Comparing partitions of different units based on same questionnaire

Genane Youness[1] and Gilbert Saporta[2]

[1]Department of statistics, ISAE, CNAM-Lebanon, CEDRIC/CNAM,
BP 11-4661, Beirut, Lebanon, gyouness@cnam.fr

[2]Chaire de Statistique Appliquée- CEDRIC/CNAM,
292 Saint Martin roads, 75003 Paris, France, gilbert.saporta@cnam.fr

Outline

- Introduction
- Concept for comparing two partitions having the same variables.
- How to define the null Hypothesis: “ the partitions are identical”.
- Latent profile model.
- Association Indices: Rand, Kappa, Redundancy indices.
- projecting partition's Procedure.
- Algorithm for projecting partitions.
- Application on simulated and real data.
- Conclusion

Introduction

Two partitions

- Same individuals
- Different variables

- Différent individuals
- Same variables

- Same individuals
- Same variables

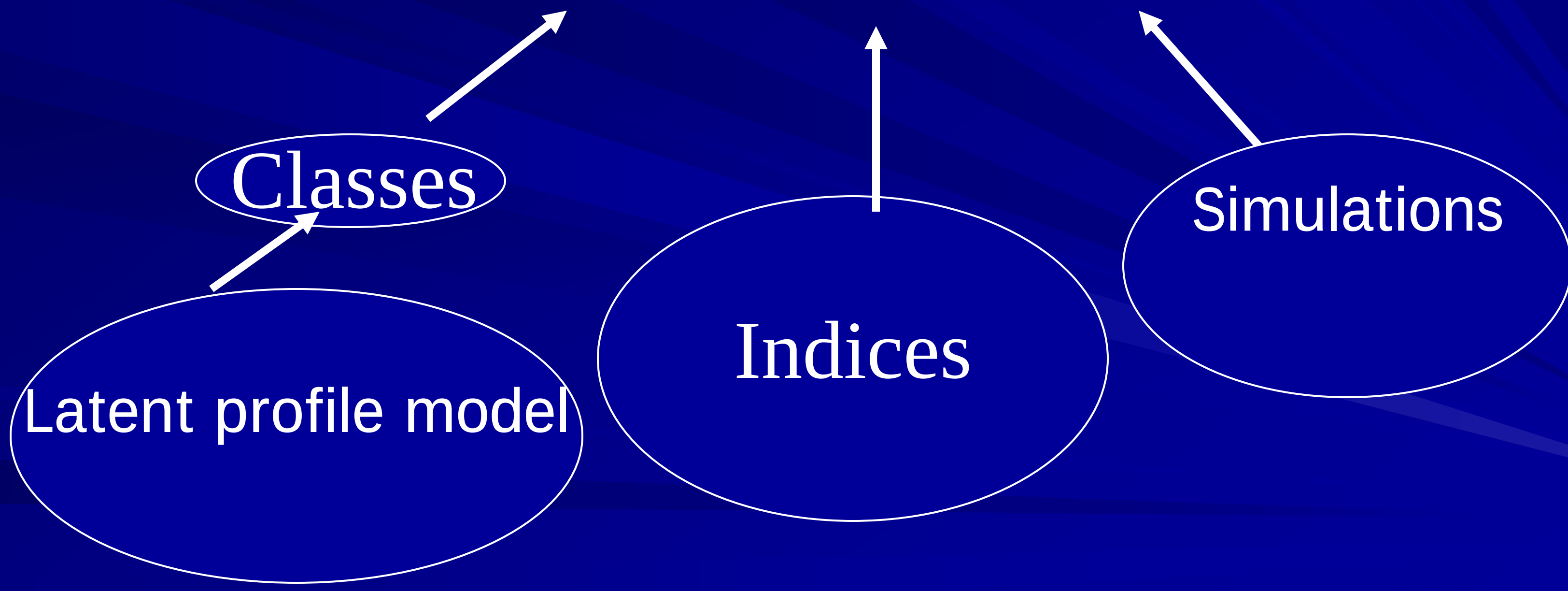
Comparing two partitions

- Two partitions of the same variables: two sets of individuals, same occasion… Are they significantly different?
- Compute an association measure *M* and its critical value.
- A probability distribution for *M* is found under the hypothesis of identical partitions?
- The H_0 is rejected as soon as *M* < critical value.

