Rank-Based Family of Probability Laws for Testing Homogeneity of Variable Groupings

Manuel L. Esquível [1]

[1]Nova SST and Nova Math, **New University of Lisbon (NUL)**, Portugal

Joint work with **Nadezhda Krasii**, DSTU, Rostov-on-Don, Russia & Nova Math NUL and **Pedro P. Mota**, Nova SST and Nova Math NUL, **Célia Nunes** DM & CMA, UBI and **Kwaku Opoku-Ameyaw**, Cocoa Research Institute of Ghana & CMA, UBI and **João T. Mexia**, Nova Math NUL

10th International Conference on Stochastic Methods, *devoted to 90th anniversary of the Department of Probability Theory of Lomonosov Moscow State University.* 1 – 5 June, 2025, Raduga – Divnomorskoye, Russia

Outline

Motivation – I: Data: Cocoa cultivation data experiment in Ghana

Data: Cocoa plant varieties V_i; Soils:1, . . . , 4; Plant specimen variables: PH, SD, DM.

Variety	Soil	Plant height	Stem diameter	Dry matter
V1	1.	68.25	11.51	44.725
V1	2.	43.6333	6.26333	13.59
V1	3.	68.3333	11.0167	38.3167
V1	4.	76.9333	11.5267	50.77
V1	1.	76.75	10.945	57.365
V1	2.	55.6333	8.75	23.245
V1	3.	59.	10.77	28.33
V1	4.	66.	10.3233	39.225
V1	1.	64.7667	10.9167	23.1767
V1	2.	47.925	8.8275	16.12
V1	3.	72.25	10.2	22.595
V1	4.	78.7333	10.2425	35.2267
V2	1.	74.8	10.39	49.18
V2	2.	42.5333	7.62	11.865
V2	3.	69.2667	11.4967	45.55
V2	4.	84.05	11.7	47.86
V2	1.	63.1	10.8767	43.165
V2	2.	58.9667	8.96	17.37
V2	3.	65.7667	11.26	56.115
V2	4.	71.	11.3467	44.785
V2	1.	67.9333	10.7833	27.3333
V2	2.	55.1667	9.44667	13.305
V2	3.	75.075	10.77	22.8067
V2	4.	79.675	10.3	30.7233
V3	1.	58.6667	10.8267	31.19
V3	2.	45.25	7.73	19.285

Motivation – II – More on Data, variable grouping

Varieties: genetic characteristics induce natural grouping of variables

Table: I – List of cocoa plant varieties in the experiment.

Code	Variety	Code	Variety	Code	Variety
V_1	$T85/799 \times PA7/808$	V_2	$T60/887 \times CRG8914$	V_3	$PA150 \times EQ3338$
V_4	$PA150 \times PA88$	V_5	$PA150 \times CRG0314$	V_6	Standard Variety
V_{11}	$PA150 \times CRG3019$	V_{12}	$T63/967 \times IMC60$	V_{13}	$T63/967 \times EQ78$
V_{14}	$T63/967 \times CRG2022$	V_{15}	$T63/967 \times CRG9066$	V_{16}	$T63/967 \times CRG0314$

A first natural grouping of variables – by genetic ascendant

12 different varieties: $V_1, \ldots, V_6$ and $V_{11}, \ldots, V_{16}$,
4 from the ascendant PA150 (V_3, V_4, V_5, V_{11}),
5 from T63/967 (V_{12}, V_{13}, V_{14}, V_{15}, V_{16})
and 3 other varieties (V_1, V_2, V_6) with **NO** common ascendant.

Question on variety grouping

Does grouping of data variables for varieties **with a common ascendant** brings additional information?

The grouping discrete distributions – I – Definition

Parametric discrete probability laws (Reference **[1]**)

Grouping probability laws; Parameters: $N,\ p,\ n_1, \dots, n_p$

$\mathcal{G}(N, p, n_1, \dots, n_p)$ is the probability law defined by:

1. The set of integers $\mathcal{S} = \{1, 2, \dots, N\}$;
2. $\boldsymbol{S}_p$ set of all partitions of $\mathcal{S}$ in p subsets $\{\mathcal{S}_1, \dots, \mathcal{S}_p\}$ s.t.:

$$\#\mathcal{S}_k = n_k \ , \ k = 1, \dots, p \text{ and } n_1 + \cdots + n_p = N \ .$$

3. Values of $Q_{\mathcal{G}}$, defined on $\boldsymbol{S}_p$:

$$Q_{\mathcal{G}}(\mathcal{S}_1, \dots, \mathcal{S}_p) = \sum_{k=1}^{p} \sum_{i,j \in \mathcal{S}_k} |i - j| \ , \ \{\mathcal{S}_1, \dots, \mathcal{S}_p\} \in \boldsymbol{S}_p \ ,$$

with corresponding frequencies (determined!).

