaneous rate p_0 can be seen in Table 27. The decrease is much weaker compared with the total order alternative.

Class_naïve	*.942*	.968	.982	.995	.997	.999
Class_boot	*.953*	.973	.985	.994	.998	.999
Power	*.479*	.773	.904	.972	.991	.999
non-centrality Δ	0.03	0.05	0.07	0.09	0.11	0.13

Table 28: Classification rates and power for several non-centralities Δ (Alternative: 0.01/0.01/0.01/Δ, n_j=100)

The increase of classification rates with increasing non-centrality Δ is much steeper than the power increase (see Table 28). For a very large non-centrality Δ the classification is almost sure.

Alternative	k	n_j	Power	Class_naiv	Class_boot
.01/.01/.07	3	100	.818	.981	.991
.01/.07/.07		100	.728	.863	.845
.01/.01/.01/.07	4	100	.855	.973	.978
.01/.01/.07/.07		100	.910	.863	.850
.01/.07/.07/.07		100	.749	.788	.762
.11/.11/.11/.17		100	*.365*	*.816*	*.830*
.11/.11/.17/.17		100	*.442*	*.699*	*.691*
.11/.17/.17/.17		100	*.342*	*.671*	*.659*
.11/.11/.11/.17		150	*.485*	*.852*	*.861*
.11/.11/.17/.17		150	.559	.770	.765
.11/.17/.17/.17		150	*.448*	*.751*	*.737*
.40/.40/.40/.46		800	.839	.922	.924
40/.40/.46/.46		800	.938	.894	.890
40/.46/.46/.46		800	.840	.899	.902
.01/.01/.01/.01/.07	5	100	.865	.970	.977
.01/.01/.01/.07/.07		100	.956	.871	.862
.01/.01/.07/.07/.07		100	.934	.832	.820
.01/.07/.07/.07/.07		100	.721	.747	.730
.01/.01/.01/.01/.01/.07	6	100	.884	.964	.972
.01/.01/.01/.01/.07/.07		100	.971	.861	.847
.01/.01/.01/.07/.07/.07		100	.976	.835	.819
.01/.01/.07/.07/.07/.07		100	.950	.825	.809
.01/.07/.07/.07/.07/.07		100	.676	.729	.711

Table 29: Classification rates and power for several locations of the change point and k=3,4,5

With larger change points *s* the classification rates increases monotonously, but the rates for the parametric bootstrap are better than for the naïve approach only for the largest possible change point. The incorrect classification is asymmetrical in direction of overestimation the change point q, while underestimation is rather small.

Alternative	q	M1	M2	M3	M4	M5	Cum over.	Cum under.
.01/.01/.01/.01/.01/.07	5	.000	.000	.001	.027	.972	-	0.028
.01/.01/.01/.01/.07/.07	4	.000	.002	.012	.847	.139	0.139	0.014
.01/.01/.01/.07/.07/.07	3	.000	.011	.819	.119	.051	0.17	0.011
.01/.01/.07/.07/.07/.07	2	.004	.809	.117	.038	.032	0.187	0.004
.01/.07/.07/.07/.07/.07	1	.711	.135	.052	.050	.053	0.29	-

Table 30: Asymmetrical cumulative false classification rates (n_j=100)

After a significant test on a change point, the estimation of the change point can be performed by model selection. A selection between transition models in logistic regression is recently available (1). The estimation of a change point in likelihood ratio ordering based on a AIC-test, with the disadvantage of a yes/no decision for a change point without information of the correct classification. Multiple change points can be identified by testing the sub-spaces right and/or left of the change point q by conditional testing according to Xiong and Barmi (2002)[19].

In practice the change point definition is relative to the pattern of proportions. In Table 31 the switch from q=3 to q=2 reveals a monotonic increase of the estimation of model 2. This increase is weaker for the switch from q=3 to q=1 according to the asymmetrical effect described in Table 31.

Alternative	Switch	Power	M1	M2	M3
.01/.01/.01/.07	q=4➔3	.862	.000	.021	.979
.01/.01/.02/.07		.809	.001	.114	.885
.01/.01/.03/.07		.809	.001	.268	.731
.01/.01/.04/.07		.820	.005	.454	.541
.01/.01/.05/.07		.851	.004	.629	.367
.01/.01/.06/.07		.885	.002	.770	.229
.01/.01/.07/.07		.907	.004	.843	.153
.01/.01/.01/.07	q=4➔2	.862	.000	.021	.979
.01/.02/.02/.07		.727	.009	.074	.917
.01/.03/.03/.07		.639	.084	.156	.761
.01/.04/.04/.07		.603	.222	.188	.590
.01/.05/.05/.07		.599	.422	.241	.337
.01/.06/.06/.07		.662	.610	.196	.193
.01/.07/.07/.07		.728	.764	.147	.089

Table 31: Model classification rates for the bootstrap approach for switching the change point from q=4 to q=3 respective q=2 (n_j=100)

6.3. EXTREME UNBALANCED EXPOSURE DATA

Particularly in environmental studies, many data for unexposed and low to medium exposed persons exist, but rarely data with high exposure exist - fortunately from an ethical point of view. However, extreme unbalanced 2 by *k* tables results and the statistical outcome depends on the rare high exposed data seriously. Costantino et al. (1995)[17] published a case-control study for respiratory cancer possibly caused by long-term

exposure to coke oven emissions, with a sample size of 10198 in the unexposed group but 487 in the highest exposure group (see Table 2). A more extreme example is the study on child cancer and magnetic fields from high voltage installations where sample size is 2 in the highest exposure group but 6457 in the unexposed group Olsen et al. (1993)[13].(see Table 3). For moderate unbalanced designs the power is not seriously influenced (Bretz and Hothorn (2002)[29], but here we look really on extremely unbalanced designs and the power decreases seriously with such a degree of unbalancedness although the total sample size is constant. Accordingly the correct classification rate decreases. If the total sample size is increased to achieve the same power, the correct classification is in the same magnitude of the balanced case. Interesting is that the correct classification increases if the change point decreases, i.e. a less unbalanced two-sample design results. Epidemiologically is a change point at a high exposure based on rare data very vague, however becomes more stable for medium-to-high exposure based on some more data for theses groups together.

