V –Combinatorial Inference in GDA

Far better an approximate answer to the right question . . .
than an exact answer to the wrong one.
J. Tukey

Introduction to Inductive Data Analysis

- *Descriptive procedures*
  (means, variances, eigenvalues, etc.).
  1) They do not depend on sample size.
  2) They lead to *descriptive conclusions*.
- *Inference procedures*
  (significance tests, confidence intervals, etc.).
  attempt to extend descriptive conclusions.
  1) They depend on sample size.
  2) They lead to *inductive conclusions*.

Paradigm: $\chi^2 = n\Phi^2$

"Test statistic = sample size × descriptive statistic"

Statistical modeling

The statistical modeling must be as assumption–free as possible.

1. Instead of *normal modeling*, prefer *combinatorial framework*. Combinatorial inference relaxes the frequentist framework and bases inference on proportions of samples.
2. Instead of *general modeling* (e.g. "general linear model") prefer *specific modeling*, that is, put the statistical model on the *specific data set* relevant to the hypothesis of interest.

Performing a permutation test

1. Outline the *effect of interest*, then define the *group of permutations* of observations that is consistent with the absence of effect ("null hypothesis").
2. Choose a suitable *test statistic* and compute it for the observed data.
3. Determine the *permutation distribution* of the test statistic by calculating the values of the test statistic for all possible rearrangements or for a large sample thereof.
4. Determine the *p–value*: calculate the proportion of rearrangements for which test statistic values are more extreme than or as extreme as the observed one.

Two types of permutation tests:

- *exact tests* (exhaustive method or Monte Carlo method)
- *approximate tests*

Permutation modelling

keep assumptions at a lower level, avoiding those that are difficult to justify or to interpret;
they do not depend on assumptions on the distribution of observations,

Permutation tests are distribution–free and nonparametric.

exchangeability

Chance formulations and tests of randomness

Randomization and permutation

VI.2 Inductive Data Analysis Philosophy

Statistical procedures should dig out "what the data have to say", and depend as little as possible on gratuitous hypotheses, unverifiable assumptions, etc.
The combinatorial framework is the most in harmony with geometric data analysis methods.

- Typicality tests
  comparing the mean point of a subcloud to the mean point of a reference cloud;
  comparing the mean point of a subcloud to a reference point.
- Homogeneity tests

Two steps:

(1) *descriptive analyses*, that is, looking at the importance of effects and *stating* the descriptive conclusions;
(2) *inductive analyses* the main objective of which is to *corroborate* (whenever possible) descriptive conclusions.

VI.3 Combinatorial Typicality Tests

Typicality situations
Committee — Gifted children — Robespierre — Target example

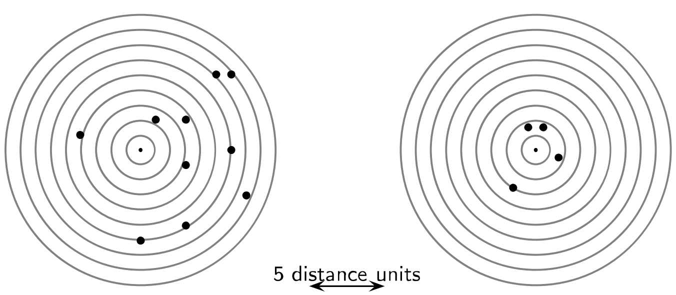

Figure: *Target example*. Target with 10 impacts and target with 4 impacts.

We wonder if the target shooter of the set of 4 impacts is the same as that of the set of 10 impacts?

The typicality problem
"Can the group of observations be assimilated to the reference population, or is it atypical of it?"

Finite sampling: Typicality test for the mean
Sample space: set of *samples* of size n_c of the population of size n, i.e. the set of all $\binom{n}{n_c}$ n_c–element subsets.

Statistic of interest: *Mean*
we compute

- the proportion of samples whose means are $\geq$ the observed one;
- the proportion of samples whose means are $\leq$ to the observed one.

The *combinatorial p–value* of the test is the smaller of these two proportions.

Example: Committee

1 ↦ *58*	4 ↦ *64*	7 ↦ *67*
2 ↦ *61*	5 ↦ *64*	8 ↦ *70*
3 ↦ *64*	6 ↦ *67*	9 ↦ *70*

Committee of 3 members with mean age 69.

