An overall look at the cumulative chi squared statistics and related methods

L. D'Ambra
University of Naples - Federico II
dambra@unina.it

Outline

- Notation
- Taguchi index
- CCS-type test: Cumulative Chi.square Statistic
- Solution columns equiprobable
- Distribution CCS-type test
- Cumulative Correspondence Analysis
- Likelihood ratio
- Links with Leti's index
- Accumulation Analysis
- Application in several context
- Some applications

Notation

- Let X and Y be categorical variables with $i = 1, \ldots, I$ and $j = 1, \ldots, J$ categories, respectively, where Y is supposed to have an ordinal nature with increasing scores;
- Let $\mathbf{N} = (N_{ij})$ be a $I \times J$ contingency table under the product-multinomial model;
- Let N_{ij} be the random variable which counts the number of observations that fall into the cross-category $i \times j$;
- $N_{i\bullet}$ and $N_{\bullet j}$ represent the counts for the categories i and j, respectively;
- We denote p_{ij} (elements of matrix $\mathbf{P}$) the probability of having an observation fall in the i-th row and j-th column of the table;
- $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ denote the probabilities
- We indicate with F_s the cumulative distribution evaluated in s: $F_Y(s) = \mathbb{P}[Y \leq s] = \sum_{j=1}^{s} p_{\bullet j} = p_{\bullet s}$

The Taguchi's statistic

As a simple alternative to Pearson's test, Taguchi (1966, 1974) proposed a measure of the association which takes into account the presence of an ordinal categorical variable by considering the cumulative sum of cell frequencies across this variable (column)

$$T_E = \sum_{s=1}^{J-1} \frac{1}{D_s(1-D_s)} \sum_{i=1}^{I} \frac{1}{N_{i\bullet}} \left(\frac{Z_{is}}{N_{i\bullet}} - D_s \right)^2$$

with $0 \leq T_E \leq [n(J-1)]$ where

- $Z_{is} = \sum_{j=1}^{s} N_{ij}$ and $Z_{\bullet s} = \sum_{j=1}^{s} N_{\bullet j}$ are the cumulative count and the cumulative column total up to the s-th column category, respectively, with $s = 1, \ldots, J-1$;
- $D_s = Z_{\bullet s}/n$ denotes the cumulative column proportion.

The Taguchi's statistic

- Takeuchi and Hirotsu (1982) shown that this statistic performs better than Pearson's chi-squared statistic when there is an order in the categories on the columns of the contingency table and it is more suitable for studies (such as clinical trials) where the number of categories within a variable is equal to (or larger than) 5;
- In the same paper, the T_E power has been also compared against several score statistics (e.g. the two-sided Wilcoxon test) showing its good power against ordered alternatives.

The Taguchi's statistic

Takeuchi and Hirotsu (1982) and Nair (1986, 1987) showed that the T statistic is linked to the Pearson chi-squared statistic

$$T_E = \sum_{s=1}^{J-1} \chi_s^2$$

where χ_s^2 is Pearson's chi-squared for the $I \times 2$ contingency tables obtained by aggregating the first s column categories and the remaining categories $(s+1)$ to J, respectively.
For this reason, the Taguchi's statistic T_E is also called the **cumulative chi-squared statistic** (hereafter *CCS*).

The Taguchi's statistic

The s-th collapsed $I \times 2$ contingency table:

$Y_{1:s}$	$Y_{(s+1):J}$	Total
Z_{1s}	$N_{1\bullet} - Z_{1s}$	$N_{1\bullet}$
Z_{2s}	$N_{2\bullet} - Z_{2s}$	$N_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
Z_{is}	$N_{i\bullet} - Z_{is}$	$N_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
Z_{Is}	$N_{I\bullet} - Z_{Is}$	$N_{I\bullet}$
$Z_{\bullet s}$	$n - Z_{\bullet s}$	n

$$T_E = \sum_{s=1}^{J-1} \chi^2_s = n \sum_{s=1}^{J-1} \phi_s = n \sum_{s=1}^{J-1} \tau_s$$

τ_s is the Goodman-Kruskal index for the s-th $I \times 2$ contingency table.

$$\chi^2_s = \sum_{i=1}^{I} \frac{(Z_{1s} - \frac{N_{1\bullet} \times Z_{\bullet s}}{n})^2}{\frac{N_{1\bullet} \times Z_{\bullet s}}{n}} + \frac{\{(N_{1\bullet} - Z_{1s}) - \frac{[N_{1\bullet} \times (n - Z_{\bullet s})]}{n}\}^2}{\frac{[N_{1\bullet} \times (n - Z_{\bullet s})]}{n}}$$

The new class of CCS-type tests

Nair (1986, 1987) generalizes T_E by considering the class of *CCS*-type tests

$$T_{CCS} = \sum_{s=1}^{J-1} w_s \left[\sum_{i=1}^{I} N_{i\bullet} \left(\frac{Z_{is}}{N_{i\bullet}} - D_s \right)^2 \right]$$

corresponding to a given set of weights $w_s > 0$. Examples of possible choices for w_s are showed in following table:

w_s	Index		
$1/J$	$T_{CCS} = T_N$	Nair	Nair
$1/[D_s(1-D_s)]$	$T_{CCS} = T_E$	Taguchi	
$p_{\bullet j}$	$T_{CCS} = W_I^2$	Cramer-von Mises	D'Ambra Amenta
$p_{\bullet j}/[D_s(1-D_s)]$	$T_{CCS} = A_I^2$	Anderson-Darling	

$$T_{CCS} = \sum_{s=1}^{J-1} w_s [D_s(1-D_s)] \chi_s^2$$

The distribution of the T_{CCS}

The properties of the *CCS*-type tests have been deeply studied by Nair (1986, 1987) by means a matrix decomposition of this statistic into orthogonal components.

- $\mathbf{A} = \mathbf{M} - (\mathbf{d}_{J-1} \times \mathbf{1}^T)$ is a matrix of dimension $((J-1) \times J)$ where $\mathbf{M}$ is a $(J-1) \times J)$ lower unitriangular matrix and $\mathbf{d}_{J-1} = [D_1, \ldots, D_{J-1}]^T$,
- $\mathbf{n}_i = [N_{i1}, \ldots, N_{iJ}]^T$ is a vector of the counts for the categories i.

The distribution of the T_{CCS}

The *CCS* statistic T_{CCS} can be also written as

$$T_{CCS} = \sum_{i=1}^{I} \frac{\mathbf{n}_i^T \mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{n}_i}{N_{i\bullet}} = \sum_{i=1}^{I} \frac{\mathbf{n}_i^T \mathbf{Q} \mathbf{D}_\lambda \mathbf{Q}^T \mathbf{n}_i}{N_{i\bullet}} = \sum_{s=1}^{J-1} \lambda_s \mathbf{v}_s^T \mathbf{v}_s$$

where

- $\mathbf{W}$ is the $[(J-1) \times (J-1)]$ diagonal matrix of weights w_s;
- λ_s is the s-th non zero eigen value of $\mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{D}_J$;
- $\mathbf{Q}$ are the eigenvectors of matrix $\mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{D}_J$.

and the i-th element of $\mathbf{v}_s$ is

$$v_{is} = \frac{1}{\sqrt{N_{i\bullet}}} \mathbf{q}_s^T \mathbf{n}_i \qquad s = 1, \ldots, J-1 \qquad i = 1, \ldots, I$$

The distribution of the T_{CCS}

The random vector of the summands associated to the i-th row $\mathbf{n}_i$ is

$$\mathbf{v}_i = \frac{1}{\sqrt{N_{i\bullet}}}\mathbf{Q}^T\mathbf{n}_i \qquad i = 1, \ldots, I$$

Given the row and column probabilities, $\mathbf{n}_i$ can be approximated by a multinomial distribution which has a limiting multinormal distribution as $n \to \infty$. Then, since $\mathbf{D}_\lambda$ and $\mathbf{A}$ only depend on the weights and the column probabilities, $\mathbf{q}_i$ also has a limiting multinormal distribution as $n \to \infty$. The expected value for v_{is} is

$$\mathbb{E}[v_{is}] = \frac{1}{\sqrt{N_{i\bullet}}}\mathbf{q}_s^T\mathbb{E}[\mathbf{n}_i] = \sqrt{N_{i\bullet}}\mathbf{q}_s^T(p_{\bullet 1}, \ldots, p_{\bullet J})^T = 0 \quad \begin{matrix} s = 1, \ldots, J-1 \\ i = 1, \ldots, I \end{matrix}$$

since the columns of $\mathbf{Q}$ are orthogonal with respect to the column probabilities.