Sample sizes	Shape	power	Class_naiv	Class_boot
200/200/200/200	.05/.05/.05/.10	.682	.930	.935
540/200/40/20	.05/.05/.05/.10	*.251*	*.755*	*.758*
200/200/200/200	.05/.05/.10/.10	.792	.843	.831
540/200/40/20	.05/.05/.10/.10	*.425*	*.698*	*.687*
200/200/200/200	.05/.10/.10/.10	.603	.795	.783
540/200/40/20	.05/.10/.10/.10	.755	.856	.854
400/400/400/400	.05/.05/.05/.10	.915	.965	.971
1340/200/40/20	.05/.05/.05/.10	*.266*	*.749*	*.749*
400/400/400/400	.05/.05/.10/.10	.968	.920	.916
1340/200/40/20	.05/.05/.10/.10	*.438*	*.674*	*.667*
400/400/400/400	.05/.10/.10/.10	.903	.907	.904
1340/200/40/20	.05/.10/.10/.10	.832	.882	.883
9740/200/40/20	.05/.05/.05/.10	*.252*	*.696*	*.702*

Table 32: Model classification rates for the bootstrap approach for extreme unbalanced designs

Unbalanced designs where the smallest sample size occurs at the informative groups (large change point s) reveal a clearly reduced classification rate, but the decrease compared with the balanced design is much weaker than the related power loss. A further reduction occurs for “in-between”-change points as long the sample size of the pooled informative groups is still smaller than in the lower exposure groups because of the interaction between the location of the change point and the sample size pattern. A further substantial increase of the sample size for the unexposed group has nearly no influence on the classification rate.

6.4 ASYMPTOTIC POWER

According to Bretz and Hothorn (2002)[29] the asymptotic power for the change point test is:

$$\text{power} = P\left(T_a^m \geq z_{q,1-\alpha} \middle| H_A\right)$$

$$= P\left(\max_{1\leq l\leq q}\left\{T_{l,a}^s\right\} \geq z_{q,1-\alpha} \middle| H_A\right)$$

$$= 1 - P\left(\max_{1\leq l\leq q}\left\{T_{l,a}^s\right\} < z_{q,1-\alpha} \middle| H_A\right)$$

$$= 1 - P\left(T_{1,a}^s < z_{q,1-\alpha} \wedge \ldots \wedge T_{q,a}^s < z_{q,1-\alpha} \middle| H_A\right)$$

$$\text{power} = P\left(T_a^m \geq z_{q,1-\alpha} \middle| H_A\right)$$

$$= P\left(\max_{1\leq l\leq q}\left\{T_{l,a}^s\right\} \geq z_{q,1-\alpha} \middle| H_A\right)$$

$$= 1 - P\left(\max_{1\leq l\leq q}\left\{T_{l,a}^s\right\} < z_{q,1-\alpha} \middle| H_A\right)$$

$$= 1 - P\left(T_{1,a}^s < z_{q,1-\alpha} \wedge \ldots \wedge T_{q,a}^s < z_{q,1-\alpha} \middle| H_A\right)$$

$$= 1 - P\left(\frac{T_{1,a}^s - E_A\left(T_{1,a}^s\right)}{\sqrt{V_A\left(T_{1,a}^s\right)}} < \frac{z_{1-\alpha} - E_A\left(T_{1,a}^s\right)}{\sqrt{V_A\left(T_{1,a}^s\right)}} \wedge \ldots \wedge \frac{T_{q,a}^s - E_A\left(T_{1,a}^s\right)}{\sqrt{V_A\left(T_{q,a}^s\right)}} < \frac{z_{1-\alpha} - E_A\left(T_{q,a}^s\right)}{\sqrt{V_A\left(T_{q,a}^s\right)}}\right)$$

$$= 1 - \Phi_q\left(\left(\boldsymbol{z}_{q,1-\alpha} - \boldsymbol{e}\right)\text{diag}\left(\tfrac{1}{v_1}, \ldots, \tfrac{1}{v_q}\right); \boldsymbol{0}, \boldsymbol{R}\right).$$

Now, $\boldsymbol{z}_{q,1-\alpha} = \left(z_{q,1-\alpha}, \ldots, z_{q,1-\alpha}\right)$ stands for the q-variate normal $100(1-\alpha)-$equipercentage point under H_0 $e = (E_A(T_{1,a}^s),...,E_A(T_{q,a}^s))$

and $\boldsymbol{v} = \left(v_1, \ldots, v_q\right) = \left(\sqrt{V_A\left(T_{1,a}^s\right)}, \ldots, \sqrt{V_A\left(T_{q,a}^s\right)}\right)$ are the vectorially summarized expectations and variances, as introduced in section 2. The asymptotic correlation matrix **R** derived above can be of a complicated structure.