List of the 84 samples of size 3 with their means

123	61	146	63	234	63	259	65	358	66	469	67
124	61	147	63	235	63	267	65	359	66	478	67
125	61	148	64	236	64	268	66	367	66	479	67
126	62	149	64	237	64	269	66	368	67	489	68
127	62	156	63	238	65	278	66	369	67	567	66
128	63	157	63	239	65	279	66	378	67	568	67
129	63	158	64	245	63	289	67	379	67	569	67
134	62	159	64	246	64	345	64	389	68	578	67
135	62	167	64	247	64	346	65	456	65	579	67
136	63	168	65	248	65	347	65	457	65	589	68
137	63	169	65	249	65	348	66	458	66	678	68
138	64	178	65	256	64	349	66	459	66	679	68
139	64	179	65	257	64	356	65	467	66	689	69
145	62	189	66	258	65	357	65	468	67	789	69

Distribution of the statistic *Mean*

				457				
				456				
				357				
			345	356	567			
			257	347	467	579		
			256	346	459	578		
		245	247	267	458	569		
		235	246	259	367	568		
		234	237	258	359	479		
		157	236	249	358	478		
		156	167	248	349	469		
		147	159	239	348	468		
	145	146	158	238	279	379	679	
	135	137	149	179	278	378	678	
125	134	136	148	178	269	369	589	
124	127	129	139	169	268	368	489	789
123	126	128	138	168	189	289	389	689
61	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*

Valeurs m	*61*	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*	
$\mathrm{n}(M=m)$	3	5	11	14	17	14	13	5	2	[84]
$\mathrm{p}(M=m)$	0.036	0.060	0.131	0.167	0.202	0.167	0.155	0.060	0.024	[1]

$$\overline{p} = 2/84;\ \underline{p} = 84/84$$

Combinatorial p–value $2/84 = 0.024 < 0.025$

Conclusion

Properties of Sampling Distribution

Mean = mean of the reference population;
standard–deviation $= \frac{v}{n_c} \times \frac{n-n_c}{n-1}$

The test–value is the deviation of the observed mean to the reference mean divided by te standard deviation SD.

Approximate typicality test:
Sampling distribution of Z is approximately a standard normal distribution (with mean 0 and standard deviation 1).

Application in MCA: typicality of class c (for mean)

Reference population: projected cloud of n active individuals onto a principal axis (with mean 0 and variance λ).
Group of observations: a subcloud of interest of size n_c.

Typicality of the group of observations (for the mean), with size n_c and coordinate $\overline{y}^c$ of the modality mean-point

Test statistic (variance $= v = \lambda$)
test–value is $T = \frac{\overline{y}^c}{\lambda}\sqrt{n_c \times \frac{n-1}{n-n_c}}$

Hypergeometric Typicality Test for a Relative Frequency

n observations, a observations possess a character of interest
hence $f_{\text{obs}} = \frac{a}{n}$
Reference population: size N, frequency $\varphi_0 = A/N$
Test statistic: F
Number of samples with $(F = a/n)$: $\binom{A}{a} \times \binom{N-A}{n-a}$

Observed upper level $p_{\text{sup}} = \dfrac{\sum_{a'=a}^{n} \binom{A}{a'} \times \binom{N-A}{n-a'}}{\binom{N}{n}}$

Remarks on Combinatorial inference

Typicality test and descriptive statistics
From typicality tests to frequentist inference
Summarizing:
Conceptually, combinatorial inference is a direct extension of descriptive statistics.

Combinatorial inference is the first stage of Inductive Data Analysis.

Geometric Typicality Tests

Paradigmatic situations
Drug effect —Target paradigm
Comparing the mean point of a cloud to a reference point

Permutation space
Combinatorial p–value
Conclusion

VI.5 Homogeneity Permutations Tests

Homogeneity situation
Pedagogy — Visual acuity — members of the Governing Council of the ECB
Comparing the mean points of subclouds

Experimental designs
Independent groups design (or between–subjects design):
each subject is assigned to only one condition of the independent variable. (see, for instance, the *Pedagogy* situation).

Repeated measures design (or within–subjects design):
each subject is assigned to each condition.

Steps of homogeneity tests

1. *Permutation set*
   The permutation group is applied to the baseline data set in order to generate all *possible data sets* that have the same *design structure* as the observed one.
2. *Combinatorial p–value*
   A statistic is chosen and then is calculated for each possible data set as well as for the observed one.
   The proportion of possible data sets for which the value of the statistic is more extreme than, or as extreme as, the observed one defines the combinatorial p–value.
3. *Conclusion*
   Given a reference level α (called α–level), we state the conclusion as:
   - if $p \leq \alpha$, the groups are said to be *heterogeneous* at level α, for the property of interest;
   - if $p > \alpha$, the groups cannot be declared heterogeneous at level α.

References

- Le Roux B., Bienaise S & Durand J-L. (to appear) *Combinatorial Inference in Geometric Data Analysis*. Chapman & Hall/CRC.
- Bienaise S., Le Roux B. (2018) Combinatorial typicality test in Geoemtric Data Analysis, *Statistica applicata – Italian Journal of Applied Statistics*, Vol 29 (2-3) `http://sa-ijas.stat.unipd.it/en/Vol29_2-3it.html`
- Le Roux B., & Rouanet H. (2004). *Geometric Data Analysis; From Correspondence Analysis to Structured Analysis*. Dordrecht: Kluwer (chapter 8 p.297-332, chapter 9 p.365-394).
- Rouanet H., Bernard J-M., Bert M-C., Lecoutre B. & M-P., Le Roux B. (1998,2000). *New Ways in Statistical Methodology: From significance Tests to Bayesian Inference*, Peter lang, Bern.