Grouping discrete distributions – II – First example – i

Cocoa plant variable data grouping from variety common ascendent

$\mathcal{G}(9,2,4,5)$; Parameter values: $N=9,\ p=2,\ n_1=4,\ n_2=5$

Number of partitions of $\{1,2,\ldots,9\}$ into two subsets of 4 and 5 elements:

$$\#\boldsymbol{S}_2 = \frac{9!}{4!\cdot 5!} = 126$$

Table: **II** – Values taken by $Q_{\mathcal{G}(9,2,4,5)}$ and corresponding probabilities

Code	**Variety**	**Code**	**Variety**	**Code**	**Variety**		
$(60,\ \frac{1}{63})$	$(74,\ \frac{1}{63})$	$(84,\ \frac{2}{63})$	$(90,\ \frac{1}{21})$	$(92,\ \frac{2}{63})$	$(94,\ \frac{1}{63})$	$(100,\ \frac{1}{14})$	$(102,\ \frac{2}{63})$
$(106,\ \frac{2}{21})$	$(108,\ \frac{10}{63})$	$(110,\ \frac{1}{21})$	$(112,\ \frac{5}{126})$	$(114,\ \frac{10}{63})$	$(116,\ \frac{1}{6})$	$(118,\ \frac{4}{63})$	$(120,\ \frac{1}{126})$

On the probability law $\mathcal{G}(9,2,4,5)$

For 126 partition configurations only 16 values taken by $Q_{\mathcal{G}(9,2,4,5)}$
No discernible pattern of regularities neither for values taken nor for the probabilities

Grouping discrete distributions – II – First example – ii

Statistic $\mathcal{G}(9,2,4,5)$: Generic values; Computation of Observed values

$\mathcal{G}(9,2,4,5)$; Parameter values: $N = 9,\ p = 2,\ n_1 = 4,\ n_2 = 5$

$\mathcal{S} = \{1,2,\ldots,9\}$, $\#S_1 = 4$, $\#S_5 = 5$, $k = 1,2,3$ the variables

$$Q^{S_1,S_2}_{\mathcal{G}(9,2,4,5)} = \sum_{i,j\in\{1,\ldots,4\}} \left|r^1_{k,i} - r^1_{k,j}\right| + \sum_{i,j\in\{5,6,\ldots,9\}} \left|r^2_{k,i} - r^2_{k,j}\right|$$

$r^1_{k,i}$ – a generic rank, for variable k, for $i \in S_1$

$r^2_{k,i}$ – a generic rank, for variable k, for $i \in S_2$

Computation of Observed values of statistic $\mathcal{G}(9,2,4,5)$

For each variable, Plant Height, Stem Diameter, Dry Matter, and each soil $s_1,\ldots,s_4$:

(1) - Sum the observations for each variety with common ascendent $PA150$

(2) - Sum the observations for each variety with common ascendent $T63/967$

(3) - Consider the ranks, in $\{1,\ldots,9\}$, of (1) to be $\mathcal{S}_1$ in (2) to be $\mathcal{S}_2$

(4) - Compute $Q^{S_1,S_2}_{\mathcal{G}(9,2,4,5)}$

Grouping discrete distributions – II – First example – iii

Statistic $\mathcal{G}(9,2,4,5)$: Quantiles; Observed values; Test results interpretation

Quantiles for the probability law $\mathcal{G}(9,2,4,5)$

From Table: **II** $q_{0.05} = 84$; $q_{0.1} = 90$

Table: **III** Observed values for the 3 variables and for the 4 soils S_i

	$Q_{1,obs}$(Plant Height)	$Q_{2,obs}$(Stem Diameter)	$Q_{3,obs}$(Dry Matter)
S_1	112	108	92
S_2	60*	90**	74*
S_3	108	116	108
S_4	110	114	90**

* - Reject H_0 (null hypothesis) at $q = 0.05$; ** - Reject H_0 at $q = 0.1$.

Interpretation of the test results (Reference **[2]**)

Reject H_0 = grouping is **not** significant for soil S_2 (3 variables).
Soil S_2: grouping gives statistically significant **homogeneity**.