A R-code makes power calculation for arbitrary sample site pattern, shapes of the dose response and dimensions *k* available. In Figure 2 the power for increasing global sample sizes N was calculated for two rather unbalanced sample size patterns *(Pattern P2 {0.8*N, 0.125*N, 0.05*N, 0.02*N, 0.005*N} and Pattern P1 {0.6*N, 0.2*N, 0.15*N, 0.03*N, 0.02*N})* and two change points (q=4 and 3). The power of the balanced design approaches 1 and the power loss due to unbalancedness is seriously particularly when the change point is to the last exposure group.

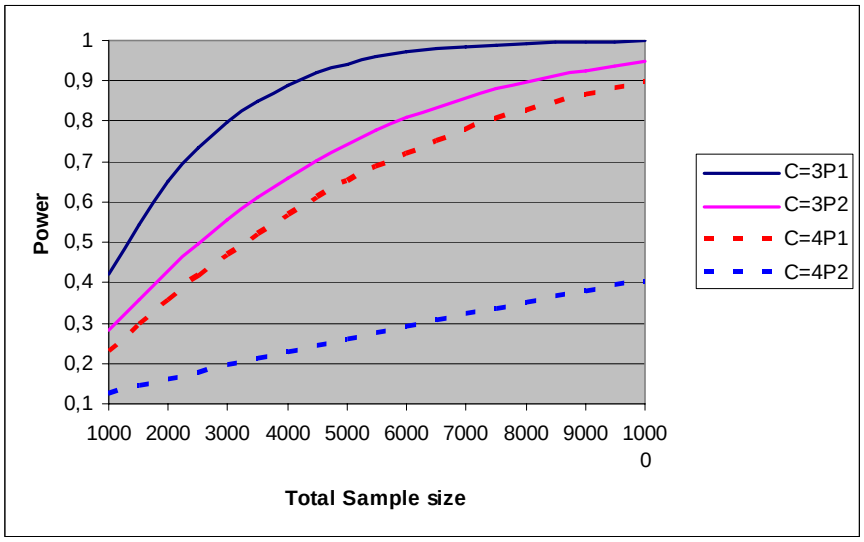

Figure 2: Power for unbalanced designs

6.5 SIMULTANEOUS CONFIDENCE INTERVALS

The general multiple contrast test in chapter 4 can be simplified written fort he change point problem:

$$t_{MC}^{change\ point} = \max_q \left\{ \frac{\bar{p}_q^* - \bar{p}_q}{\sqrt{p(1-p)\sum_{j=1}^{k} c_{qj}^2 / n_j}} \right\} \quad \text{where} \quad \bar{p}_q^* = \frac{\sum_{j=q+1}^{k} n_j p_j}{\sum_{j=q+1}^{k} n_j}, \quad \bar{p}_q = \frac{\sum_{j=1}^{q} n_j p_j}{\sum_{j=1}^{q} n_j}$$

and c_{sj} see above. Because k-1 change points are possible is this test (k-1)-variate normal distributed with correlation matrix R based on the $\rho_{ij} \quad j = 1,...,k, i = j+1,...,k-1$ from chapter 4. For one-sided increasing dose-responses the lower one-sided simultaneous (1-α) confidence intervals are in analogy to Hirotsu and Srivastava (2000)[43]:

$$\pi_q^* - \pi_q \geq (\bar{p}_q^* - \bar{p}_q) - z_{k,R,1-\alpha} \sqrt{p(1-p)\sum_{j=1}^{k} c_{qj}^2 / n_j} \quad , q = 1,...,k-1$$

7. A MODEL BASED FRAMEWORK FOR MULTIPLE CONTRAST TESTS AND CORRECTION FOR ADDITIONAL CONFOUNDING.

7.1 SCORE TEST IN THE LOGISTIC MODEL

A disadvantage of the multiple contrast test in the chapters 1-6 is its limitation on simple one-way layouts. In this chapter the relation between the multiple contrast test and the score test in a logistic model is described to allow correction for additional confounders.
Consider k binomials $y_j \sim Bin(n_j, \pi_j)$, $j = 0,\ldots,k$ and introduce the parametrization

$$\alpha_j = \text{logit}\left(\frac{\pi_j}{1-\pi_j}\right)$$

The following models are relevant for the developments here.

Full model. M: $\boldsymbol{\alpha} \in R^{k+1}$

Isotonic regression. $H_2: \alpha_0 \le \alpha_1 \le \cdots \le \alpha_k$.

Regression on a number of contrast vectors. $H_1: \alpha_j = \alpha + \beta_1 a_j + \beta_2 b_j$

For simplicity we shall here consider two contrast vectors *a* and *b* satisfying $\sum_j a_j n_j = \sum_j b_j n_j = 0$, but the results are not restricted to this situation.

Note also that the isotonic regression model can be specified as a regression model with *k* contrast vectors corresponding to the *k* simple comparisons (group 1,…,*j*-1 versus group *j*,…,*k* for *j* = 2,…,*k*).

