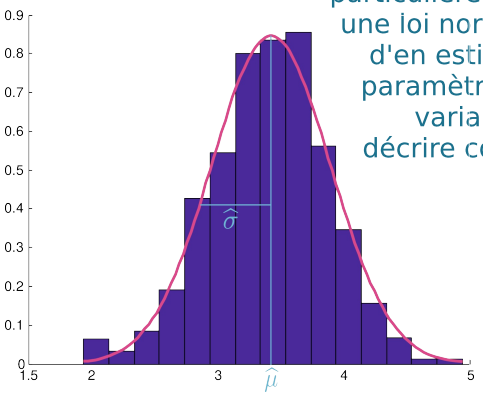
Estimation non-paramétrique

La manifestation du hasard est décrite mathématiquement par des ***lois de probabilités***. En statistique, la loi régissant le phénomène que l'on observe est inconnue, et on cherche à retrouver des caractéristiques de cette loi à partir des observations effectuées.

Paramétrique vs non-paramétrique

Dans le cadre de la ***statistique paramétrique***, on suppose que la loi recherchée a une forme particulière (par exemple une loi normale). Il suffit d'en estimer quelques paramètres (moyenne, variance...) pour la décrire complètement.



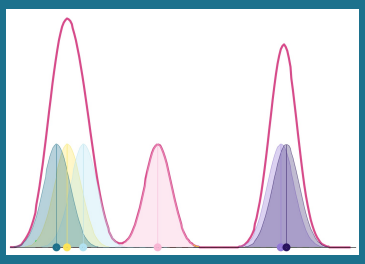
On a pesé 1100 bébés à la naissance. La forme de l'histogramme obtenu suggère un ajustement par une loi normale (courbe rose), dont on estime la moyenne μ et l'écart-type σ.

Si l'on n'a pas d'a priori sur la forme de la loi inconnue, on cherche à estimer des fonctions, et non plus des paramètres. C'est l'objet de la ***statistique non-paramétrique***, qui nécessite moins de connaissances préalables de la loi. En contrepartie, il faut plus de données pour obtenir une précision d'estimation équivalente à celle du cadre paramétrique.

La forme de l'histogramme des longueurs de pétales de 150 iris ne correspond pas à une loi connue : un estimateur non-paramétrique est plus approprié. La courbe rose représente un estimateur à noyau.

Estimation à noyau de la densité

Dans un histogramme, on regroupe les observations en classes dont on représente la fréquence. Celui-ci s'interprète comme une fonction, constante sur chacune des classes, qui estime la densité de la loi. Mais cette fonction est discontinue.

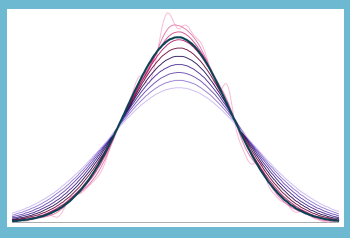


L'***estimation à noyau*** consiste à remplacer les barres de l'histogramme par des bosses, typiquement des courbes en cloche, centrées sur chacune des observations. L'estimateur est formé par la somme (ou plutôt la moyenne) des courbes en cloche. On obtient cette fois une fonction continue.

$$\widehat{f}_h : x \longmapsto \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Compromis biais-variance

Toute la difficulté du problème réside dans le choix de la largeur des cloches : des cloches trop étroites entraînent une trop grande variabilité, tandis que des cloches trop larges induisent un trop mauvais ajustement.
La largeur optimale, appelée ***oracle***, est inaccessible : elle dépend de la loi (inconnue) des observations. La ***sélection de modèle*** propose des méthodes de calibration permettant une adaptation à chaque jeu de données. La pertinence des méthodes est justifiée par un résultat mathématique, les ***inégalités oracles***.



$$\mathbb{E}\left[\left\|\widehat{f}_{\widehat{h}} - f\right\|^2\right] \leq C_1 \min_{h \in \mathcal{H}} \mathbb{E}\left[\left\|\widehat{f}_h - f\right\|^2\right] + \frac{C_2}{n}$$

LMRS
Laboratoire de Mathématiques Raphaël Salem
cnrs
UNIVERSITÉ DE ROUEN NORMANDIE