VI — Inductive Data Analysis

Brigitte.LeRoux@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~lerb/
rouanet@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~rouanet/

1 From description to inference

• *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.).

They attempt to extend descriptive conclusions.

1) They depend on sample size.

2) They lead to *inductive conclusions*.

Statistical modeling: as assumption-free as possible.

1) Instead of normal modeling,

prefer *combinatorial framework*.

2) Instead of general modeling (e.g. "general linear model")

prefer *specific modeling*,

i.e. put the statistical model on the *specific data set* relevant to the hypothesis of interest.

2 Significance Tests in MCA

Problem: for an axis, compare the mean of subcloud k (subcloud of the individuals who have chosen modality k) to 0 (overall mean).

- *Test-values* (combinatorial framework)

Test statistic: $T_k = \sqrt{n-1}\; y^k \sqrt{n_k/(n-n_k)}$

with n size of the overall cloud; n_k size of subcloud k; y^k coordinate of modality k on axis.

For large n and n_k/n far from 0 and 1: If $|T_k| \geq z_\alpha$, the mean of subcloud k differs significantly from 0 (the overall mean) at level α, hence the *conclusion*: for the axis, with respect to the mean, subcloud k is *atypical* (at level α) of the overall cloud of individuals.

Property: $T_k^2 = (n-1)\;\cos^2\theta_k$ ($\cos^2\theta_k$ is the Qlt of modality k)

Culture example. For axis 1, compare the mean of each age class to the overall mean.

Test values. 16.3 ($p < .001$); 5.7 ($p < .01$) ; 1.5 NS;

-0.4 NS ; -6.5 ($p < .01$); -17.8 ($p < .001$)

Conclusions

• For Axis 1, with respect to the mean, Class 1 is highly atypical of the whole cloud, on the right side of axis; Class 6 is highly atypical on the left side.

• Class 2 is fairly atypical of the whole cloud, on the right side of axis; Class 5 is fairly atypical on the left side.

• One cannot assert that Class 1 is atypical of the whole cloud on the right side of axis; that Class 6 is atypical on the left side.

• *t–tests* (traditional sampling framework)

Specific statistical modeling: Each age class is a random sample of the corresponding population age class. True means are the means of classes in the population on axis 1.

Results. 18.2 ($p < .001$); 6.0 ($p < .01$) ; 1.6 NS;

-0.4 NS ; -6.4 ($p < .01$); -16.6 ($p < .001$)

Conclusions

• One is nearly certain that the true mean on axis 1 of Class 1 is on the right side; that the true mean of Class 6 is on left side.

• One is fairly certain that the true mean of Class 2 is on the right side of axis 1; that the true mean of Class 5 is on left side.

• One cannot ascertain that the true mean of Class 3 is on the right side of axis 1; that the true mean of Class 4 is on left side.

Remarks

- Both tests lead to concordant conclusions.
- *The two Warnings about significance tests*

Warning 1. Evidence of effect (statistically significant) is not proof of large effect (especially for a large sample).

Warning 2. No evidence of effect (non-significant) is not proof of no effect, or even of small effect (especially for a small sample).

(Traditional inference handles poorly the smallness hypothesis)

Confidence intervals for the true means at level .05 (traditional sampling framework):

$[+0.34; +0.43]$; $[+0.08; +0.15]$; $[-0.01; +0.07]$

$[-0.06; +0.04]$; $[-0.24; -0.13]$; $[-0.46; -0.36]$

Bayesian framework

The Bayesian framework allows assessing the probabilities (i.e. measures of uncertainty) of the hypotheses of interest. Assuming "noninformative" (i.e. neutral) prior distributions, the following Bayesian reinterpretations hold for the inference on means:

1) Observed significance levels (p-values): If p denotes the usual two-sided observed significance level, the probability that the true mean lies on the side of the observed mean is $1 - p/2$.

2) The probability that the true mean lies inside the observed confidence interval at level α is $1 - \alpha$.

3 Confidence ellipses

For the modality mean point k in a principal plane, the approximate confidence ellipse at level α is the inertia ellipse with coefficient kappa

$\kappa = \sqrt{\chi^2_\alpha}/\sqrt{n_k}$

where χ^2_α is the α upper value of the χ^2 distribution with 2 d.f.

For $\alpha = .05$, $\chi^2_\alpha = 5.991$. The confidence ellipse at level .05 can be obtained by shrinking the concentration ellipse (i.e. inertia ellipse with $\kappa = 2$) by the factor $\sqrt{5.991}/2\sqrt{n_k} = 1.22/\sqrt{n_k}$.

The greater the size class, the smaller the κ coefficient, the smaller the ellipse.