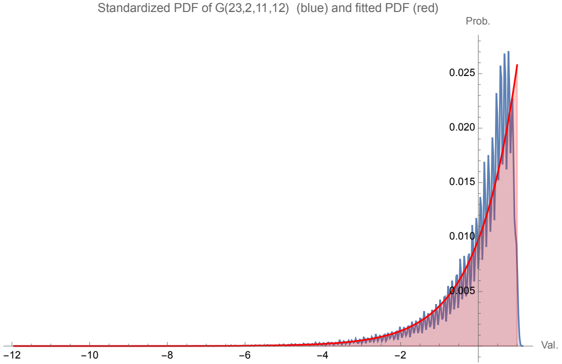
Grouping discrete distributions – III – Challenges – i

Need of massive computational work and unrecognisable probability laws.

Computational challenges for moderate large values of parameters

Number of evaluations of statistic $Q^{S_1,S_2}_{\mathcal{G}(N,2,n_1,N-n_1)}$ is: $2^N - 1$!

Figure: I – Fitting a Gaussian (red) to the standardised law of $X \frown \mathcal{G}(23, 2, 11, 12)$ (blue)



Grouping discrete distributions – III – Challenges – ii

Towards an asymptotic result: observed properties – i

On ratios of moments for standardised grouping distributions

Y_N, standardised $X \frown \mathcal{G}(N, 2, n_1, n_2)$, $N = 13, 15, 17, 19, 21, 23$:

(1) - consider $\left| \frac{\mathbb{E}[Y_N^{n+1}]}{\mathbb{E}[Y_N^n]} \right|$ for $n = 2, \ldots, 60$ and fit

(2) - logistic functions of the form, $f_{L,a}(x) = \frac{L}{1+e^{-ax}}$ for $a > 0$,

(3) - observe that $L \approx \frac{N}{2}$ (Table: **IV** and also Figure: **II**).

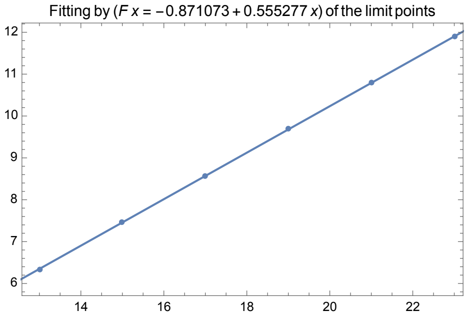
Table: **IV** – L values for moment ratios: standardised $X \frown \mathcal{G}(N, 2, n_1, n_2)$, $N = 13, 15, 17, 19, 21, 23$

$\mathcal{G}(13, 2, 6, 7) : L \approx 6.34319$	$\mathcal{G}(15, 2, 7, 8) : L \approx 7.45872$	$\mathcal{G}(17, 2, 8, 9) : L \approx 8.57213$
$\mathcal{G}(19, 2, 9, 10) : L \approx 9.68292$	$\mathcal{G}(21, 2, 10, 11) : L \approx 10.79090$	$\mathcal{G}(23, 2, 11, 12) : L \approx 11.89560$

Grouping discrete distributions – III – Challenges – iii

Towards an asymptotic result: observed properties – ii

Figure: **II** – Linear fitting of L points: $X \frown \mathcal{G}(N, 2, n_1, n_2)$ with $N = 13, 15, 17, 19, 21, 23$



On observed and then plausible hypothesis for an asymptotic result

(1) - $\mathbb{E}[Y_m^n] = (-1)^n \, |\mathbb{E}[Y_m^n]|$

(2) - $\left|\frac{\mathbb{E}[Y_m^{n+1}]}{\mathbb{E}[Y_m^n]}\right| = \frac{p_m}{N_m} + \epsilon_n^m$ with $\lim_{n\,,\, m \to +\infty} \left(\frac{p_m}{N_m} + \epsilon_n^m\right) = c$

Grouping discrete distributions – III – Challenges – iv

Towards an asymptotic result: observed properties – iii

Questions on plausible hypothesis for an asymptotic result

Asymptotics of ratios of standardised grouping laws, stable?
Right hand behaviour of residuals, towards non zero value?
Should we take $L \approx -0.871073 + 0.555277 * N$? (see Figure: **II**)