Concept for comparing two partitions

H_0: « Two partitions are identical »

H_1: « They are not identical »



How to define H_0: « Two partitions are identical »

- Two partitions are close to each other if observations come from the same underlying common partition P,
- The two observed partitions are noisy realisations of the common one.
- Model for a common partition: latent profil model or mixture of probability distributions.

Latent profile model

- Particular mixture model based on hypothesis of locale independency conditioned by latent classes

$$f(\mathbf{x}) = \sum_k \pi_k \prod_j f_k(x_j / k)$$

- Serve to generate partitions
- Used by Green et Krieger 1999.
- Observed variables are numeric and latent variables are qualitative. (Bartholomew and Knott 1999)

Association indices

Notations

P_1 et P_2 partitions of the same individuals with *p* et *q* classes

- K_1, K_2 : disjunctives tables (n,*p*) et (n,*q*)

$k_1(i,P_1)=1$ if $i \in P_1$; 0 otherwise

- $c^{k}_{ii'} = \begin{cases} 1 & \text{if i and i' are in the same class of } P_k, \\ 0 & \text{otherwise} \end{cases}$ $\quad C_1 = K_1 K_1' \quad C_2 = K_2 K_2'$

- Contingency table N (p,q) in term of n_{uv}

$N=K_1' K_2$

Rand index

- Rand's raw index is the proportion of agreements

$$R' = \frac{2\sum_u \sum_v n_{uv}^2 - \sum_u n_{u.}^2 - \sum_v n_{.v}^2 + n^2}{n^2}$$

- Corrected rand index R_C (Hubert & Arabie 1985): with an hypothesis of random partitions

  (R_C could be <0)

$$R_C = \frac{R - R_{esp}}{R_{max} - R_{esp}} = \frac{n^2 \cdot \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}{\frac{1}{2} \cdot n^2 \cdot (\sum_u n_{u.}^2 + \sum_v n_{.v}^2) - \sum_u n_{u.}^2 \cdot \sum_v n_{.v}^2}$$

- Asymmetric Rand R_A (Chavent 2001):
  - P_1 is more 'refined' than P_2
  - measure the inclusion of P_1 in P_2

$$R_A(P_1, P_2) = \frac{n^2 + \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2}{n^2}$$

Kappa coefficient

- (Cohen 1960) , compute nominal scale agreement between two raters defined as “ the proportion of agreement after chance agreement is removed from consideration”.

$$\kappa = \frac{n \sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.} n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.} n_{.i}}$$

- Condition for use:
  - Two partitions having the same number of clusters *p=q=k.*
  - Identify the classes from Kappa maximum because the labeling of clusters is totally arbitrary.

Redundancy index

- RI (Stewart and Love 1968)
  - $W_{ij} = X_i X'_j$
  - Is a weighted average of the squared multiple correlation between components of X_1 and X_2