Homogeneity. $H_0: \alpha_j = \alpha$

The likelihood function based on the data $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)$ becomes

$$L(\boldsymbol{\pi}) = L(\boldsymbol{\alpha}) = C\left[\prod_j \pi_j^{y_j}\left(1-\pi_j\right)^{y_j}\right] = C\exp\left\{\sum_j \alpha_j y_j - n_j \ln\left(1+e^{\alpha_j}\right)\right\},$$

so the log-likelihood function can be written as

$$l(\boldsymbol{\alpha}) = c + \sum_j \alpha_j y_j - \sum_j n_j \ln\left(1+\exp(\alpha_j)\right).$$

Under H_1 the log-likelihood function reduces to

$$l(\alpha, \beta_1, \beta_2) = c + \alpha T_0 + \beta_1 T_1 + \beta_2 T_2 - \sum_j n_j \ln\left(1-\exp\left\{\alpha + \beta_1 a_j + \beta_2 b_j\right\}\right),$$

with sufficient statistic given by $\boldsymbol{T} = (T_0, T_1, T_2) = \left(\sum y_j, \sum y_j a_j, \sum y_j b_j\right)$

Similarly, under H_0 the likelihood function becomes , $l(\alpha) = c + \alpha T_0 - N\ln\left(1-\exp\{\alpha\}\right)$ where $N = \sum_j n_j$ and the sufficient statistic is simply T_0.

Following Gart & Tarone (1983)[44] hypotheses about the parameters (β_1, β_2) should be assessed in the distribution of the *T* given T_0, equivalently in the distribution (T_1, T_2) given T_0. This is the score statistic based on the conditional likelihood function. Under the hypothesis of homogeneity, $\beta_1 = \beta_2 = 0$, the conditional distribution of (T_1, T_2) given T_0 becomes a multivariate hypergeometric distribution and the mean and the variance of (T_1, T_2) can be obtained immediately as

$$E(T_1) = E(T_2) = 0$$

$$Var(T_1) = \frac{N}{N-1}\sum n_j a_j^2 p(1-p),$$

$$Var(T_2) = \frac{N}{N-1}\sum n_j b_j^2 p(1-p)$$

$$Cov(T_1, T_2) = \frac{N}{N-1} \sum n_j a_j b_j p(1-p),$$

where $p = \frac{T_0}{N} = \frac{\sum y_j}{\sum n_j}$.

Alternatively, one may want to consider the score statistic based on the unconditional likelihood function. This leads to the distribution of (T_1, T_2) under the hypothesis of homogeneity. Mean and variances are then obtained by omitting the finite sample correction in the formulas above. The usual score test is therefore obtained as a χ^2-test of the form $S'V^{-1}S$, where *S* is the vector of score statistics and *V* is the variance-covariance matrix under H_0. Each of components of the score statistic is a single contrast test statistic based on the particular contrast and a global test of the hypothesis of homogeneity versus the regression model can therefore also based on the maximum of the scaled components of the score statistic.

7.2 CORRECTION FOR CATEGORICAL CONFOUNDERS VIA STRATIFICATION

We now turn to a stratified version of the problem above. In each of *m* strata, e.g. defined from the one or several categorical confounders, consider *k* binomials $y_{ij} \sim Bin(n_{ij}, \pi_{ij})$, $i = 1,\ldots m, j = 1,\ldots,k$ and introduce the logit parametrization

$$\alpha_{ij} = \text{logit}\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

Besides the full model in which the logit parameters are allowed to vary freely we consider the following models.

Isotonic regression with each strata. $H_2: \alpha_{11} \le \alpha_{12} \le \cdots \le \alpha_{1K}, \ldots, \alpha_{m1} \le \alpha_{m2} \le \cdots \le \alpha_{mk}$

Regression on a number of contrast vectors within strata. $H_1: \alpha_{ij} = \alpha_i + \beta_1 a_j + \beta_2 b_j$, where we again for simplicity consider two contrast vectors, only.

Homogeneity within strata. $H_0: \alpha_{ij} = \alpha_i$

The log-likelihood function becomes

$$l(\boldsymbol{\alpha}) = c + \sum_i \sum_j \alpha_{ij} y_{ij} - \sum_i \sum_j n_{ij} \ln\left(1 + \exp(\alpha_{ij})\right)$$

and under the regression hypothesis the log-likelihood function reduces to

$$l(\boldsymbol{\alpha}, \beta_1, \beta_2) = c + \sum_i \alpha_i T_{i0} + \beta_1 T_1 + \beta_2 T_2 - \sum_i \sum_j n_{ij} \ln\left(1 - \exp\left\{\alpha_i + \beta_1 a_j + \beta_2 b_j\right\}\right)$$

with sufficient statistic given by

$$\boldsymbol{T} = (T_{10}, \ldots T_{m0}, T_1, T_2) = \left(\sum_j y_{1j}, \ldots, \sum_j y_{mj}, \sum_i \sum_j y_{ij} a_j, \sum_i \sum_j y_{ij} b_j\right)$$

Similarly, under H_0 the likelihood function becomes $l(\boldsymbol{\alpha}) = c + \sum_i \alpha_i T_{i0} - \sum_i N_i \ln\left(1 - \exp\{\alpha_i\}\right)$ where $N_i = \sum_j n_{ij}$ and the sufficient statistic is simply $\boldsymbol{T_0} = (T_{10}, \ldots T_{m0})$.