The distribution of the T_{CCS}

The covariance:

$$\mathbb{C}[v_{is}, v_{i's'}] = \frac{n}{n-1}\left[\mathbf{I}_I - \sqrt{p_{i\bullet}p_{i'\bullet}}\right] \otimes \mathbf{I}_{J-1}$$

The v_{is} are asymptotically i.i.d. with a $N(0,1)$ distribution as $n \to \infty$, $s = 1, \ldots, J-1$, $i = 1, \ldots, I$.
Nair interpreted the first two components $\mathbf{v}_1^2$ and $\mathbf{v}_2^2$ as tests for location and dispersion effects, respectively.

Limiting distribution

The limiting distribution of T_{CCS} is a linear combination of iid chi-squared distributions $\chi^2_{s,(I-1)}$ with $(I-1)$ degrees of freedom (df)

$$T_{CCS} = n \times \sum_{s=1}^{J-1} \lambda_s \left(\sum_{i=1}^{I} \mathbf{v}_s^T \mathbf{v}_s \right) \to^d_{H_0} \sum_{s=1}^{J-1} \lambda_s \chi^2_{s,(I-1)}$$

where $\chi^2_{s,(I-1)}$ is the chi-squared distribution for the s-th component $(s = 1, \ldots, J-1)$ and λ_s are elements of $\mathbf{D}_\lambda$.

The distribution of the T_{CCS}

By using Satterthwaite's two-moment approximation (1946), the asymptotic distribution of T_{CCS} can be then approximated (Nair, 1986, 1987)

$$T_{CCS} \sim d(I-1) \times \chi^2_{(v)}$$

with

- $v = \frac{(I-1)\times(J-1)}{\rho}$ degrees of freedom, with $\rho = \frac{\sum_{s=1}^{J-1} \lambda_s^2}{\sum_{s=1}^{J-1} \lambda_s}$;
- $d = \frac{1}{(I-1)}\rho$;

it proves that $\rho > 1$.

Explicit solutions when the column are equiprobables

Solutions of

$$\mathbf{A}^T\mathbf{W}\mathbf{A} = \mathbf{Q}\mathbf{D}_\lambda\mathbf{Q}^T$$

can be not find numerically. Let consider the special case of equiprobable column categories that is when we have $p_{\bullet j} = 1/J$:

Author	Index	w_s	λ_s
Nair	T_N	$1/J$	$[4J(\sin^2(\frac{s\pi}{2J})]^{-1}$
Taguchi	T_E	$[D_s(1-D_s)]^{-1}$	$J^2[s(s+1)]^{-1}$
Cramer-von Mises	W^2	$p_{\bullet j}$	$[4J(\sin^2(\frac{s\pi}{2J})]^{-1}$
Anderson-Darling	A^2	$p_{\bullet j}[D_s(1-D_s)]^{-1}$	$J[s(s+1)]^{-1}$

Explicit solutions when the column are equiprobables

For $(I \times J)$ with $J > 2$ tables with $w_s = \frac{1}{d_s(1-d_s)}$

- eigenvectors are given by the Chebichev polinomials
- the first squared component (location or linear) is proportional to the Kruskal-Wallis statistic for contingency tables
- the second squared component (dispersion or quadratic) is the generalization of Mood's statistic (1954) for grouped data.

For $(2 \times J)$ tables we have two components:

- The first component (linear) of Taguchi's statistics is equivalent to Wilcoxon statistics
- The second component (quadratic) is equivalent to Mood's test (1954)
- Under the row multinomial model the Nair CSS statistics ($w_s = 1/J$) is decomposed into components where the s-th component can detect cosinusoidal deviations in the s-th moment

T_{CCS} in matrix notation and link with CA

Statistic T_{CCS} can be expressed in matrix notation by

$$T_{CCS} = trace(\mathbf{D}_I^{-\frac{1}{2}}\mathbf{N}\mathbf{A}^T\mathbf{W}\mathbf{A}\mathbf{N}^T\mathbf{D}_I^{-\frac{1}{2}})$$

where:

- **W** diagonal matrix of weights of order $(J-1)\times(J-1)$;
- **A** matrix of order $(J-1)\times J$ involving the cumulative column proportions

$$\mathbf{A} = \begin{bmatrix} 1-D_1 & -D_1 & \dots & -D_1 & -D_1 \\ 1-D_2 & 1-D_2 & \dots & -D_2 & -D_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-D_{J-1} & 1-D_{J-1} & \dots & 1-D_{J-1} & -D_{J-1} \end{bmatrix}$$

T_{CCS} in matrix notation and link with CA

Considering that $\mathbf{A} = \mathbf{M} - (\mathbf{d}_{J-1} \times \mathbf{1}^T)$ and $\mathbf{d}_{J-1} = \mathbf{Mc}$, where

- $\mathbf{M}$ is a $(J-1) \times J$ lower unitriangular matrix;
- $\mathbf{d}_{J-1} = [D_1, \ldots, D_{J-1}]^T$;
- $\mathbf{c}$ and $\mathbf{r}$ vectors of the column and row marginal frequencies of matrix $\mathbf{P}$, respectively

we have:

$$\begin{aligned} T_{CCS} &= n \times trace[\mathbf{D}_I^{-\frac{1}{2}} \mathbf{N} \mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{N}^T \mathbf{D}_I^{-\frac{1}{2}}] \\ &= n \times trace[\mathbf{D}_I^{-\frac{1}{2}} (\mathbf{P} - \mathbf{r}\mathbf{c}^T) \mathbf{M}^T \mathbf{W} \mathbf{M} (\mathbf{P} - \mathbf{r}\mathbf{c}^T)^T \mathbf{D}_I^{-\frac{1}{2}}] \end{aligned}$$

if $\mathbf{M}^T \mathbf{W} \mathbf{M} = \mathbf{D}_J^{-1}$

$$\chi^2 = n \times trace[\mathbf{D}_I^{-\frac{1}{2}} (\mathbf{P} - \mathbf{r}\mathbf{c}^T) \mathbf{D}_J^{-1} (\mathbf{P} - \mathbf{r}\mathbf{c}^T)^T \mathbf{D}_I^{-\frac{1}{2}}] \qquad (CA)$$

T_{CCS} in matrix notation and link with CA

- Taguchi's statistic is also at heart of a cumulative extension of correspondence analysis (TCA) (Beh et al., 2011) when cross-classified variables have an ordered structure;
- TCA is based on an SVD of the centred matrix

$$\mathbf{B} = \mathbf{D}_I^{-\frac{1}{2}}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{M}^T\mathbf{W}^{\frac{1}{2}}$$

- The CCS statistic can be decomposed as a sum of squared singular values μ_i

$$T_{CCS} = n \times ||\mathbf{B}||^2_{D_I} = \sum_{i=1}^{I} \mu_i^2$$

Using different weights w_s we propose a family of Ordinal Correspondence Analysis.