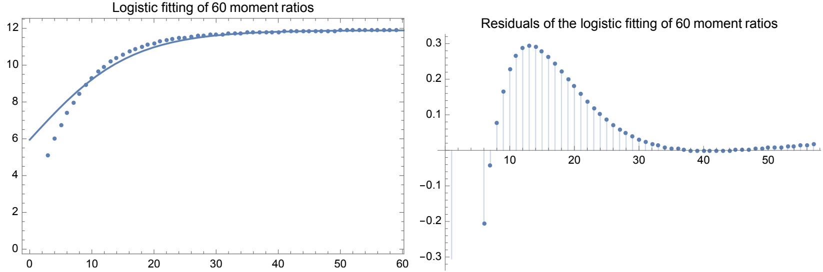


Figure: **III** Logistic fitting of 60 absolute values of moment ratios for standardised law of $X \frown \mathcal{G}(23, 2, 11, 12)$ (**left**) and residuals (**right**).

Grouping discrete distributions – IV – Asymptotic law?– i

A negative asymptotic result (via standardised laws): formulation (Reference **[3]**)

Theorem: asymptotic behaviour for $\mathcal{G}(N, p, n_1, \ldots, n_p)$, large N, p

If $\lim_{m\to+\infty} \frac{p_m}{N_m} = c$ with,

$$\sum_{k\geq 1} \frac{|\epsilon_k^m|}{p_m/N_m} < +\infty \; , \; d = \lim_{m\to+\infty} \prod_{k=2}^{+\infty} \left(1 + \frac{\epsilon_k^m}{p_m/N_m}\right) \; ,$$

then there is a unique generalised function $\lambda_{cd}(x)$—that is not a probability measure—with moment generating function $\mathcal{M}_{\lambda_{cd}}$

$$\mathcal{M}_{\lambda_{cd}}(t) := \langle \lambda_{cd}(x), e^{tx} \rangle = 1 + \frac{d}{c^2}\left(e^{-ct} - 1 + ct\right) + (1-d)\frac{t^2}{2} \; ,$$

for t in a non-empty interval centred at zero.

Grouping discrete distributions – IV – Asymptotic law?– ii

A negative asymptotic result:proof and interpretation (Reference **[3]**)

Idea of the proof, under the hypothesis:

Apply: convergence of moments $\Rightarrow$ weak convergence:

(1) - there exist $\widetilde{\mathcal{G}}^{\infty}$ limit law for $(\widetilde{\mathcal{G}}(N_m, p_m, n_{m,1}, \dots, n_{m,p_m}))_{m\geq 1}$

(2) - For Y_c standardised from $X \frown \mathcal{G}^{\infty}$ we have, for $n \geq 3$:

$\mathbb{E}\left[Y_c^0\right] = 1\ ,\ \mathbb{E}\left[Y_c^1\right] = 0\ ,\ \mathbb{E}\left[Y_c^2\right] = 1\ ,\ \mathbb{E}\left[Y_c^n\right] = d(-c)^{n-2}$

(3) - Summing: $\mathcal{M}_{Y_c}(t) = 1 + \frac{d}{c^2}\left(e^{-ct} - 1 + ct\right) + (1-d)\frac{t^2}{2}$

(4) - If $c \neq 0$ (for $d = 1$ all the subgroups with $\# = 1$):

$\lambda_{cd}(x) = \left(1 - \frac{d}{c^2}\right)\delta_0(x) + \frac{d}{c^2}\delta_{-c}(x) - \frac{d}{c}\delta_0'(x) + \frac{1-d}{2}\delta_0''(x)$

(5) - If $c = 0$ (at least one subgroup $\# = \infty$):

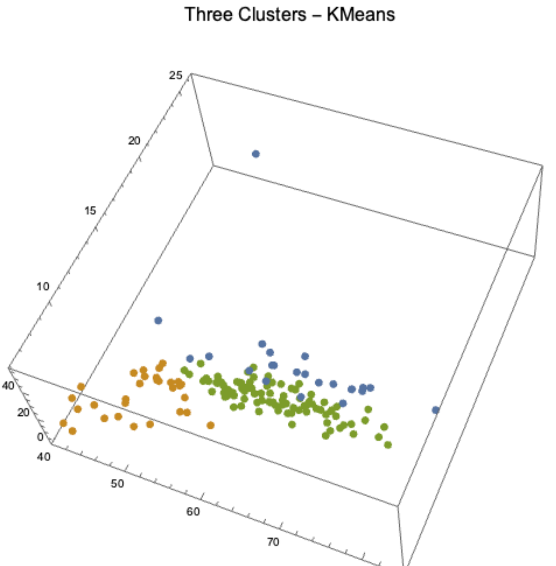
$\lambda_0(x) = \delta_0(x) + \frac{1}{2}\delta_0''(x)$.