Again, following Gart & Tarone (1983)[44], hypotheses about the parameters (β_1, β_2) should be assessed in the distribution of the *T* given $\boldsymbol{T_0}$, equivalently in the distribution (T_1, T_2) given $\boldsymbol{T_0}$. Under the hypothesis of homogeneity, $\beta_1 = \beta_2 = 0$, the conditional distribution is a sum of multivariate hypergeometric distributions and the mean and the variance of (T_1, T_2) are therefore obtained as

$$E(T_1) = \sum_i \frac{T_{i0}}{N_i} \sum_j n_{ij} a_j$$

$$E(T_2) = \sum_i \frac{T_{i0}}{N_i} \sum_j n_{ij} b_j$$

$$Var(T_1) = \sum_i \left[\frac{N_i}{N_i - 1} \frac{T_{i0}(N_i - T_{i0})}{N_i} \left\{ \sum_j a_j^2 \frac{n_{ij}}{N_i} - \left(\sum_j a_j \frac{n_{ij}}{N_i} \right)^2 \right\} \right],$$

$$Var(T_2) = \sum_i \left[\frac{N_i}{N_i - 1} \frac{T_{i0}(N_i - T_{i0})}{N_i} \left\{ \sum_j b_j^2 \frac{n_{ij}}{N_i} - \left(\sum_j b_j \frac{n_{ij}}{N_i} \right)^2 \right\} \right]$$

$$Cov(T_1, T_2) = \sum_i \left[\frac{N_i}{N_i - 1} \frac{T_{i0}(N_i - T_{i0})}{N_i} \left\{ \sum_j a_j b_j \frac{n_{ij}}{N_i} - \left(\sum_j a_j \frac{n_{ij}}{N_i} \right) \left(\sum_j b_j \frac{n_{ij}}{N_i} \right) \right\} \right]$$

The usual score test can also here be obtained as a χ^2-test of the form $S' V^{-1} S$, where *S* is the vector of score statistics and *V* is the variance-covariance matrix under H_0. A multiple contrast test of the hypothesis of within stratum homogeneity can constructed from the maximum of the scaled components of the score statistic.

7.3 CORRECTION FOR CONTINUOUS COVARIATES

The same approach can in principle be used to correct for a continuous covariate *z*, but a closed form expression for the (conditional) moments may be difficult, or impossible, to derive. Consider the simplest situation with no strata and a single continuous covariate. The binomial variate y_j is a sum of Bernoulli contributions $y_{lj}, l = 1, ..., n_j$. Let z_{lj} the value of the covariate measured together with y_{lj} then the sufficient statistic for the regression model becomes $T^* = (T_0, T_z, T_1, T_2)$, where $T_z = \sum_{lj} z_{lj} y_{lj}$. The relevant null hypothesis is, however, no longer homogeneity, but $\alpha_{lj} = \alpha + \beta z_{lj}$ and the mean and

variance of the contrasts (T_1, T_2) given (T_0, T_z) is in general complex even if individuals in same group share the same value of the covariate, i.e. if $z_{lj} = z_j$.

8. EXAMPLES

The primary cardiac arrest data in Table 1 reveals at clear change point at q=3 with a classification rate of 86% (the observed power is 0.99), i.e. drinking more than 4.9 cups coffee the day makes problems with PCA. The maximum simultaneous lower confidence limit is for the sub-set [4 vs. {3,2,1}] and with 0.104 medically relevant large. This finding agrees with the estimation of the change point.

Coffee intake in cups	j	$\hat{\pi}_j$	p	CI	
<1/week	1	0.346	<0.0001	{4,3,2} vs.1	-.02
1/week …1.9/day	2	0.322		{4,3} vs.{1,2}	.035
2…4.9/day	3	0.391		4 vs.{3,2,1}	**.104**
≥ 5 /day	4	0.544			

Table 1[modified]: PCA and coffee consumption

The coke oven emission data in Table 2 reveals at weak change point at q=2 with a classification rate of only 43%. Because the observed power is rather high (0.999), a change point model seems to be not appropriate for these data. The maximum simultaneous lower confidence limit is for the sub-set [5 vs. {4,3,2,1}] and with 0.030 medically relevant small. This finding disagrees with the estimation of the change point.

Exposure /mg/m^3-mo	j	$\hat{\pi}_j$	p	CI	
Unexposed	1	0.008		{5,4,3,2} vs.1	*.003*
1-199	2	0.014	<0.0001	{5,4,3} vs.{1,2}	*.012*
200-399	3	0.022		{5,4} vs.{3,2,1}	*.022*
400-599	4	0.030		5 vs.{4,3,2,1}	*.030*
>599	5	0.045			

Table 2[modified]: Respiratory cancer and PAH

The magnet field cancer data in Table 3 reveal a change point q=8 with a classification rate of 0.74 (observed power 0.999). The false classification is nearly concentrated on q=7 (0.26) .
The maximum simultaneous lower confidence limit is for the sub-set [10 vs. {1,2,3,4,5,6,7,8,9}] and with 0.563 medically relevant large. This finding disagrees with the estimation of the change point.

Exposure µTesla.	j	$\hat{\pi}_j$	p	CI	
0-0.05	1	0.263	0.002	{10,9,8,7,6,5,4,3,2} vs.1	-.716
0.051-0.101	2	0		{10,9,8,7,6,5,4,3} vs.{1,2}	-.410
0.101-0.15	3	0.4		{10,9,8,7,6,5,4} vs.{1,2,3}	-.327

0.151-.20	4	0.25		{10,9,8,7,6,5} vs.{1,2,3,4}	-.246
0.201-0.25	5	0.25		{10,9,8,7,6} vs.{1,2,3,4,5}	-.139
0.251-0.30	6	0		{10,9,8,7} vs.{1,2,3,4,5,6}	.108
0.301-0.35	7	0		{10,9,8} vs.{1,2,3,4,5,6,7}	.343
0.351-0.85	8	0.5		{10,9} vs.{1,2,3,4,5,6,7,8}	.534
0.851-1.6	9	1		10 vs.{1,2,3,4,5,6,7,8,9}	.563
>1.61	10	1			

Table 3[modified]: Child cancer and magnetic fields

The hypoglycemia data in Table 4 reveal a change point q=1 with a classification rate of 0.55 already at the lowest level of insulin. Because the observed power is not too high (0.84) this change point model may be relevant, i.e. for insulin levels over 0.25 hypoglycemia is present. The maximum simultaneous lower confidence limit is for the sub-set [{5,4,3,2} vs. 1] and with 0.044 medically not too large. This finding agrees with the estimation of the change point.