Cumulative Correspondence Analysis and inferential tools

The cumulative extension of correspondence analysis is performed by applying a generalised singular value decomposition (GSVD) to the matrix with different weights system w_s, in particular:

$$GSVD(\mathbf{B})_{D_I;I} \Rightarrow \mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

with: $\mathbf{U}^T\mathbf{D}_I\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$. In particular, the total inertia can be expressed in terms of $\mathbf{B}$ such that

$$\frac{T_{CCS}}{n} = ||\mathbf{B}||^2_{D_I} = trace(\mathbf{B}^T\mathbf{D}_I\mathbf{B}) = \sum_{i=1}^{I}\sum_{s=1}^{J-1} p_{i\bullet} b_{is}^2$$

To visually summarise the association between the row and the column categories, we define the row and column principal coordinates by

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma} \qquad \mathbf{G} = \mathbf{V}\mathbf{\Sigma}$$

Cumulative Correspondence Analysis and inferential tools

The squared Euclidean distance of the s-th **column coordinate** from the origin of the plot is (A. D'Ambra and P.Amenta, 2021):

$$d_s^2(s;0) = w_s D_s(1-D_s)\frac{X_s^2}{n} = \theta_s \frac{X_s^2}{n} \Rightarrow r_s = \sqrt{\frac{\theta_s}{n}\chi^2_{\alpha;2}}$$

The $\frac{T^{KS}_{CCS}}{d(I-1)}$-statistic can be then expressed in terms of the predictor (row) coordinates such that:

$$\frac{1}{d(I-1)}\sum_{i=1}^{I}\sum_{m=1}^{M} N_{i\bullet} f_{im}^2 \sim \chi^2_v \Rightarrow r_i \sim \sqrt{\frac{d(I-1)}{N_{i\bullet}}\chi^2_{\frac{2v}{(I-1)(J-1)}}}$$

By considering these confidence circles, we can identify which (if any) I categories significantly contribute to the dependence structure between the row and aggregated column categories.

A unified framework of CAs coping with ordinal data.

$$SVD[\mathbf{D}_I^{-\frac{1}{2}}\mathbf{L}(\mathbf{P}-\mathbf{D}_I\mathbf{1}\mathbf{1}^T\mathbf{D}_I)\mathbf{A}^{-\frac{1}{2}}\mathbf{R}^T\mathbf{W}^{\frac{1}{2}}] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Method		Row	Column	**L**	**R**	$\mathbf{W} = diag(w_s)$	**A**	Index	Author
CA		N/O	N/O	**I**	**I**	$w_s = 1$	$\mathbf{D}_J$	ϕ^2	1
NSCA		N/O	N/O	**I**	**I**	$w_s = 1$	**I**	$n(\tau)$	2
T_{CCS}	T_E	N/O	O	**I**	**M**	$w_s = \frac{1}{[d_s(1-d_s)]}$	**I**	$\frac{T_E}{n}$	3
	T_N	N-O	O	**I**	**I**	$w_s = \frac{1}{J}$	**I**	$\frac{T_N}{n}$	4
	W	N-O	O	**I**	**I**	$w_s = p_{\bullet j}$	**I**	$\frac{W^2}{n}$	4
	A	N-O	O	**I**	**I**	$w_s = \frac{p_{\bullet j}}{[d_s(1-d_s)]}$	**I**	$\frac{A^2}{n}$	4
DA	H	O	O	**L**	**M**	$w_s = \frac{1}{[d_s(1-d_s)]}$	**I**	$\frac{H}{n}$	5

with **L** and **M** uni-triangular matrix, DA - Double Cumulative.

- 1 - J.P.Benzecrì; 2 - L. D'Ambra - C.N.Lauro;
- 3 - Beh-L. D'Ambra-Simonetti; 4 - A.D'Ambra-P-Amenta-L.D'Ambra;
- 5 - L.D'Ambra-Beh-Cammiantiello

New tools for the interpretation of T_{CCS}

It is possible to view T_{CCS} statistic weighted sum of the Goodman-Kruskal index (1954)

$$T_{CCS} = \sum_{s=1}^{J-1} w_s D_s (1 - D_s) \chi_s^2 = n \sum_{s=1}^{J-1} w_s D_s (1 - D_s) \tau_s$$

where τ_s is the Goodman-Kruskal index for the s-th $I \times 2$ contingency table

- This new formulation of T_{CCS} highlights that it reflects also a unidirectional association between the categorical variables (row versus column), in addition to the symmetrical association structure that is instead required by the Pearson chi-squared statistic

A likelihood ratios interpretation

T_{CCS} can be viewed as a approximate sum of likelihood ratios

- Let L_s be the log-likelihood function (unconstrained model) of the s-th $I \times 2$ table obtained by aggregating the first s column categories and the remaining $(s+1)$ to J

$$L_s = \sum_{i=1}^{I} n_{i\bullet} \left[\frac{z_{is}}{n_{i\bullet}} \ln\left(\frac{p_{is}}{n_{i\bullet}}\right) + \left(1 - \frac{z_{is}}{p_{i\bullet}}\right) \ln\left(1 - \frac{p_{is}}{p_{i\bullet}}\right) \right]$$

- $L_s^{H_0}$ is the log-likelihood function under $H_0 : \frac{p_{is}}{p_{i\bullet}} = p_{\bullet s}$

$$L_s^{H_0} = \sum_{i=1}^{I} n_{i\bullet} \left[\frac{z_{is}}{n_{i\bullet}} \ln(p_{\bullet s}) + \left(1 - \frac{z_{is}}{n_{i\bullet}}\right) \ln(1 - p_{\bullet s}) \right]$$

- If $p_{is}/p_{i\bullet}$ and $p_{\bullet s} = \sum_{i=1}^{I} p_{is}$ are unknown, then they can be replaced by the maximum likelihood estimates $z_{is}/n_{i\bullet}$ and $d_s = z_{\bullet s}/n$, respectively

A likelihood ratios interpretation

- We obtain the following likelihood ratio

$$LR_s = 2\sum_{i=1}^{I} n_{i\bullet}\left[\frac{z_{is}}{n_{i\bullet}}\ln\left(\frac{z_{is}/n_{i\bullet}}{d_s}\right) + \left(1-\frac{z_{is}}{n_{i\bullet}}\right)\ln\left(\frac{1-z_{is}/n_{i\bullet}}{1-d_s}\right)\right]$$

- LR_s and χ^2_s have the same limiting null chi-squared distribution with $(I-1)$ df. In fact, they are asymptotically equivalent and $\chi^2_s - LR_s$ converges in probability to zero.
- This implies that the sum of the LR_s is approximately equivalent to the T_{CCS} statistic

$$T_{CCS} = \sum_{s=1}^{J-1} w_s D_s(1-D_s)\chi^2_s \cong \sum_{s=1}^{J-1} w_s D_s(1-D_s)LR_s$$

A T_{CCS} theoretical framework

It is possible to develop a theoretical framework showing a new link between an unifying index of the heterogeneity, unalikeability and variability measures with the class of CCS-type statistics T_{CCS}

- Let $d_{jj'}$ be a positive quantity reflecting the quantification of the diversity between the categories j and j' of the categorical variable Y;
- Let

$$\bar{D} = \sum_{j=1}^{J} \sum_{j'=1}^{J} d_{jj'} p_{\bullet j} p_{\bullet j'}$$

be the the unifying index of the heterogeneity, unalikeability and variability measures [Leti (1983), Zanella (1989)]

A T_{CCS} theoretical framework

- If we consider nominal categories then $\bar{D}$ amounts to

$$\bar{D} = d\left(1 - \sum_{j=1}^{J} p_{\bullet j}^2\right) \qquad \textit{if } d{=}1 \Rightarrow GIH = 1 - \sum_{j=1}^{J} p_{\bullet j}^2$$

- $\bar{D}$ turns out to be proportional to Gini's index of heterogeneity also known as the Gini–Simpson coefficient or the Gibbs–Martin (or Blau) index
- In the presence of ordered categories, then it's possible to show:

$$\bar{D} = 2d \sum_{s=1}^{J-1} F_s(1 - F_s) \qquad \textit{if } d{=}1 \Rightarrow D^* = 2 \sum_{s=1}^{J-1} F_s(1 - F_s)$$

- We highlight that $\bar{D}$ amounts also to Leti's coefficient of unalikeability (or diversity), introduced as a measure of dispersion of categorical data (Leti, 1983)

Grilli and Rampichini's decomposition

Grilli and Rampichini (2002) decompose the Leti's coefficient of unalikeability $D^*/2$ according to the well-known principle of the between- and within-group variance decomposition of a quantitative variable

$$\begin{aligned}\frac{D^*}{2} &= \sum_{s=1}^{J-1} F_s(1-F_s) \\ &= \underbrace{\sum_{i=1}^{I} p_{i\bullet} \sum_{s=1}^{J-1} F_{s|i}(1-F_{s|i})}_{D_W} + \underbrace{\sum_{i=1}^{I} p_{i\bullet} \sum_{s=1}^{J-1} (F_{s|i}-F_s)^2}_{D_B}\end{aligned}$$

where $F_{s|i} = \mathbb{P}(Y \leq s|X=i) = \sum_{j=1}^{s} p_{ij}/p_{i\bullet} = p_{is}/p_{i\bullet}$ is the cumulative distribution of the conditional variable $(Y|X=i)$ evaluated in s and $F_s = \sum_{i=1}^{I} p_{i\bullet} F_{s|i} = \sum_{i=1}^{I} p_{is} = p_{\bullet s}$.