Negative conclusion from observed characteristics of cases studied:

Do not lead to valid hypothesis for an asymptotic behaviour of grouping laws.

Grouping discrete distributions – V – Second Example – i

Grouping by *KMeans* clustering of variable values (Reference **[3]**).



Grouping discrete distributions – V – Second Example – ii

KMeans clustering grouping.

For each soil and variable: varieties distribution by cluster

Consistent grouping inside the same soils!

Table: **V** – Clusters of varieties according to variables and soils

Types of Soil	**Plant Height**	**Stem Diameter**	**Dry Matter**
Soil 1: cluster varieties [a]	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$
Soil 2: cluster varieties	**1** : $\{V_{14}\}$ [b] **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\{V_{14}\}$ [b] **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\{V_{14}\}$ [b] **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$
Soil 3: cluster varieties	**1** : $\emptyset$ **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\emptyset$ **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\emptyset$ **2** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **3** : $\{V_2, V_5, V_{16}\}$
Soil 4: cluster varieties	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$	**1** : $\mathcal{V} \setminus \{V_2, V_5, V_{16}\}$ **2** : $\emptyset$ **3** : $\{V_2, V_5, V_{16}\}$

[a] $\mathcal{V} = \{V_1, V_2, V_3, V_4, V_5, V_6, V_{11}, V_{12}, V_{13}, V_{14}, V_{15}, V_{16}\}$. [b] One observation of V_{14} in Cluster 2.

Grouping discrete distributions – V – Second Example – iii

Test statistic and results interpretation.

Table: **VI** – Values of the grouping statistic: 3 variables and 4 soils.

Types of Soil	Plant Height	Stem Diameter	Dry Matter
Soil 1	360	324	344
Soil 2	332	344	332
Soil 3	248 [a]	284 [a]	268 [a]
Soil 4	320	312	352

[a] Quantiles: $q_{\mathcal{G}(12,2,9,3):0.05} = 296 = q_{\mathcal{G}(13,3,9,3,1):0.05}$.

Interpretation of the test results (Reference [3])

$H_0 \approx$ grouping produces inhomogeneous groups.
Reject H_0 (Soil 3): the within-group ranks are homogeneous.
Soil 3: homogeneous production characteristics in 3 variables.
Sets of ranks analysis to determine if production is good or bad;
e.g.: varieties V_5 and V_{16} may be inadequate for Soil 3.

Conclusions and future work

Ongoing research on grouping distributions

Conclusions of the ongoing research

- Identification of the asymptotic law for large value of parameters of grouping distributions not yet available.
- A different set of hypothesis needed.

Future immediate work

- Two suggested approximations of the probability law of grouping distributions via approximate (sampling) sequence of moments: **discrete laws** and **normal mixtures**.
- Numerical semigroup studies for sums of independent grouping distributions (preliminary version of preprint available).

Main Suport References (article link on the title)

Mathematical and Statistical results: [1], More experimental results [2–3]

[1] Manuel L. Esquível and Nadezhda P. Krasii and Pedro P. Mota and Célia Nunes and Kwaku Opoku-Ameyaw, Rank-Based Family of Probability Laws for Testing Homogeneity of Variable Grouping. Mathematics 2025, 13(11), 1805, published online May 28, 2025.

[2] Kwaku Opoku-Ameyaw and Célia Nunes and Manuel L. Esquível and João Tiago Mexia CMMSE: a nonparametric test for grouping factor levels: an application to cocoa breeding experiments in acidic soils. *Journal of Mathematical Chemistry*, (**61**):652–672, published online December, 9, 2022.

[3] Kwaku Opoku-Ameyaw and Célia Nunes and Manuel L. Esquível and João Tiago Mexia. Grouping factor levels in cocoa breeding experiments. Submitted, March 6, 2025.

Acknowledgements and Funding

INSTITUTIONS: DSTU, Russia & NUL & UBI, Portugal; FUNDING:

This presentation was partially funded by Portuguese funds through the Fundação para a Ciência e a Tecnologia, FCT I.P., via projects UIDB/00297/2020 and UIDP/00297/2020 of the Nova Math NUL
© Nova SST, New University of Lisbon (NUL).

REPÚBLICA PORTUGUESA

NOVA id FCT
Associação para a Inovação e Desenvolvimento da FCT

NOVAMATH
CENTER FOR MATHEMATICS + APPLICATIONS

fct Fundação para a Ciência e a Tecnologia