

DISCRIMINATION (K7)

(07 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Dans certaines **situations statistiques**, il peut être utile d'associer une **unité statistique** à une **catégorie** préétablie d'unités du même type, cette catégorie faisant partie d'un ensemble donné de catégories. Il faut donc pouvoir reconnaître une « **famille** » à laquelle (ou un « groupe » auquel) peut appartenir l'unité.

(i) Une unité n'est pas imputée à une catégorie de façon arbitraire : l'imputation se base sur des **critères** d'appartenance. Ces derniers sont des **variables**, parfois appelées « **descripteurs** », caractérisant les unités afin de les comparer et de les rapprocher (cf **Statistique descriptive**). Ces variables peuvent être considérées comme des **variables aléatoires**. Elles doivent être communes à toutes les unités, mais aussi être observables (cf **observabilité**, **inobservable**). Ce sont donc des **variables qualitatives** ou des **variables quantitatives**.

(ii) On considère une **partition** en k classes d'une **population** donnée (ou, alternativement, k populations données d'**unités statistiques**), sur laquelle (lesquels) on peut observer divers descripteurs constituant une liste L .

Le **statisticien** observe une nouvelle unité à travers ses descripteurs, dont la « liste » U peut :

- (a) soit être incluse dans L , auquel cas les comparaisons se font sur la liste U ;
- (b) soit contenir L , auquel cas les comparaisons se font sur la liste L .

On appelle **problème de discrimination**, ou **problème de classement**, ou encore **problème d'affectation**, le problème de détermination de la classe de la partition (resp de l'une des k populations) à laquelle appartient cette unité (cf **analyse discriminante**). C'est donc toujours, au mieux, sur la liste minimale $M = \min(L, U)$ que la discrimination s'effectue ; ce peut être sur une sous-liste de M .

(iii) Les données d'un problème de discrimination sont donc généralement :

(a) d'une part, les k classes de la partition considérée (resp de k **échantillons** resp extraits des k populations), ainsi que des observations correspondantes des L descripteurs sur les unités ;

(b) d'autre part, l'**unité supplémentaire** (ou **unité « nouvelle »**) ainsi que l'observation correspondante de ses propres descripteurs U .

(iv) Le principe de la discrimination est basé sur l'évaluation d'une « **distance** » entre l'unité nouvelle et chacune des classes de la partition (resp chacun des k échantillons) afin d'affecter cette unité à l'une des k populations correspondantes.

Ce problème peut s'interpréter comme un problème de **reconnaissance des formes**.