Decomposition of a "weighted" version of Leti's coefficient

Let 's consider now a "weighted" version of Leti's coefficient of unalikeability

$$\bar{D} = 2\sum_{s=1}^{J-1} w_s F_s(1 - F_s)$$

which subsumes the ordinary Leti's index for $w_s = d$.

We can then decompose the 'weighted' Leti's coefficient $\bar{D}$ according to the well-known principle of between- and within-group variance decomposition of a quantitative variable

Decomposition of a "weighted" version of Leti's coefficient

$$
\begin{aligned}
\frac{\bar{D}}{2} &= \sum_{s=1}^{J-1} w_s F_s(1-F_s) \\
&= \underbrace{\sum_{i=1}^{I} p_{i\bullet} \sum_{s=1}^{J-1} w_s F_{s|i}(1-F_{s|i})}_{D_W} + \underbrace{\sum_{i=1}^{I} p_{i\bullet} \sum_{s=1}^{J-1} w_s (F_{s|i}-F_s)^2}_{D_B} \\
&= D_W + \sum_{j=1}^{J-1} w_s \sum_{i=1}^{I} p_{i\bullet} \left(\frac{p_{is}}{p_{i\bullet}} - F_s \right)^2 = D_W + T_{CCS}
\end{aligned}
$$

The cumulative chi-squared statistic T_{CCS} is then a part of the 'weighted' Leti's coefficient $\bar{D}/2$.

This decomposition subsumes that developed by Grilli and Rampichini for $w_s = 1$, $\forall s$.

Decomposition of a "weighted" version of Leti's coefficient

Interesting properties arise if we realtivized T_{CCS}/n to its maximum (A. D'Ambra and P.Amenta, submitted):

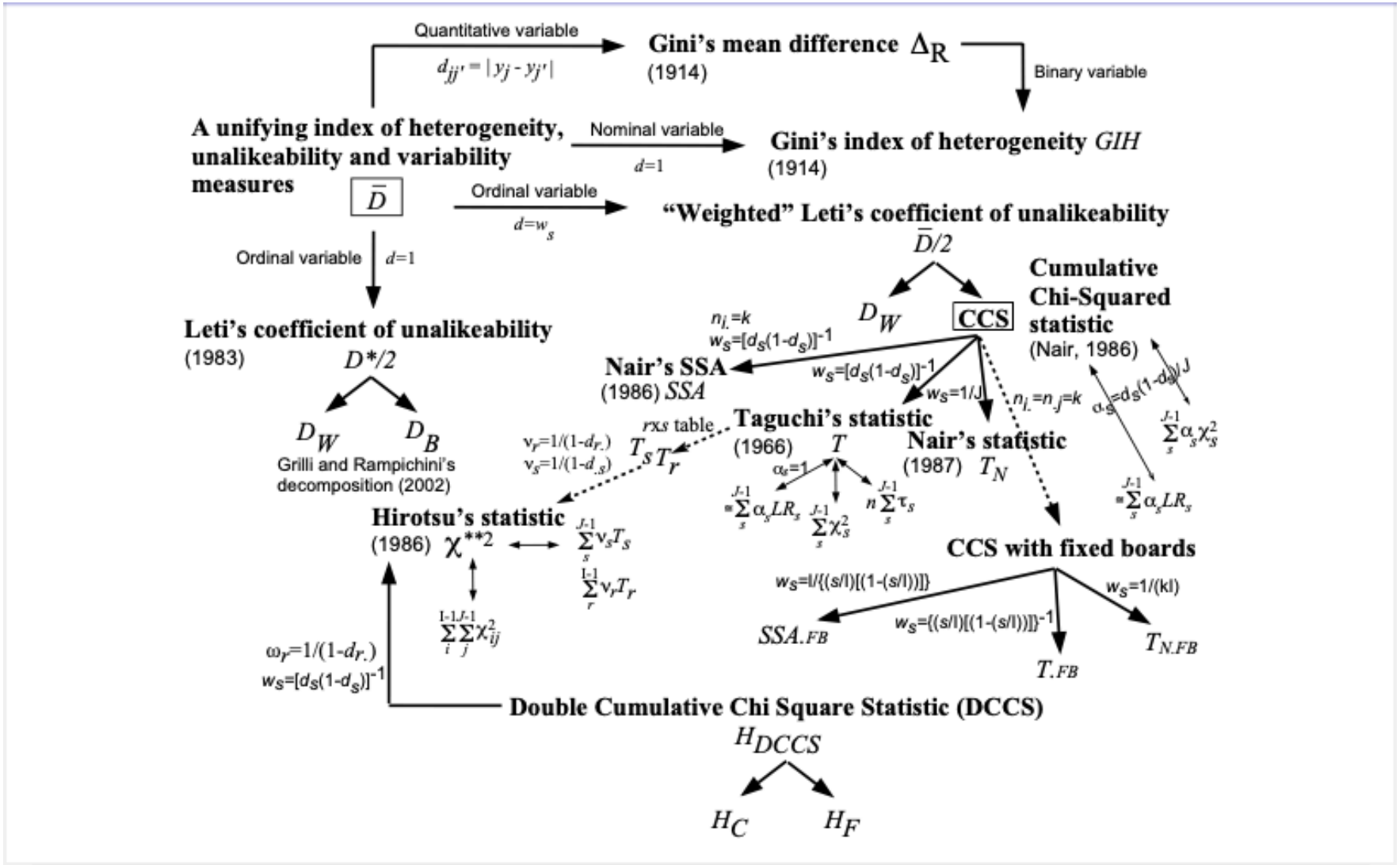
$$\delta = \frac{2T_{CCS}}{n\bar{D}}$$

which can be considered the extension of the Goodman-Kruskal's τ for a criterion categorical variable with an ordinal nature. It's possible to show that δ index is a weighted mean of τ: indeed, we can write

$$\delta = \frac{\sum_{s=1}^{J-1}\{w_s F_s(1-F_s)\}\tau_s}{\sum_{s=1}^{J-1}\{w_s F_s(1-F_s)\}} = \sum_{s=1}^{J-1}\frac{b_s}{\sum_{s=1}^{J-1} b_s}\tau_s = \sum_{s=1}^{J-1} l_s \tau_s$$

with $\sum_{s=1}^{J-1} l_s = 1$. Other proprieties are in (A. D'Ambra and P.Amenta, submitted).

A recapitulative scheme of all indices



Accumulation Analysis (ANOVA for ordinal data)

Accumulation analysis (henceforth abbreviated as AA) is a method proposed by Taguchi (1974) for analyzing ordered categorical data from industrial experiments.
Taguchi's statistic T_E was originally developed to test the hypothesis of homogeneity against monotonicity in the treatment effects within a one-way Anova model in industrial experiments. An $I \times J$ contingency table with row multinomial model with equal row totals ($N_{i\bullet} = k$) observations per level of a factor A with I levels) has been then obtained. For this kind of model, Nair (1986, 1987) shows that the sum of squares for the factor A is given by

$$SSA = n \sum_{s=1}^{J-1} \frac{1}{kD_s(k - kD_s)} \sum_{i=1}^{I} \left(Z_{is} - kD_s \right)^2$$

which is also a special case of the T_{CCS} statistic with fixed and equal row totals.

