Cours 10 - Éléments empiriques (2/4) (20 mars 2024)

Alignement pédagogique

Mise en application

Objectifs visés et activités associées

Objectif général : 4. Appliquer les méthodes de base en sciences de l'information pour analyser des données

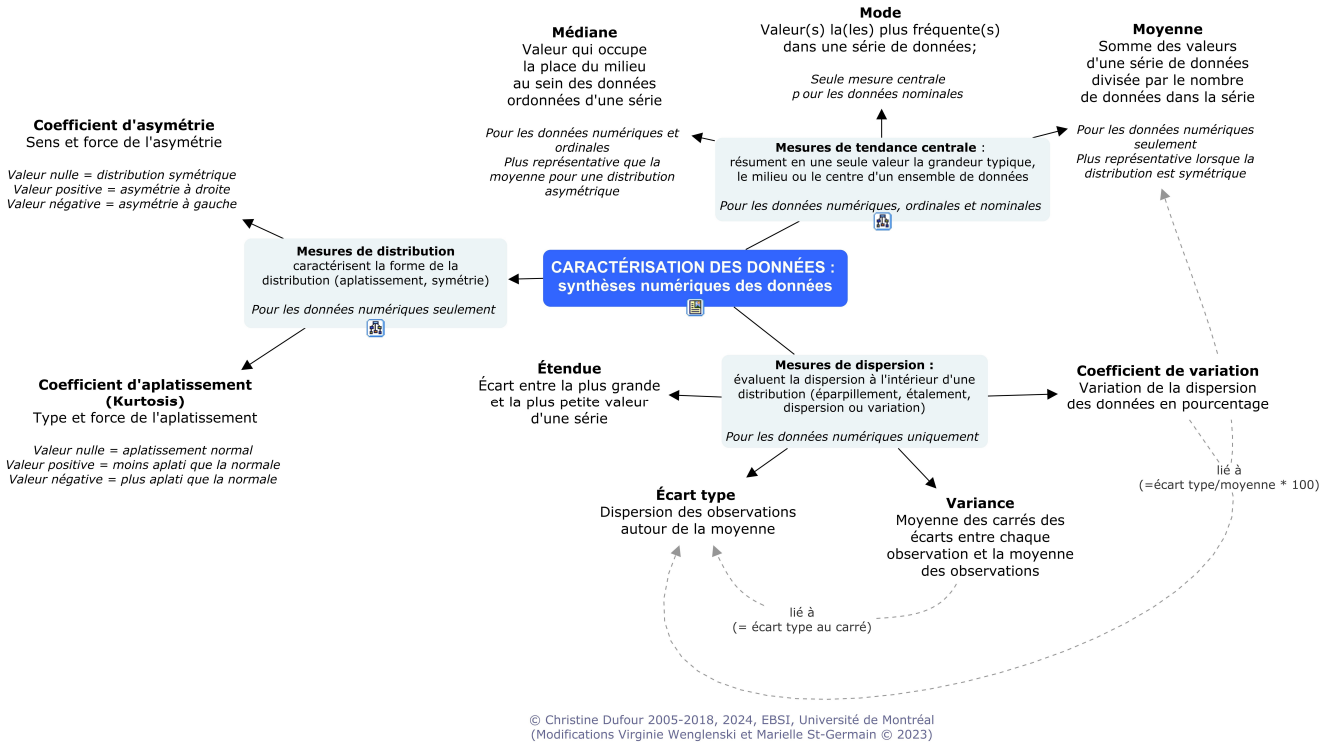
Objectif spécifique : 4b. Appliquer les méthodes statistiques de base pour analyser des données quantitatives

Activité : TP3 et Devoir 4

Cours 10 - Éléments empiriques (2/4) (20 mars 2024)

Cartes conceptuelles

Phase 3 : Éléments empiriques > Analyse des données > Analyses statistiques > Statistiques descriptives > Caractérisation des données



https://reseauconceptuel.umontreal.ca/rid=1YRFSB0P5-2DLML8K-9KS4/sci6007_c10_pe_analyse_statistique_descriptives_mesures.cm

Synopsis

La caractérisation des données passe par l'utilisation de mesures statistiques. Une mesure statistique (par exemple la moyenne ou l'étendue) permet de résumer, en une seule valeur, une facette d'un ensemble de données numériques. Trois facettes des données peuvent être représentées : (1) leur centralité, (2) leur distribution, et (3) leur dispersion. L'idée à retenir est, d'une par, que de n'utiliser qu'une seule facette est