Insulin level	j	$\hat{\pi}_j$	p	CI	
<.25	1	0.091	0.018	{5,4,3,2} vs.1	**.044**
0.251-0.49	2	0.221		{5,4,3} vs.{1,2}	.003
0.5-0.74	3	0.322		{5,4} vs.{3,2,1}	-.066
0.75-0.99	4	0.366		5 vs.{4,3,2,1}	-.198
>1	5	0.207		-.149	

Table 4[modified]: Percentage hypoglycemia and daily insulin level

The abortion data in Table 5 reveal a change point q=3 with a classification rate of 0.999 (observed power 0.968). I.e. for the 20-25 years old women a clearly increased risk of abortion exists if the man is older than 34 years. The maximum simultaneous lower confidence limit is for the sub-set [4 vs. {3,2,1}] and with 0.424 medically relevant large. This finding agrees with the estimation of the change point.

Males age	j	$\hat{\pi}_j$	p	CI	
<25	1	0.127	<.0001	{4,3,2} vs.1	-.215
25-29	2	0.103		{4,3} vs.{1,2}	-.063
30-34	3	0.047		4 vs.{3,2,1}	**.424**
35-39	4	0.583			

Table 5[modified]: Abortion rate and male age

The respiratory cancer data in Table 6 reveal a change point q=2 with a classification rate of 0.70 (observed power 0.92). The maximum simultaneous lower confidence limit is for the sub-set [{6,5,4,3,2} vs. 1] and with 0.058 medically not too large. This finding disagrees with the estimation of the change point.

Sulfidic nickel exposure	j	$\hat{\pi}_j$	p	CI	
Unexposed	1	0.149	0.0031	{6,5,4,3,2} vs.1	.058
Low	2	0.225		{6,5,4,3} vs.{1,2}	.044

Low-medium	3	0.336		{6,5,4} vs.{3,2,1}	-.019
Medium	4	0.313		{6,5} vs.{4,3,2,1}	-.046
Medium-high	5	0.299		6 vs.{5,4,3,2,1}	-.080
High	6	0.329			

Table 6[modified]: Lung cancer and cumulative exposure to sulfidic nickel

The mental retardation data in Table 7 reveal a clear change point at q=4 with a classification rate of 0.94 (observed power 0.999). I.e. severe mental retardation occurs only at rather high doses over 0.995 Gy. The maximum simultaneous lower confidence limit is for the sub-set [5 vs. {4,3,2,1}] and with 0.434 medically relevant large. This finding agrees with the estimation of the change point.

Radiation Dose /Gy	j	$\hat{\pi}_j$	p	CI	
<0.005	1	0.008	<0.0001	{5,4,3,2} vs.1	-.025
0.005-0.095	2	0.014		{5,4,3} vs.{1,2}	.017
0.095-0.495	3	0.009		{5,4} vs.{3,2,1}	.203
0.495-0.995	4	0.093		5 vs.{4,3,2,1}	**.434**
>0.995	5	0.462			

Table 7[modified]: Severe mental retardation and organ specific radiation dose

The eye cataract data in Table 8 reveal a clear change point at q=3 with a classification rate of 0.998 (observed power 0.999). I.e. severe eye cataracts occurs at doses over 0.995 Gy, the same change point like for the mental retardation data in Table 7. The maximum simultaneous lower confidence limit is for the sub-set [{7,6} vs. {1,2,3,4,5}] and with 0.496 medically relevant large. This finding disagrees with the estimation of the change point.

Radiation Dose /Gy	j	$\hat{\pi}_j$	p	CI	
<0.005	1	0.000	<0.0001	{7,6,5,4,3,2} vs.1	.082
0.005-0.494	2	0.027		{7,6,5,4,3} vs.{1,2}	.191
0.495-0.994	3	0.088		{7,6,5,4} vs.{1,2,3}	.369
0.995-1.994	4	0.376		{7,6,5} vs.{1,2,3,4}	.488
1.995-2.994	5	0.628		{7,6} vs.{1,2,3,4,5}	.496
2.995-3.994	6	0.711		7 vs.{1,2,3,4,5,6}	.460
>3.995	7	0.690			

Table 8[modified]: Severe cataracts depending on DS86 eye organ dose

9. CONCLUSIONS

In this report the analysis of 2 by *k* tables from epidemiological exposure studies is described. Although a likelihood ratio test is available, multiple contrast tests approximate its power with the advantage of simple distribution under the null- and alternative hypothesis. Not only the global decision for or against a trend is of interest. Information on the particular type of the alternative is welcome. For the multiple contrast test the best

contrast approach can be used for identification of the type of alternative, whereas a parametric bootstrap is suitable for estimation that variability. For the total order alternative at least for dimensions k > 3 miss-identification occurs, particularly for the linear shape. It seems to be more difficult to identify a linear shape compared with step shape (Fairley et al. (1987)[45]). For step alternatives the bootstrapped best contrast behaves quite well for different dimensions, non-centralities, samples sizes, and rate of the unexposed group (due to the asymmetry in binomial testing). The alternative model selection approach using the information criterion under order restriction was investigated. Although both approach are different – e.g. the best contrast approach controls the false positive error rate, while this is not an issue in model selection, both approaches behave quite similar for several conditions for the design k=3.