Accumulation Analysis

Sum Square Total	$SST = nI(J-1)$
Sum Square Error	$SSE = SST - SSA$

To obtain mean squares (MS), AA uses $(I-1)(J-1)$ and $(J-1)I(n-1)$ as degrees of freedom for SSA and SSE, respectively. Since:

$$\mathbb{E}[SSE] = \frac{n(n-1)I^2(J-1)}{(In-1)} \approx (J-1)I(n-1)$$

Finally, A calculates an F-like statistic given by $F_A = MSA/MSE$. Thus, AA is an ANOVA-like procedure. The use of MSE in the denominator of $F_A = MSA/MSE$ is unnecessary. There is really no notion of "error" in this situation, since SSA in

$$SSA = n\sum_{s=1}^{J-1} \frac{1}{kD_s(k-kD_s)} \sum_{i=1}^{I} \left(Z_{is} - kD_s \right)^2$$

has already been standardized by $kD_s(1-D_s)$ (Nair, 1986)

Accumulation Analysis

Ad's simplicity and similarity to ANOVA is appealing. Unfortunately, it does not possess ANOVA's property of independent sums of squares. Noticing that $SSE = constant - SSA$, Nair (1986) and Box and Jones (1986) pointed out the undesirable property that SSE, depends on the effect of factor A.
For different examples and multifactor setting (see Hamada and Wu, 1998)

An extension of Nair's SSA: consumer preference

- We consider the analysis of consumer preference studies (see Anderson, 1959 and Schach, 1979) suppose we have I independent varieties of a product (treatments) we want ranked by k consumers (blocks) using Schach's method we obtained the resulting table is an $I \times I$ contingency table. We want to test the hypothesis of homogeneity of treatments.
- Let's consider a two way squared contingency table $\mathbf{N}$ with both border totals fixed: $I = J$ and $N_{i\bullet} = N_{\bullet j} = k$.

	1	...	j	...	I	Total
Treatment 1	N_{11}	...	N_{1J}	...	N_{1I}	k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Treatment i	N_{i1}	...	N_{iJ}	...	N_{iI}	k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Treatment I	N_{I1}	...	N_{IJ}	...	N_{II}	k
Total	k	...	k	...	k	kI

An extension of Nair's SSA: consumer preference

- We can define then a new subclass of CCS-type tests (those with fixed border totals)

$$T_{CCS.FB} = k \sum_{s=1}^{J-1} w_s \sum_{i=1}^{I} \left(\frac{Z_{is}}{k} - \frac{s}{I} \right)^2$$

- This subclass allows us to get new subsumed fixed borders (.FB) versions of the previous indices T_N , T_E and SSA, respectively.
- SSA in this case is Anderson's statistics:

$$A = \frac{I-1}{I} X^2$$

Application in several context

- Dose-response crinical trial (Hirotsu, 1997 - Australian & New Zealand Journal of Statistics)
- Statistical analysis of Pharmacological data (Matsumoto, 1997 - Journal of Statistics Medicine)
- Association between risk of disease and point sources of pollution (Lagazio, Marchi, Biggeri, 1996 - Statistica Applicata)

Empirical study

EMPIRICAL

STUDY

Comparison of X^2 and T_{CCS}: illustrative example

Consider the following table which analyzes the number of errors in an assessment test for two groups of students

	Number of errors								
	8-14	15-19	20-22	23-27	28-31	32-35	36-41	41-50	
	11	17	21	25	29	33	37	45	Tot.
G_A	16	35	61	52	23	7	4	3	201
G_B	6	10	12	13	12	15	12	22	102
T	22	45	73	65	35	22	16	25	303

The Pearson and LR statistics are $X^2 = 75.2$ and $LR = 74.5$ respectively. The statistics show there is evidence of association.

Comparison of X^2 and T_{CCS}: illustrative example

Looking at the table of cumulative cell probabilities we note that the first row grows faster than the second

	Number of errors							
	8-14	15-19	20-22	23-27	28-31	32-35	36-41	41-50
	11	17	21	25	29	33	37	45
G_A	0.080	0.254	0.557	0.816	0.930	0.965	0.985	1.0
G_B	0.059	0.157	0.275	0.402	0.520	0.667	0.784	1.0

an appropriate follow-up test would compare the hypothesis of homogeneity against trend in the row distributions. For both these situations, ordinal variables or trend alternatives, alterations to the usual chi-squared tests have been made to increase the power of the test. Without loss of generality, we are interested in detecting monotone increasing trends.

Comparison of X^2 and T_{CCS}: illustrative example

We suspected that an appropriate follow-up test would compare the hypothesis of homogeneity against de- creasing trend in the row distributions. In particular, we test the hypotheses

$$H_0 : \psi_{1s} = \psi_{2s} \qquad H_1 : \psi_{1s} \neq \psi_{2s}$$

where ψ_s denote the s-th cumulative column proportion.

	Mean	Median
A	22.55	21
B	30.29	29

Comparison of X^2 and T_{CCS}: illustrative example

We calculate the first two components of the statistics and p-values.

Component	df	Value	P-Value
$\mathbf{v}_1^T \mathbf{v}_1$	1	61.50	< 0.0001
$\mathbf{v}_2^T \mathbf{v}_2$	1	9.31	0.0023
Remainder	5	4.40	0.4938
X^2	7	75.21	< 0.0001

$$X^2 = \sum_{s=1}^{J-1} \sum_{i=1}^{I} v_{is}^2$$

It is computed using orthogonal polynomials. Since these polynomial are defined from distributional moments, the components v_{is}^2 will detected deviation in central moments: the first components detect linear deviation for the central moment, the subsequent components will detect deviation for the some type for their correspondent central moment (dipersion, skewness and so on). With a p-value significantly smaller than 0.0001, the Pearson test suggest there is an association. The highly significant first and second components suggest there are associations between the medians (Location) and dispersions. In fact, with the median of 21 and 29 cm for the Group A and Group B respectively, we can say that the Group A have significant smaller median of the Group B. Additionally, with sample standard deviations of 6.22 and 10.34 cm, the the Group A have significant less variability then the the Group B.

Comparison of X^2 and T_{CCS}: illustrative example

Component	df	Value	P-Value
$\lambda_s \mathbf{v}_1^T \mathbf{v}_1$	0.439	101.427	< 0.0001
$\lambda_s \mathbf{v}_2^T \mathbf{v}_2$	0.439	1.035	0.1295
Remainder	2.195	0.923	0.9230
T_E	3.069	103.385	< 0.0001

The $T_E/[d(I-1)]$ (adjusted Taguchi statistics) reject the hypothesis of homogeneity of the row distributions are valued at 103.385 with p-values of < 0.0001. The first component is highly significant, suggesting the existence of an association between the medians (Location).
We notice that the second component for the CCS component do not reject the hypothesis of homogeneity in the dispersions while the one for the Pearson statistic does. This is because the Taguchi and Nair CCS statistics' second components determine if the scale parameters differ from the null due to polynomial and cosinusoidal deviations, respectively, while the Pearson statistic's second component determine if the scale parameters differ from the null due to any type of deviation.

Comparison of X^2 and T_{CCS}: illustrative example

Consider the comparison of live versus televised modes of instruction. The Table give the letter grades from a course taught using the two modes of instruction (Nair, 1987).

Grades	A	B	C	D	E	Total
Live	16	30	22	4	8	80
Televised	11	19	28	8	14	80

	Grades					
Row profile	A	B	C	D	E	
Live	0.200	0.375	0.275	0.050	0.100	1.000
Televised	0.138	0.238	0.350	0.100	0.175	1.000

Cumulated	Grades					
row profile	A	B	C	D	E	Median
Live	0.200	0.575	0.850	0.900	1.000	B
Televised	0.138	0.375	0.725	0.825	1.000	C

Comparison of X^2 and T_{CCS}: illustrative example

Results:

		Index	Statistic	P-Value	$\sum_{i=1}^{I} v_{i1}^2$	P-Value
	X^2	7.085	7.085	0.131		
T_{CCS}	T_N	0.521	9.000	0.016	6.282	0.012
	T_E	13.162	8.000	0.028	5.336	0.021
	W^2	0.729	9.132	0.011	6.790	0.009
	A^2	3.462	8.877	0.015	6.312	0.012

From Table, we see that the power of T_{CCS} can be attributed only to the first components v_1^2 (Location or Linear).

