Analyse de données à l'aide du logiciel R

Vincent Vandewalle
Maître de conférences en mathématiques appliquées
Université Lille 2

IUT de Roubaix, Département STID

Jeudi 10 décembre 2013

- Quelle est la population étudiée ?
  - Les lycéens de terminale ?
  - Les lycéens du lycée Saint Rémi de Roubaix ?
- Quelles sont les données disponibles ?
  - Les données sont-elles disponibles pour toute la population ?
  - Dispose-t-on seulement d'un échantillon ? Cet échantillon est-il "représentatif" ?

Quelle est la nature des variables ?

- Quantitative :
  - Discontinue (nombre d'enfants)
  - Continue (taille)
- Qualitative :
  - Nominale (Catégories Socio Professionnelles)
  - Ordinale (degré de satisfaction)

Analyse d'une variable quantitative : aspects descriptifs

- Représentations graphiques :
  - Histogramme (diagramme en bâtons si discontinue)
  - Courbes cumulatives (en escalier si discontinue)
- Calcul d'indicateurs :
  - Tendance centrale : moyenne, médiane, classe modale
  - Position : minimum, maximum, quartiles
  - Dispersion : étendue, variance, écart-type, écart inter-quartiles
  - Dispersion relative : coefficient de variation, écart inter-quartiles relatif

Analyse d'une variable qualitative : aspects descriptifs

- Représentations graphiques :
  - Diagramme en bâtons (dans l'ordre des modalités si ordinale, dans l'ordre des fréquences si nominale)
- Tableaux :
  - Fréquences cumulées croissantes et décroissantes (dans l'ordre des modalités si ordinale, dans l'ordre des fréquences si nominale)

Analyse d'une variable : aspects inférentiels

- Donner des réponses aux questions :
  - Quel candidat risque de remporter l'élection présidentielle ?
  - Quelle est la quantité moyenne versée par une machine à embouteiller par bouteille ?
  - La quantité moyenne versée par une machine à embouteiller est-elle supérieure à 1 litre par bouteille ?
- Outils utilisés :
  - Lois de probabilités usuelles
  - Approximations de lois

Les questions de la statistique : approches exploratoires

- Lien entre deux variables qualitatives
  - Y-a-t'il indépendance entre les deux variables ?
  - Quelles sont les modalités qui s'attirent ? Qui se repoussent ?
- Lien entre p ($p > 2$) variables qualitatives
- Lien entre deux variables quantitatives
- Lien entre p ($p > 2$) variables quantitatives
- Faire des classes à partir de p variables quantitatives ($p \geq 1$)

Les questions de la statistique : approches prédictives (1/2)

- Explication d'une variable quantitative par une autre variable quantitative

$$y = f(x)$$

- Une variable quantitative fonction de plusieurs variables quantitatives

$$y = f(x_1, x_2, \ldots, x_p)$$

- Une variable quantitative fonction d'une variable qualitative

$$y = f(x)$$

Les questions de la statistique : approches prédictives (2/2)

- Une variable quantitative fonction de p variables qualitatives

$$y = f(x_1, x_2, \ldots, x_p)$$

- Une variable qualitative en fonction de variable(s) qualitative(s) et ou quantitative(s)

$$y = f(x_1, x_2, \ldots, x_p, x_{p+1}, x_{p+2}, \ldots, x_{p+q})$$

Utilisation du logiciel R

- R : un logiciel statistique spécialisé
- Rcmdr : interface graphique de R de type clique-bouton
- Rstudio : environnement de programmation convivial de R