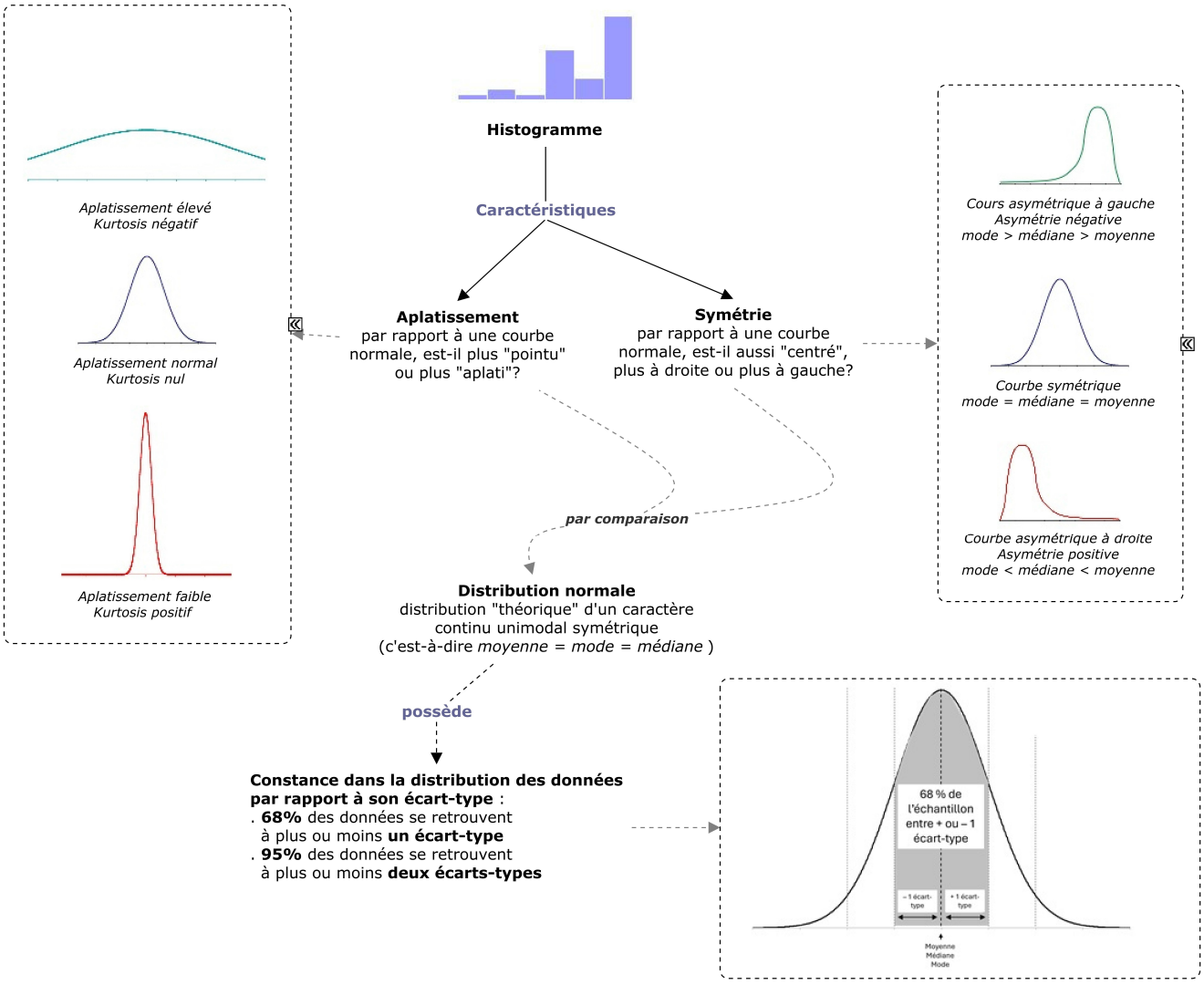
réducteur quant à la représentation obtenue des données, d'autant plus que, d'autre part, certaines mesures statistiques peuvent être plus ou moins représentatives selon la distribution des données. Il est donc important, lorsque l'on s'intéresse aux mesures statistiques, de les considérer comme un ensemble et non individuellement.

Les mesures de tendance centrale visent à représenter en une seule valeur la grandeur typique (mode), le milieu (médiane) ou le centre d'équilibre (moyenne) d'un ensemble de données.

Les mesures de dispersion, de leur côté, vont caractériser la dispersion des données, que ce soit la différence entre la plus grande et la plus petite valeur (étendue), la dispersion des données autour de la moyenne (écart type) (ou son carré, la variance), ou la variation en pourcentage de la dispersion (coefficient de variation). Ce dernier (coefficient de variation) est fort utile pour s'aider à interpréter l'écart type.

Les mesures de distribution vont décrire la forme de la distribution par rapport à une courbe normale sur deux aspects (aplatissement et la symétrie). Ces dernières s'interprètent plus facilement lorsqu'on les applique à un histogramme.