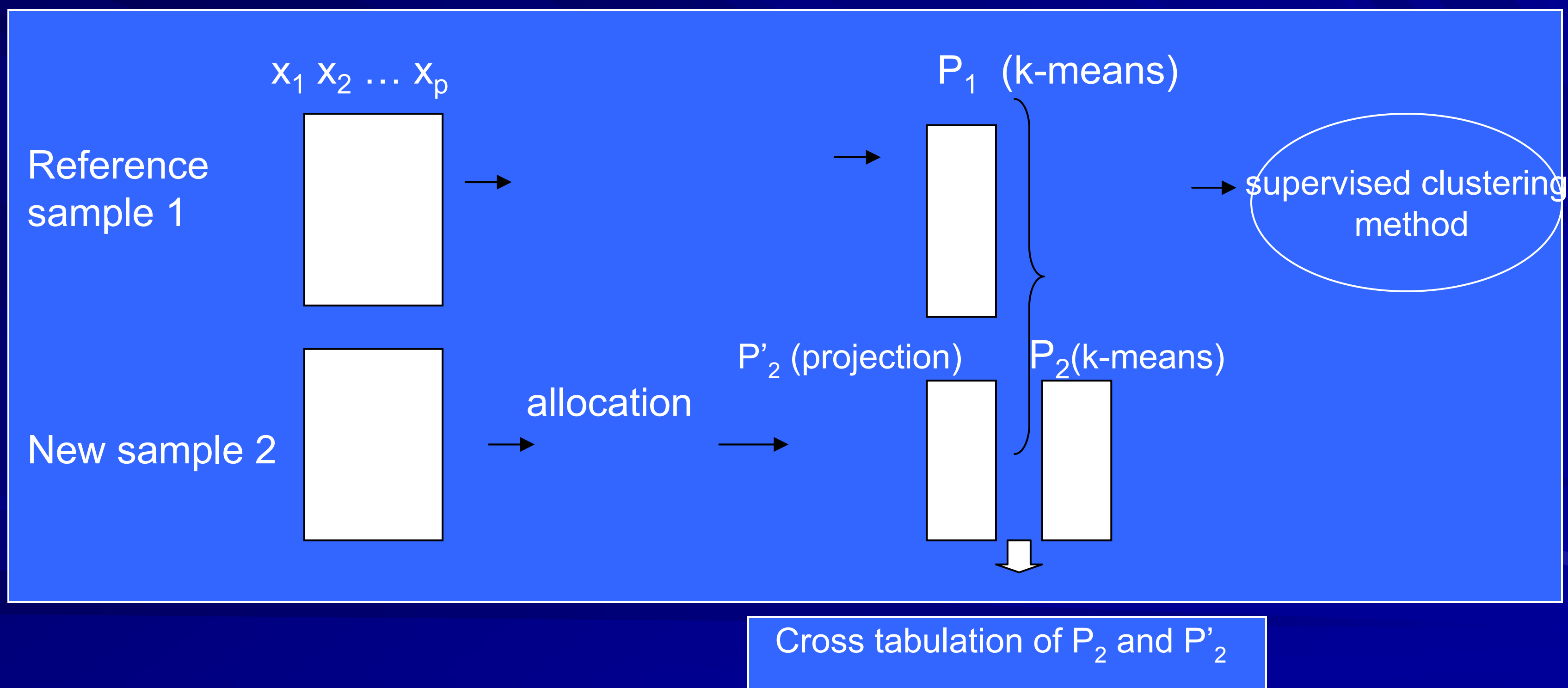
$$RI(X_1, X_2) = \frac{\mathrm{trace}(W_{12} W_{22}^{-1} W_{21})}{\mathrm{trace}(W_{11})}$$

- If X_1, X_2 indicator variables,
  - $RI = \tau_b$ (Goodman and Kruskal 1979)

$$\tau_{bP_2/P_1} = \frac{\sum_u \sum_v \frac{n_{uv}^2}{n.n_{u.}} - \sum_v \left(\frac{n_{.v}}{n}\right)^2}{1 - \sum_v \left(\frac{n_{.v}}{n}\right)^2}$$

Procedure of projecting a partition on another one

- Projecting a partition upon a reference one consists in allocating the units of the second set in the clusters defined by the reference partition using some discriminant analysis technique



Algorithm for projecting partitions

- Generate the sizes n_1, n_2,….,n_k of the clusters according to a multinomial distribution $M(n, \pi_1, \pi_2,…., \pi_k)$.
- For each cluster, generate ni values from a random normal vector with p independent components. The first data set I_1 of n_1 units is obtained.
- The same independent normal variables are used to generate the second data set I_2 of n_2 units. The initial data $I=I_1 + I_2$. Obtain P_1.
- Classifying I_2 into the clusters of P_1 by linear discriminant analysis to obtain P'_2.
- Computing association indices for P_2 and P'_2 of the same set I_2.
- Randomly permuting the observations of I.
- Again N times **to find the empirical sampling distribution of the association indices**.