	Value	P-Value
Association linear by linear	5.594	0.018

Empirical study

- A well-known case study of a polysilicon deposition process by Phadke (1989) is here considered.
- The polysicon layer is very important for defining the gate electrodes for the transistors in manufacturing very large scale integrated (VLSI) circuits. VLSI is the process of creating integrated circuits by combining thousands of transistor based circuits on a single chip.
  - One main problem occurring during the deposition process consists of the so called surface defects. The presence of surface defects usually degrades the performance of the integrated circuits.
- We have 6 factors, named A, B, C, D, E and F, with three levels

	Levels		
	1	2	3
A. Deposition temperature (^{0}C)	$T_0 - 25$	T_0	$T_0 + 25$
B. Deposition pressure (mttor)	$P_0 - 200$	P_0	$P_0 + 200$
C. Nitrogen flow (sccm)	N_0	$N_0 - 150$	$N_0 - 75$
D. Silane flow (sccm)	$S_0 - 100$	$S_0 - 50$	S_0
E. Setting time (min)	t_0	$t_0 + 8$	$t_0 + 16$
F. Cleaning method	None	CM_2	CM_3

- The main aim of the user is to identify the really important factors and determine their levels to improve process quality.

Empirical study

Several techniques have been proposed for the analysis of ordered categorical data with a focus on quality improvement in industrial settings: (e.g.)

- The accumulation analysis (AA) introduced by Taguchi;
- Nair (1986) suggested a different scoring scheme (SCORE) for AA;
- Jeng and Guo (1996) suggested a weighted probability scoring scheme (WPSS) and a single performance measure MSD (e.g. mean square deviation) derived from WPSS to reach an optimal solution;
- Asiabar and Ghomi (2006) suggested a technique called MEL;
- Wu and Yeh (2006) presents instead a weighted signal-to-noise ratio (WSNR) method, which was originally suggested by Taguchi.

All these proposals use very different approach to achieve the optimal combination solutions for the Phadke's data.

We point out that all these methods seems do not verify the statistical significance of each optimal factor level as well as of the optimal combination solution.

Empirical study

- This dataset has been also studied in D'Ambra et al. (2009) by means an exploratory approach based on a suitable correspondence analysis using the Taguchi's statistic T_E.
- We extend the approach used in D'Ambra et al. (2009), joining the information coming from a regression model for a binary dependent variable with the significance of the main results.
- To show how the changes of the levels of the factors affect the probability distribution of the defects, we examine factor effects adjusted at a specified level by building the cumulative table using the L_{18} orthogonal array (D'Ambra et al., 2009).
- The definition of surface defects and the cumulative categories are here listed

Categories	Description	Cumulative categories
I: 0–3 defects	No surface defects	(I)= I (0–3 defects)
II: 4–30 defects	Very few defects	(II)= I+II (0–30 defects)
III: 31–300 defects	Some defects	(III)= I+II+III (0–300 defects)
IV: 301–1000 defects	Many defects	(IV)= I+II+III+IV (0–1000 defects)
V: 1001 and more defects	Too many defects	(V)= I+II+III+IV+V(0–∞ defects)

Empirical study

To show how the changes of factor levels affect the probability distribution of defects, we examine factor effects adjusted at a specified level by building the cumulative table and using the L_{18} orthogonal array. The following tables shows the L_{18} orthogonal array and factor assignment where empty columns are identified by label "e".

	Factor Levels							
Experiment	e	**A**	B	C	D	E	e	F
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	**2**	1	1	2	2	3	3
5	1	**2**	2	2	3	3	1	1
6	1	**2**	3	3	1	1	2	2
7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	**2**	1	2	3	1	3	2
14	2	**2**	2	3	1	2	1	3
15	2	**2**	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

Empirical study

Final Table $\Longrightarrow$

	Number of observations by categories				
Factor Levels	(I)	(II)	(III)	(IV)	(V)
A1	34	40	51	53	54
A2	**7**	**22**	**34**	**41**	**54**
A3	8	14	19	32	54
B1	25	40	46	51	54
B2	20	28	36	43	54
B3	4	8	22	32	54
C1	19	30	32	39	54
C2	11	20	28	39	54
C3	19	26	44	48	54
D1	20	25	34	41	54
D2	13	31	42	44	54
D3	16	20	28	41	54
E1	21	27	28	43	54
E2	16	29	36	42	54
E3	12	20	30	41	54
F1	21	23	26	34	54
F2	21	30	40	46	54
F3	7	23	38	46	54

Empirical study

In order to achieve the sought statistical significance, we consider then a strategy combining the results coming from the Taguchi's statistic T and a regression model for binary dependent variables.
This strategy proceeds essentially in three-steps.

1. Due to the ordinal nature of the column variable, we compute the Taguchi's statistic T and all the χ^2_s in order to choose the optimal column-aggregated table. This table reflects the highest symmetrical association.
2. We perform a TCA graphical representation on a two dimensional plot which shows the distances from the origin to the row points. The nearest row point to the TCA origin axis will be used as reference category in a following regression model for binary dependent variables.
3. Finally, a logistic regression is then applied to the optimal column-aggregated table to identify the optimal combination and the sought statistical significance.

Empirical study

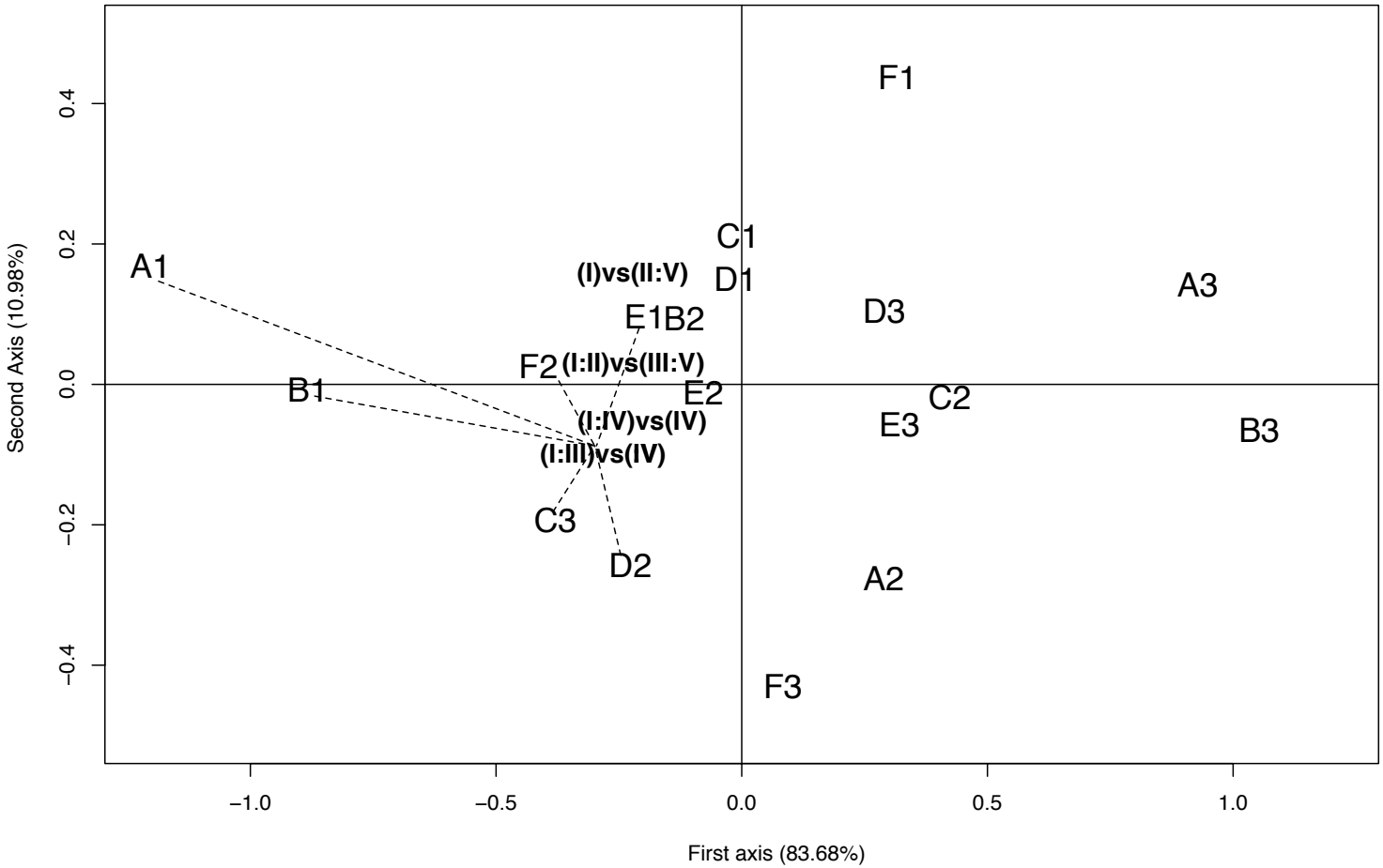
- The Taguchi statistic is $T_E = 328.920$, which is significant with 37.167 df by using the approximate asymptotic distribution of the statistic T_{CCS}.
- The following table shows the chi-squared value (or $n \times \tau$) for each cumulated table and its p-value. This table leads to choice the column-aggregated table **(I:III)vs(IV:V)** which shows the highest value. We chose it as our change-point.