A close expression for the general unbalanced design is available. The consequences of rather unbalanced designs, i.e. large number in the unexposed or low exposed groups but rare number in the high exposed groups can be calculated dependent on the expected shape. Simultaneous confidence intervals for the step alternative are available, too.
A R-algorithm for calculating the global trend test (p-value), the model identification rates, the power, the confidence interval and the model selection percentages (k=3 only) is available.

Finally the relation between the multiple contrast test and the score test in a logistic model was shown to allow correction for additional confounders.

The suitability of such a simple change point alternative in epidemiological exposure studies should be critically discussed in the near future, although some data examples seem to be appropriate.

10. REFERENCES

1. Hill AB. Principles of medical statistics. Oxford University Press, New York, (1971) 9th ed. pp.309-323.
2. Weed DL, Gorelic LS. The practice of causal inference in cancer epidemiology. *Cancer Epidemiol. Biomarkers Prev.* 1996; **5**:303-311.
3. Rosenbaum PR Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* 2003; **4**: 1-10.
4. Takebayashi T, Nishiwaki Y, Nomiyama T, .Lack of relationship between occupational exposure to carbon disulfide and endocrine dysfunction: A six-year cohort study of the Japanese rayon workers. *J Occup. Health* 2003; **2**: 111-118.
5. Barros-Dios JM, Barreiro MA, Ruano-Ravina A, et al. . Exposure to residential radon and lung cancer in Spain: A population-based case-control study. *Am. J. Epidemiol.* 2002; **156**: 548-555.
6. Royston P, Ambler G, Sauerbrei W The use of fractional polynomials to model continuous risk variables in epidemiology. *Internat. J Epidemiol*. 1999; **28**: 964-974.
7. Leuraud K, Benichou J. A comparison of several methods to test for the existence of a monotonic dose-response relationship in clinical and epidemiological studies. *Statistics in Medicine* 2001; **20**:3335-3351.

8. Pike MC, Ross RK, Spicer DV. Problems involved in including women with simple hysterectomy in epidemiologic studies measuring the effects of hormone replacement therapy on breast cancer risk. *Am J Epidemiol* 1998;**147**:718-721.
9. Baumgartner RN, Koehler KM, Gallagher D, Romero L, Heymsfield SB, Ross RR, Garry PJ, Lindeman RD. Epidemiology of sarcopenia among the elderly in New Mexico. *Am J Epidemiol* 1998; **147**:755-763.
10. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* 1995; **6**:450-454.
11. Weinmann S, Siscovick DS, Raghunathan TE, Arbogast P, Smith H, Bovbjerg VE, Cobb LA, Psaty BM. Caffeine intake in relation to the risk of primary cardiac arrest. *Epidemiology* 1997; **8**: 505-508.
12. Anraku K. An information criterion for parameters under a simple order restriction. *Biometrika* 1999; **86**: 141-152.
13. Olsen JH, Nielsen A, Schulgen G. Residence near high voltage facicilties and risk of cancer in children. *Brit. Medical J* 1993; **307**: 891-895.
14. Neuhäuser, M. and Hothorn, L.A.: An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. *Computat. Stat. & Data Analysis* 1999; **30**: 403-412.
15. Webster J, Piscitelli G, Polli A, D'Alberton A, Falsetti L, Ferrari C, Fioretti P, Giordano G, L'Hermite M, Ciccarelli E et al. Dose-dependent suppression of serum prolactin by cabergoline in hyperprolactinaemia: a placebo controlled, double blind, multicentre study. *Endocrinology* 1992; **37**:534–541.
16. Greenland S A meta-analysis of coffee, myocardial infarction, and coronary death. *Epidemiology* 1994; **4**:366-374.
17. Costantino JP, Redmond CK, Bearden A. Occupationally related cancer risk among coke oven workers: 30 years of follow-up. *J Occup Environ Med* 1995;**37**:597-604.
18. Lausen B, Lerche R, Schumacher M .Maximally selected rank statistics for dose-response problems. *Biometrical J* 2002; **44**: 131-147.
19. Xiong C, El Barmi H On detecting change in likelihood ratio ordering *J Nonparametric Statist*. 2002, **14**: 555-568.
20. Slama R, Werwatz A, Boutou O, Ducot B, Spira A, Hardle W. Does male age affect the risk of spontaneous abortion? An approach using semiparametric regression. *Am. J. Epidemiology* 2003; **157**: 815-824.
21. Grimsrud TK, Berge SR, Haldorsen T, Andersen A. Exposure to different forms of nickel and risk of lung cancer. *Am. J. Epidemiology* 2002*;* **156**: 1123-1132.
22. Otake M, Schull WJ, Lee S . Threshold for radiation-related severe mental retardation in prenatally exposed A-bomb survivors: A re-analysis. *Intern. J Radiat. Biol* 1996; **70**: 755-763.
23. Otake M, Neriishi K, Schull WJ . Cataract in atomic bomb survivors based on a threshold model and the occurrence of severe epilation. *Radiat. Res* 1999; **146**: 339-348.
24. Agresti A, Coull BA .The analysis of contingency tables under inequality constraints. *J Statist Plan Inference* 2002; **107**: 45-73.
25. Peddada SD, Prescott KE, Conaway M. Tests for order restrictions in binary data. *Biometrics* 2001; **57**: 1219-1227.
26. Bretz F, Hothorn LA . Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *ATLA-Alternat. Lab Anim* 2003; **31**: 81-96 Suppl. 1.