Simulation

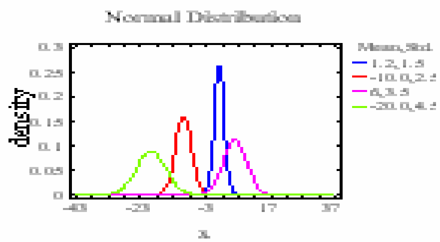
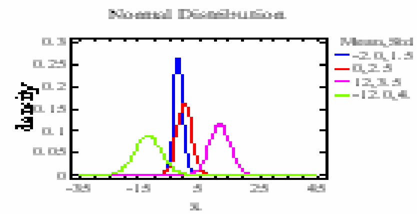
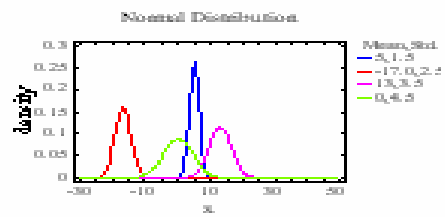
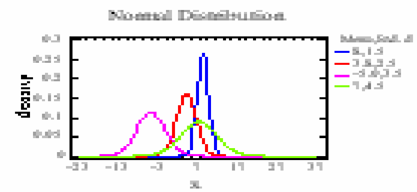
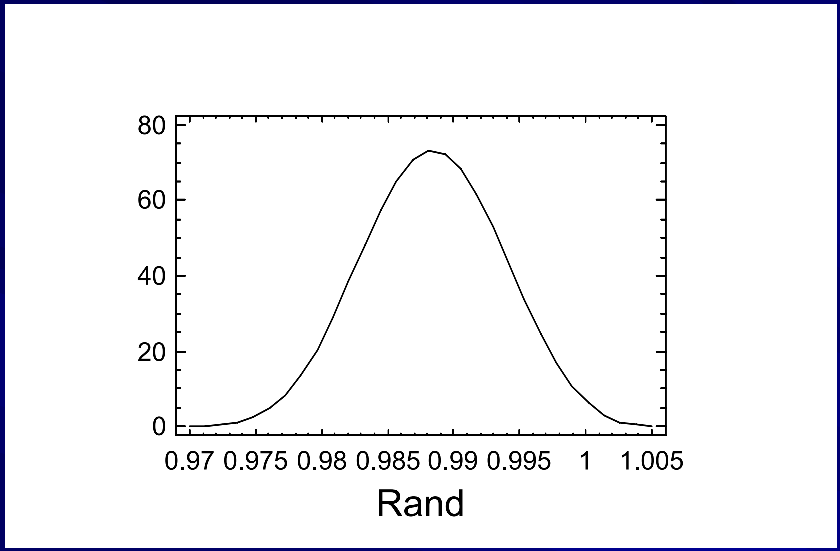
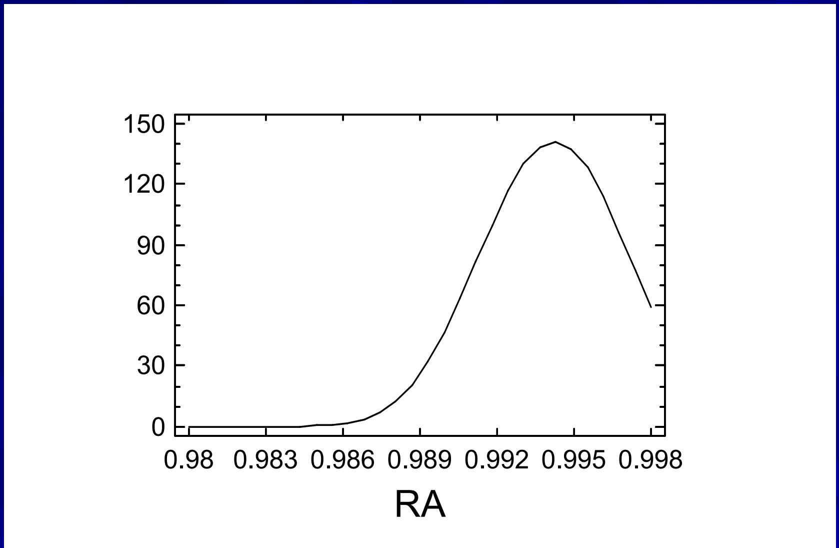
Classe 1	X1 N(1.2,1.5) X2 N(-10,2.5) X3 N(6,3.5) X4 N(-20,4.5)	
Classe 2	X1 N(-2,1.5) X2 N(0,2.5) X3 N(12,3.5) X4 N(-12,4.5)	
Classe 3	X1 N(5,1.5) X2 N(-17,2.5) X3 N(13,3.5) X4 N(0,4.5)	
Classe 4	X1 N(8,1.5) X2 N(3.8,2.5) X3 N(-5,3.5) X4 N(7,4.5)	

