VALIDATION CROISÉE (F, G)
(20 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'important concept de **validation** concerne la notion de **représentation statistique**. On considère généralement qu'un **modèle** est « valide » ssi, à la fois :

(a) son **estimation** est optimale, et sa qualité la meilleure possible (cf **adéquation**, **ajustement**, **qualité d'un ajustement**). Autrement dit, l'adéquation du modèle aux **données** (ou aux **observations**) est élevée ;

(b) il permet d'attester la vraisemblance d'un jeu d'**hypothèses statistiques** faites par l'**homme de l'art** ou par le **statisticien** (cf **théorie des tests**, **test d'hypothèses**). Autrement dit, divers tests statistiques tendent à valider des **caractéristiques** importantes du modèles : eg les valeurs de ses **paramètres**.

Un **« modèle validé »** est considéré comme étant le « vrai » modèle qui gouverne les observations, ie qui en est à l'origine. Cette attitude, cruciale, autorise diverses latitudes :

(a) **extrapolation** ou **prévision** : le modèle peut, sous certaines conditions, être appliqué à d'autres **unités statistiques** (eg individus), à d'autres périodes (eg le « futur ») ou à d'autres espaces (eg des « zones géographiques ») ;

(b) **décisions** ou **actions** : celles-ci portent sur diverses variables d'action et de contrôle, dont doivent résulter divers résultats escomptés.

(i) On considère un **modèle** appliqué à un jeu de **données** (eg **échantillon**), et l'on cherche à apprécier la qualité ou l'**efficacité** d'une **procédure statistique** (estimation, test, prévision, etc) fondée sur ce modèle.

Dans certaines situations, il est possible d'appliquer la même procédure à un (ou plusieurs) nouveau(x) autre(s) jeu(x) de données. On compare alors les « performances » entre résultats (estimateurs, statistiques de test, prédicteurs, etc) obtenus : **variabilité** ou **dispersion**, **efficacité**, **propriété asymptotique**, etc.

La démarche précédente est appelée **méthode de validation croisée**, ou **procédé de validation croisée** (cf aussi **résistance**). Le **schéma** est donc (avec un modèle M donné et S échantillons A_s , s = 1 ,..., S) :

(1) $M \leftrightarrow A_1$,..., $M \leftrightarrow A_S$.

(ii) Ce type de méthodes est à distinguer (et à comparer) aux méthodes d'étude de la **robustesse**, dans lesquelles le modèle change mais les données sont les mêmes (cf aussi **courbe d'influence**). Le schéma est donc (R modèles M_r et un échantillon A) :

(1) $M_1 \leftrightarrow A$,..., $M_R \leftrightarrow A$.