	χ^2_s of $I \times 2$ contingency tables with $[y_{(1:s)}$ vs $y_{(s+1:J-1)}]$				
	(I)vs(II:V)	(I:II)vs(III:V)	(I:III)vs(IV:V)	(I:IV)vs(V)	T_E
Values	83.209	79.265	**95.879**	70.567	328.920
df	17	17	17	17	37.167
p-values	0.000	0.000	0.000	0.000	0.000

Tabella: Chi-squared with its p-value for each cumulated table and the Taguchi's statistic.

Empirical study

- We performed a TCA, that it amounts to performing a CA of cumulative frequencies using a decomposition of Taguchi statistic T_E, to visualize the strength of association between the rows and the columns (highlighted in bold).
- The following figure shows the TCA graphical representation of the results on a two-dimensional plot explaining the **94.66%** of total inertia.



Empirical study

- Next table shows the distances from the origin to the row points in the TCA plot.
- The nearest row point to the TCA origin axis is E_2 and it will be used as reference category in a following logistic regression model.

Levels	A	B	C	D	**E**	F
1	0.079	0.042	0.004	0.009	0.003	0.017
2	0.009	0.002	0.010	0.008	**0.001**	0.007
3	0.049	0.063	0.011	0.006	0.006	0.011

Tabella: Row distances from the origin of TCA plot.

Empirical study

- The distance values between the row points and the column point (I:III)vs(IV:V) (max χ^2) on the previous two dimensional TCA plot are showed in the following table

Levels	A	B	C	D	E	F
1	**0.186**	**0.266**	0.517	0.445	**0.379**	0.604
2	0.445	0.417	0.540	**0.336**	0.433	**0.344**
3	0.699	0.641	**0.304**	0.547	0.495	0.388

  This allows us to provide a ranking for the factor importance. It suggests that the plot ranking is the combination of $A_1 - B_1 - C_3 - D_2 - E_1 - F_2$.

- We remark that this is an **explorative** solution obtained from a correspondence analysis using T_E.
- We can build up this solution by using the definition of Taguchi's statistic which be approximately computed as sum of likelihood ratios (SLR).
  Each LR is computed by applying a grouped logistic model to each aggregated subtable, using **reference categories** "E_2" (nearest to the origin axis of TCA).
- We use the logistic model for the probabilities p_{is} for the its simplicity of interpretation of the coefficients in terms of odds and odds ratios

Empirical study

- The following table shows the Likelihood Ratio (LR) and their p-values derived by a grouped logistic model on each aggregated subtable.
- This table leads to choice again the column-aggregated table (I:III)vs(IV:V) which shows the highest contribution.

	Likelihood ratio test for $H_0 : p_{is}/p_{i.} = p_{.s}$				
	(I)vs(II:V)	(I:II)vs(III:V)	(I:III)vs(IV:V)	(I:IV)vs(IV)	$SLR \approx T$
LR statistic	87.022	83.533	**103.061**	77.729	351.345
df	17	17	17	17	
p-values	0.000	0.000	0.000	0.000	

- We use the identified subtable (I:III)vs(IV:V) for a logistic regression model to choose the optimal solution.

Empirical study

- It is well known that the exponential of the logistic coefficients Exp(Coefficient) are equivalents to the odds. The Exp(Coef) values and their associated *p*-values, are then computed.

	Coefficient	Stand. Err.	Wald	DoF	p-value	Exp(Coef)
A1	2.140	0.661	10.498	1	0.001	8.500
A2	-0.163	0.403	0.162	1	0.687	0.850
A3	-1.304	0.406	10.335	1	0.001	0.271
B1	1.056	0.480	4.847	1	0.028	2.875
B2	0.000	0.408	0.000	1	1.000	1.000
B3	-1.068	0.400	7.125	1	0.008	0.344
C1	-0.318	0.400	0.634	1	0.426	0.727
C2	-0.619	0.397	2.433	1	0.119	0.538
C3	0.788	0.454	3.017	1	0.082	2.200
D1	-0.163	0.403	0.162	1	0.687	0.850
D2	0.560	0.436	1.644	1	0.200	1.750
D3	-0.619	0.397	2.433	1	0.119	0.538
E1	0.172	0.415	0.172	1	0.679	1.188
E3	-0.470	0.398	1.395	1	0.238	0.625
F1	-0.767	0.397	3.737	1	0.053	0.464
F2	0.357	0.424	0.708	1	0.400	1.429
F3	0.172	0.415	0.172	1	0.679	1.188

Empirical study

- The following table reports the computed Exp(Coef) (Odd ratios) values according to the factors and levels with their p-values.

Level	A	B	C	D	E	F
1	**8.500** (**0.01**)	**2.875** (**0.028**)	0.727 (0.426)	0.850 (0.687)	**1.188** (0.679)	0.464 (**0.053**)
2	0.850 (0.687)	1.000 (1.000)	0.538 (0.119)	**1.750** (0.200)	- -	**1.429** (0.400)
3	0.271 (**0.001**)	0.344 (**0.008**)	**2.200** (0.082)	0.538 (0.119)	0.625 (0.238)	1.188 (0.679)

- Bold values (maximal Exp(Coef) values) identify the optimal combination:

$$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$$

They can be ranked according to their Exp(Coef) values. Some of them turn out to be statistically significant too.

Empirical study

Comparative results for the optimal factor settings

	Optimal combination	
TM	$A_1 - B_1 - C_1 - D_1 - E_2 - F_2$	$\neq$
TCA	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	$=$
MEL	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	$=$
SCORE	$A_1 - B_1 - C_3 - D_2 - E_2 - F_2$	$\neq$
WSNR	$A_1 - B_1 - C_1 - D_1 - E_2 - F_2$	$\neq$
AA	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	$=$
MSD	$A_1 - B_1 - C_3 - D_2 - E_1 - F_3$	$\neq$
WPSS	$A_1 - B_1 - C_3 - D_2 - E_1 - F_3$	$\neq$
TCALR	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	$=$

- (TCALR) TCA+Logistic Regression (D'Ambra et al., 2018)
- (TM) Taguchi's two steps optimization method (Wu and Yeh, 2006)
- (AA) Accumulation analysis (Taguchi, 1986)
- (WPSS) Weighted probability scoring scheme (Jeng and Guo, 1996)
- (SCORE) Scoring scheme (Nair, 1986)
- (MSD) Single performance measure (Jeng and Guo, 1996)
- (WSNR) Weighted signal-to-noise ratio (Wu and Yeh, 2006)
- (MEL) Minimizing the expected loss (Asiabar and Ghomi, 2006)

Empirical study

- Finally, our approach builds up this result allowing to choose also another sub-optimal combination. This is obtained by selecting the best Exp(Coef) values that have a p-value lower than a predetermined one by the user.
- For instance, the sub-optimal solution with a p-value less or equal to a preselected value of **0.082** is

$$A_1 - B_1 - C_3 - F_1$$

with no significant E and F factor levels.