Tableau 2. *Les distributions par classe*

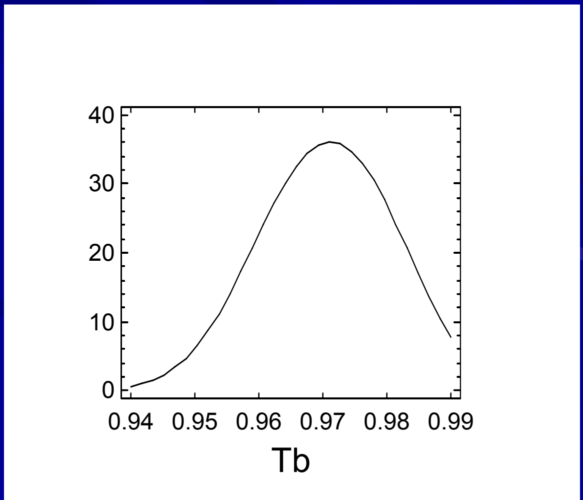
Simulation (1)

- **Identical partitions with same number of clusters**





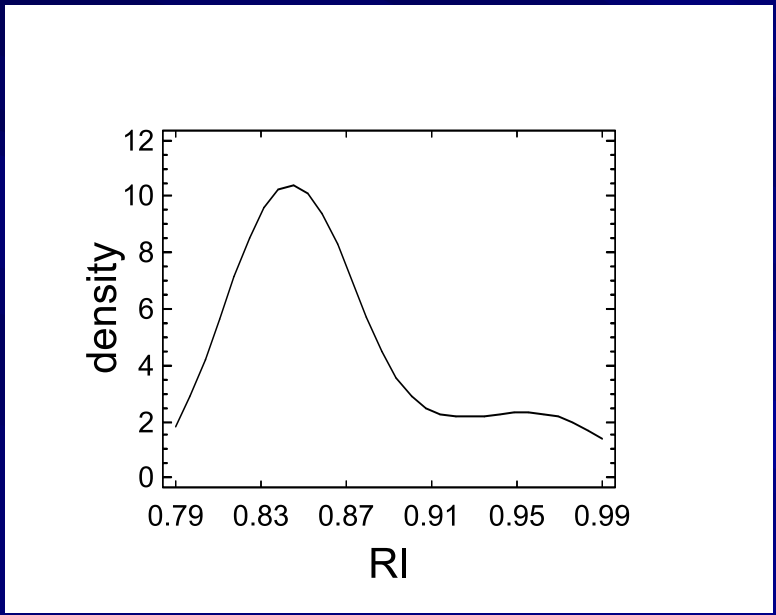
	Mean	Error type	Lower Limit	Upper Limit
Rand	0.988367	0.000163061	0.988047	0.988687
RA	0.994158	0.00008495	0.993991	0.994325
Tb	0.970632	0.000405929	0.969834	0.97143