27. Salanti G, Ulm K. Tests for trend in binary response. *Biometrical J* 2003; **45**: 277-291.
28. Robertson T., Wright F. T. and Dykstra R. L. (1988). Order restricted statistical inference. New York. Wiley.
29. Bretz, F and Hothorn, L.A. Detecting dose–response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine*. 2002; **21**:3325–3335.
30. Hothorn, L.A., and Bretz, F.: Evaluation of animal carcinogenicity studies. Cochran-Armitage trend test vs. multiple contrast tests. *Biometrical J*. 2000;**42:** 553-567
31. Akaike H A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974;**19**: 716-723.
32. Anraku K Personal communication (2003)
33. Zhao L and Peng L Model selection under order restriction. *Statistics & Probability Letters* 2002; **57**: 301-306
34. Hirotsu C, Marumo K .Changepoint analysis as a method for isotonic inference *Scandinavian J Statist* 2002; **29**: 125-138.
35. Pastor-Barriuso R, Guallar E, Coresh J .Transition models for change-point estimation in logistic regression. *Statistics in Medicine* 2003, **22**: 1141-1162.
36. Hirotsu C Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* 1982; **69**: 567-577.
37. Hirotsu C, Kuriki S, Hayter AJ . Multiple comparison procedures based on the maximal component of the cumulative chi-square statistic. *Biometrika* 1992; **79**: 381-392.
38. Neuhäuser, M. and Hothorn, L.A.: Trend tests for dichotomous endpoints with application in carcinogenicity studies. *Drug Information J*. 1997; **30**: 463-469
39. Stewart WH, Ruberg SJ .Detecting dose response with contrasts. *Statistics in Medicine* 2000; **19**: 913-921.
40. Worsley KJ On the likelihood ratio test for a shift in locations of normal populations. *JASA* 1979; **74**: 365-367.
41. Schell MJ, Singh B The reduced monotonic regression method. *JASA* 1997; **92**: 128-135.
42. Holmes CC, Heard NA Generalized monotonic regression using random change points. *Statistics in Medicine 2003,* 22: 623-638.
43. Hirotsu C, Srivastatva M.S. Simultaneous confidence intervals based an one-sided max t test. *Statistics & Probability Letters* 2000; **49**: 25-37.
44. Gart JJ, Tarone, RE. The relation between score tests and approximate UMPU tests in exponential models common in biometry. *Biometrics* 1983; **39**: 781-786.
45. Fairley D, Pearl DK, Verducci JS. The penalty for assuming that a monotone regression is linear. *Ann. Statist* 1987, **15**: 443-448.

Appendix: Analysis of the data example according to Weinmann et al. (1997)[11] (PCA and coffee consumption) using the R program bindosres

1. Generate an ASCII file

```
library(bindosres)
sink("C:\\temp\\weimann.txt")

###### CALCULATION BEST CONTRAST BOOTSTRAP APPROACH

set.seed(177908751)
NSim <- 5000

## Example weimann
p <- c(0.346, 0.322, 0.391, 0.544)
n <- c(228,149,322,158)
bp <-simulations(p, n, NSim = NSim, status = FALSE, type = "para",
            ctype = "Changepoint", sigonly = TRUE, B=1000, usequant=TRUE)
bp

#######CALCULATION SIMULTANEOUS CONFIDENCE INTERVALS
### example weimann
x <- c(79,48,126,86)
n <- c(228,149,322,158)
mytest <- bindosres.test(x, n)
mytest
confint(mytest)


######CALCULATION POWER FOR A RELATED EXPECTED VALUE AND SAMPLE SIZE PATTERN
p <-c(0.34,0.34,0.4,0.55)
n <-c(230,150,320,160)
power.bindosres(n,p,type=("Changepoint"))

sink()
```

2. Interpretation of the output

```
Simulations with  para  bootstrap
valid random samples:  5000
success rates:
[1] 0.346 0.322 0.391 0.544
sample sizes:
[1] 228 149 322 158
Power:  0.9936                    ← asymptotic power
Naive estimation:
```

```
        1         2         3
0.003824477 0.138687601 0.857487923
Bootstrap estimation:
        1         2         3
0.003623188 0.135466989 0.860909823  ← identification rates


[1]  79  48 126  86
[1] 228 149 322 158

        Asymptotical Test for Dose-Response Relationship using Changepoint
        contrasts

data:
T = 4.2337, p-value = 3.085e-05                    ← p-value
sample estimates:
Estimated dose response relationship :
Contrast number
          3                                ← estimated best contrast
    Dose0      Dose1      Dose2      Dose3
-0.3261803 -0.2131617 -0.4606581  1.0000000

         5 % 0 %
C1  0.10445797 Inf                 ← lower limit CI (in reverese order)
C2  0.03544768 Inf
C3 -0.02190627 Inf
[1] 0.994058
attr(,"error")
[1] 1.447143e-05
attr(,"msg")
[1] "Normal Completion"
```