Phase 3 : Éléments empiriques > Analyse des données > Analyses statistiques > Statistiques descriptives > Présentation en graphique > Caractérisation d'un histogramme



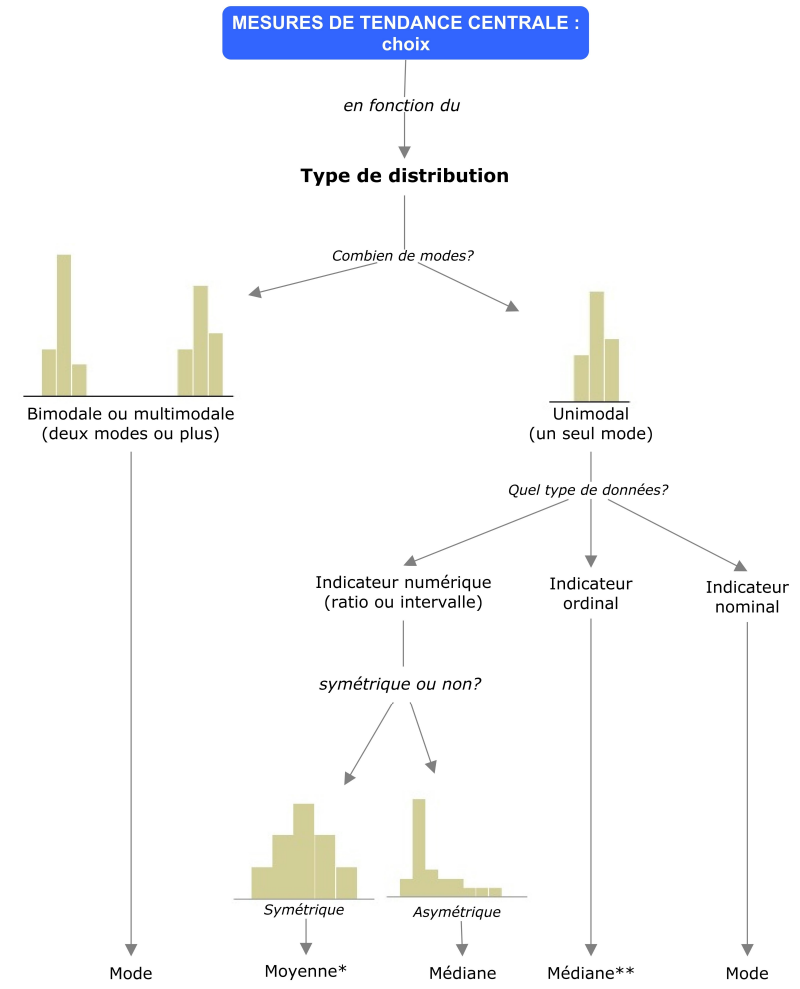
https://reseauconceptuel.umontreal.ca/rid=1YRFSB0P5-18X84RR-9KRL/sci6007_c10_histogramme_caracterisation.cmap

Synopsis

La forme de la distribution des données pour un caractère numérique, distribution souvent représentée par un histogramme, peut être décrite selon deux aspects, soit son aplatissement et sa symétrie. Le point de comparaison en ce cas est une distribution normale (en forme de cloche) qui se retrouve présente dans bien des phénomènes. La distribution peut être plus ou moins aplatie que la normale, en fonction de l'ampleur de la dispersion des données (des données avec beaucoup de dispersion ont habituellement un aplatissement plus grand que la normale). La distribution peut aussi être symétrique ou asymétrique. Une distribution asymétrique à gauche possédera quelques valeurs dans plus petites et un regroupement de valeurs plus grandes. Par exemple, s'il s'agit de la distribution des revenus dans un réseau de bibliothèques, une

distribution asymétrique à gauche signifie que quelques bibliothèques ont des revenus plus petits que les autres. L'asymétrie des données est importante à repérer comme elle aura une influence sur le choix des mesures de centralité à retenir.

Phase 3 : Éléments empiriques > Analyse des données > Analyses statistiques > Statistiques descriptives > Caractérisation des données > Choix des mesures de centralité



Source : Vaughan, 2001, p. 32 (notre traduction et adaptation)

** Dans une distribution symétrique, la moyenne étant proche du mode et de la médiane, ces derniers pourraient aussi être utilisés.*

*** Il est à noter que le mode pourrait aussi être utilisé.*

https://reseauconceptuel.umontreal.ca/rid=1YRFSB0P5-P7X8QC-9KS9/sci6007_c10_pe_analyse_statistique_descriptives_centralite_ch

Synopsis

Le choix des mesures de centralité repose sur le type de caractères. Un caractère ordinal, par exemple, pourra être représenté par la médiane principalement (on pourrait aussi exploité le mode). Un caractère nominal, pour sa part, ne peut qu'être caractérisé en termes de centralité que par le mode. C'est le caractère numérique qui peut être représenté par le plus de mesures de centralité, soit la moyenne, la médiane et le mode. Il faut cependant s'assurer de faire le bon choix!

Pour le caractère numérique, le choix repose sur la forme de sa distribution, telle qu'illustrée par un histogramme. Un caractère ayant une distribution bimodale ou multimodale - c'est-à-dire qu'il y a plusieurs

modes autour desquels se regroupent les données - est mieux représenté par le mode, la moyenne et la médiane ne voulant rien dire en ce cas. S'il possède uniquement un mode, il faut vérifier sa symétrie. Si la distribution est relativement symétrique, la moyenne et la médiane qui s'en rapprochera en ce cas sont les mesures à exploiter. Si la distribution est très asymétrique, la moyenne perd son sens et c'est la médiane en ce cas qu'il faut retenir.