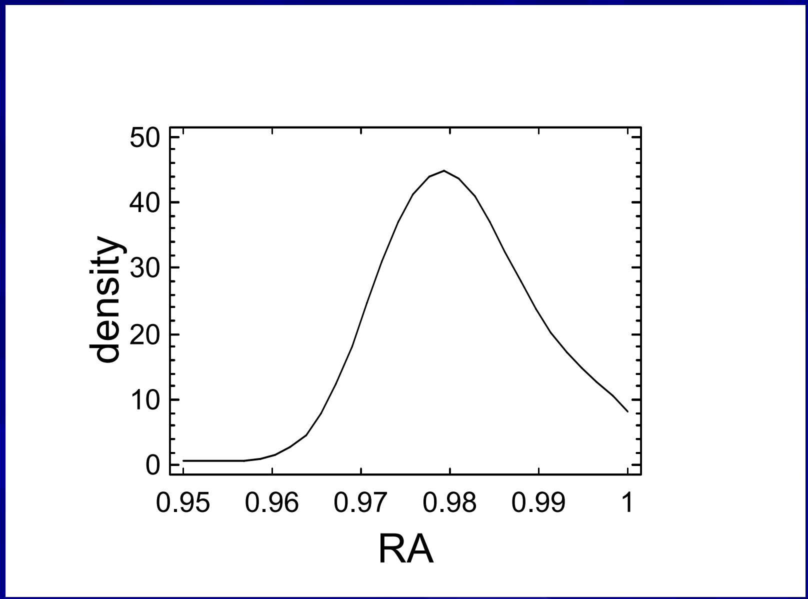


Simulation (2)

- **Identical partitions with different number of clusters**

5 classes, N=500, P'2 (projection on P1) 5 clusters, P2 (k-means) 3 clusters.





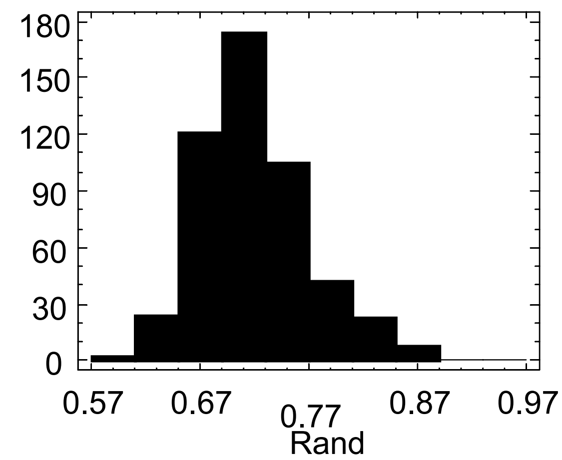
- RI vary 0.79915 to 0.988053, E(RI)= 0.867616.
- RA vary from 0.95 to 0.99824, E(RA) =0.98129.

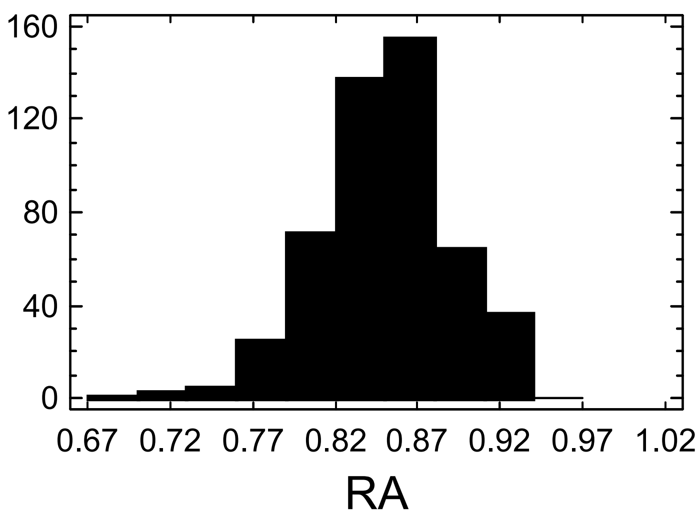
Real data (1)

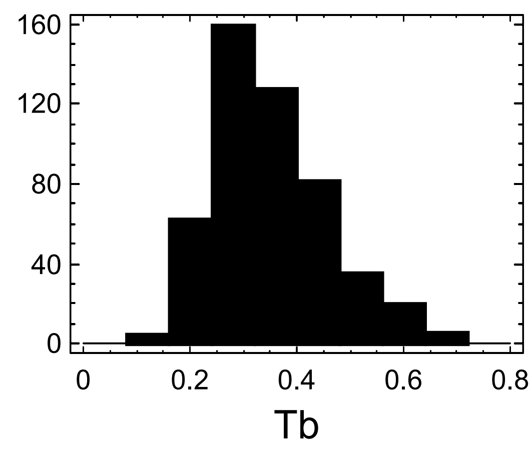
- Survey about the condition of life and the aspirations of French people (Lebart 1987).
- Goal is to compare men’s and women’s partitions.

n=624, p=14, 500 times.