Level	A	B	C	D	E	F
1	**8.500** (**0.01**)	**2.875** (**0.028**)	0.727 (0.426)	0.850 (0.687)	**1.188** (0.679)	0.464 (**0.053**)
2	0.850 (0.687)	1.000 (1.000)	0.538 (0.119)	**1.750** (0.200)	- -	**1.429** (0.400)
3	0.271 (**0.001**)	0.344 (**0.008**)	**2.200** (0.082)	0.538 (0.119)	0.625 (0.238)	1.188 (0.679)

Empirical study

Due to the methodological differences between all these proposal, we compute the likelihood ratio (LR) (for each optimal solution) of a nested model to identify the method with the highest value. These likelihood ratios are computed by using the indicator matrix of the optimal solutions of each mehod as predictive variables.

	Optimal combination	LR	df	p-value
TM	$A_1 - B_1 - C_1 - D_1 - E_2 - F_2$	49.595	6	0.000
MEL	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	**74.240**	6	0.000
SCORE	$A_1 - B_1 - C_3 - D_2 - E_2 - F_2$	72.120	6	0.000
WSNR	$A_1 - B_1 - C_1 - D_1 - E_2 - F_2$	49.595	6	0.000
AA	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	**74.240**	6	0.000
MSD	$A_1 - B_1 - C_3 - D_2 - E_1 - F_3$	71.408	6	0.000
TCALR	$A_1 - B_1 - C_3 - D_2 - E_1 - F_2$	**74.240**	6	0.000

- (TM) Taguchi's two steps optimization method (Wu and Yeh, 2006)
- (AA) Accumulation analysis (Taguchi, 1986)
- (WPSS) Weighted probability scoring scheme (Jeng and Guo, 1996)
- (SCORE) Scoring scheme (Nair, 1986)
- (MSD) Single performance measure (Jeng and Guo, 1996)
- (WSNR) Weighted signal-to-noise ratio (Wu and Yeh, 2006)
- (MEL) Minimizing the expected loss (Asiabar and Ghomi, 2006)
- (TCALR) TCA+Logistic Regression (D'Ambra et al., 2018)

Empirical study

The proposed research regards the travelling back and forth by train from Naples to Rome and it aims to choose the best scenario for users and company for this service. To this purpose we have selected four control factors (A = Journey time; B = Comfort, C = Cost; D = Frequency) each of which with 3 levels.

Level	Factors			
	[A] Journey time	[B] Comfort	[C] Cost	[D] Frequency
1	1:29	70 cm	Euro 19.50	Each four hours
2	1:54	75 cm	Euro 27.60	Each two hours
3	2:09	80 cm	Euro 36.10	Each hours

The number of different runs in a complete factorial design will be then 81.

Empirical study

To show how the changes of the levels of the factors affect the probability distribution of the satisfaction, we examine factor effects adjusted at a specified level by building the cumulative table by means of an L_9 orthogonal array.

Scenario	Factors levels			
	[A]	[B]	[C]	[D]
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

Empirical study

The contingency table between the factor levels and the evaluations of the potential users is reported in following table.

Factor		Evaluation				
		I	II	III	IV	V
A1	Low Journey time	9	49	42	20	54
A2	Middle Journey time	26	31	42	44	31
A3	High Journey time	31	56	46	34	7
B1	Low comfort	19	53	62	34	6
B2	Middle comfort	31	43	29	20	51
B3	High comfort	16	40	39	44	35
C1	Low cost	13	26	29	35	71
C2	Middle cost	8	52	65	33	16
C3	High cost	45	58	36	30	5
D1	Low frequency	42	67	41	13	11
D2	Middle frequency	11	42	47	47	27
D3	High frequency	13	27	42	38	54

Empirical study

	w_s	$D_s(1-D_s)$	X_s^2	$w_s D_s(1-D_s)X_s^2$
(I)vs(II:V)	0.200	0.110	93.660	2.060
(I:II)vs(III:V)	0.200	0.237	135.491	6.468
(I:III)vs(IV:V)	0.200	0.231	171.512	7.921
(I:IV)vs(IV)	0.200	0.145	221.309	6.416
Nair T_N				22.864

	w_s	$D_s(1-D_s)$	X_s^2	$w_s D_s(1-D_s)X_s^2$
(I)vs(II:V)	9.054	0.110	93.660	93.660
(I:II)vs(III:V)	4.215	0.237	135.491	135.491
(I:III)vs(IV:V)	4.320	0.231	171.512	171.512
(I:IV)vs(IV)	6.888	0.145	221.309	221.309
Taguchi T_E				622.972

Empirical study

	w_s	$D_s(1-D_s)$	X_s^2	$w_s D_s(1-D_s)X_s^2$
(I)vs(II:V)	0.126	0.110	93.660	1.305
(I:II)vs(III:V)	0.261	0.237	135.491	8.416
(I:III)vs(IV:V)	0.249	0.231	171.512	9.864
(I:IV)vs(IV)	0.188	0.145	221.309	6.018
Cramer von-Mises W^2				25.602

	w_s	$D_s(1-D_s)$	X_s^2	$w_s D_s(1-D_s)X_s^2$
(I)vs(II:V)	1.145	0.110	93.660	11.819
(I:II)vs(III:V)	1.098	0.237	135.491	35.491
(I:III)vs(IV:V)	1.076	0.231	171.512	42.629
(I:IV)vs(IV)	1.293	0.145	221.309	41.466
Anderson Darling A^2				131.405

Empirical study

	T_E Nair	T_N Taguchi	W^2 CvM	A^2 AD
δ	0.0756	0.0745	0.0763	0.0764
T_{CCS}	22.8644	621.9719	25.6024	131.4045
$d(I-1)$	0.0647	1.6066	0.0799	0.3617
$T_{CCS}/[d(I-1)]$	353.1314	387.1288	320.4126	363.2658
df	24.6124	27.3866	22.1200	25.0498
P.Value	0.0000	0.0000	0.0000	0.0000

Main References

- Agresti, A.: Categorical Data Analysis, 3rd edn. Wiley, Hoboken, NJ (2013)
- Beh, E.J., D'Ambra, L., Simonetti, B.: Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic. Commun. Stat. Theory Methods 40, 1620–1632 (2011)
- Cuadras, C.M., Cuadras, D.: Unified approach for the multivariate analysis of contingency tables. Open J. Stat. 5, 223–232 (2015)
- D'Ambra A., P. Amenta, E.J. Beh: Confidence regions and other tools for an extension of correspondence analysis based on cumulative frequencies. AStA Advances in Statistical Analysis (2021)
- D'Ambra, L., Amenta, P., D'Ambra, A.: Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation. Stat. Methods Appl. 27(2), 297–318 (2018)
- Grilli L, Rampichini C (2002) Scomposizione della dispersione per variabili statistiche ordinali. Stat LXI I(1):111–116

Main References

- Hirotsu, C.: Cumulative chi-squared statistic as a tool for testing goodness of fit. Biometrika 73, 165–173 (1986)
- Leti, G.: Statistica descrittiva. Il Mulino (1983)
- Nair VN (1986) Testing in industrial experiments with ordered categorical data. Technometrics 28(4):283– 291
- Satterthwaite F (1946) An approximate distribution of estimates of variance components. Biom Bull 2:110– 114
- Taguchi G (1974) A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. Saishin Igaku 29:806–813
- Takeuchi K, Hirotsu C (1982) The cumulative chi square method against ordered alternative in two-way contingency tables. Technical report, vol 29. Reports of Statistical Application Research. Japanese Union of Scientists and Engineers