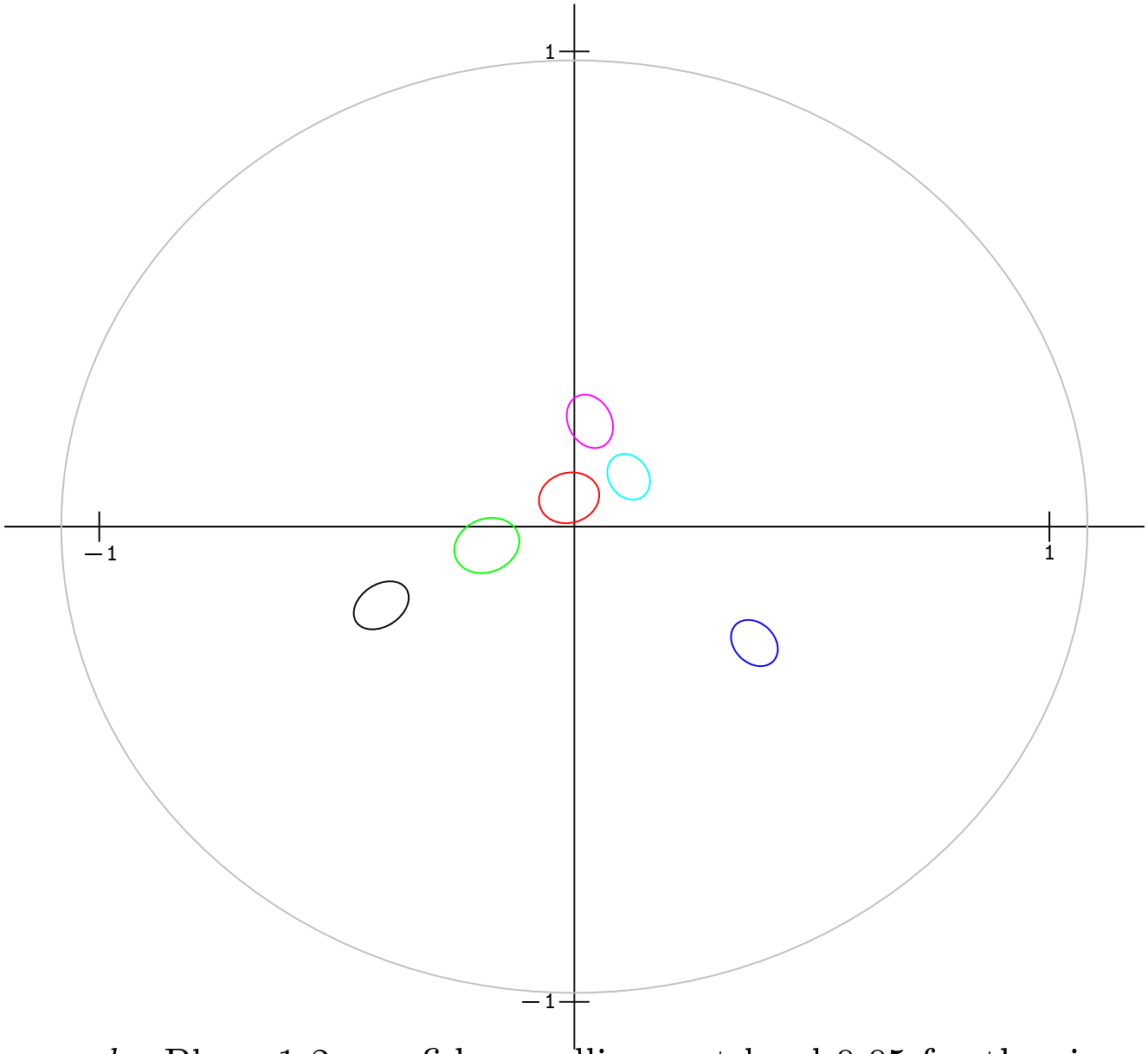
Culture example: Confidence ellipses in plane 1-2 for the mean points of the 6 age classes

κ coefficients are: $0.12; 0.10; 0.11; 0.12; 0.14; 0.11$

For each k, the confidence ellipse tells us about the location of the true mean point of class k: The greater the size class n_k, the more certain we are about the location of the true mean point around the observed mean point.

Bayesian reinterpretation of confidence ellipes

Again assuming "noninformative priors", confidences can be interpreted as probabilities. For modality k, the probability that the true mean point lies inside the ellipse at level α is equal to $1 - \alpha$.



Culture example. Plane 1-2: confidence ellipses at level 0.05 for the six age classes

References

— 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).