The upper 5% fractil under H0 is equal to 0.721 for Rand, 0.85 for RA and 0.35 for Tb

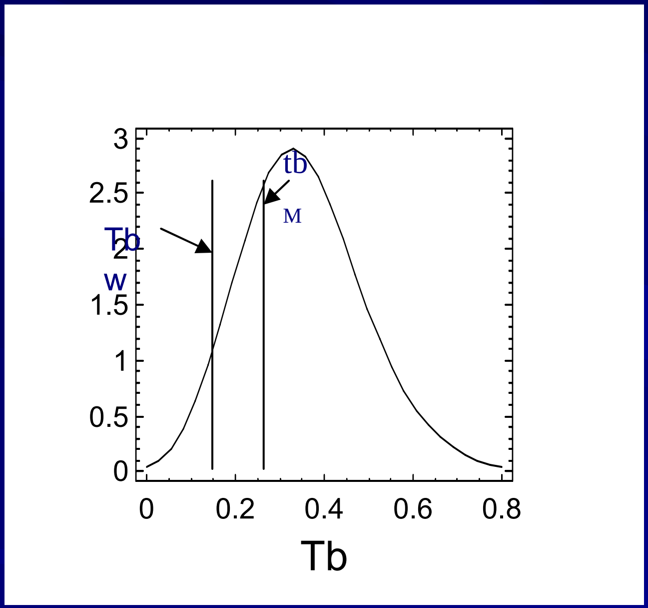


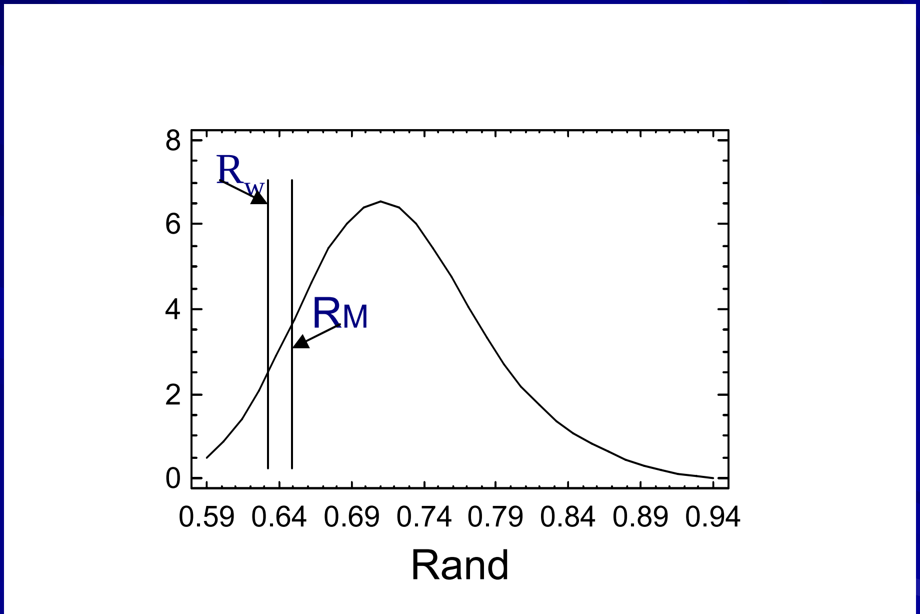




Real data (2)

- To compare men and women's partitions, sorting sex, dividing data according to sex.
- Women: Tb=0.185, Rand=0.6134. Men: Tb= 0.2582, Rand= 0.6466.





- Men and women's partitions are not considered identical since the indices have values much lower to their critical values.

Conclusion

- Conclusion.
  - Method of comparing partitions coming from two set of objects with same variables based on a projection of partitions, through supervised clustering method.
  - Critical values for distribution of indices depending on the k, n and the separation of classes.
  - Applications have proved the feasibility of our approach.

- Further studies are needed:
  - Find universal critical values.
  - Look at the meaning of classes in terms of the variables (external and internal information).