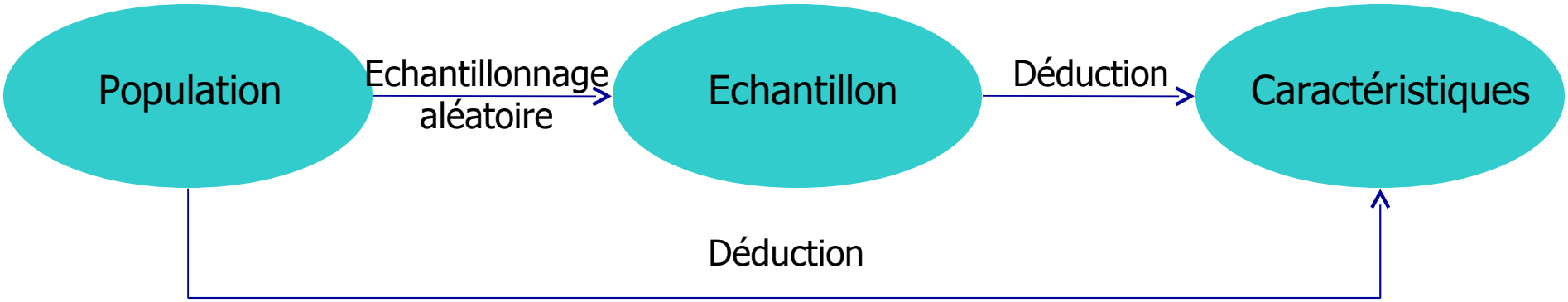
Statistique descriptive

- Objectif : Analyser les caractéristiques principales d'un ensemble (de grande taille)
- Définition : Méthode scientifique qui
  - recense un corpus de données chiffrées
  - analyse ses données pour les rendre intelligibles à l'observateur
- Exemple : études démographiques au XIX siècle

Concepts : population et échantillon

- En statistique, on utilise le terme population plutôt qu'ensemble.
- Les éléments de la population sont appelés individus ou unités statistiques.
- La population est étudiée à travers des caractères.
- Pour des raisons de coûts ou autres, le corpus porte souvent sur une partie de la population appelée échantillon.



Statistique descriptive et mathématique

- Statistique descriptive :
  - Description à travers des résumés chiffrés : moyennes, médianes, écarts-types, corrélations, …
  - Description à travers des résumés graphiques : histogrammes, diagrammes en bâton, mappings, ..
  - C'est l'objet de ce cours.
- Statistique mathématique :
  - Trouver des estimateurs non biaisés et efficients pour passer de l'échantillon à la population.
  - L'outil mathématique sous jacent est la théorie des probabilités : cours vu en Ing1-S2.
  - Ce cours sera vu en TC Ing2-S3.

Echantillonnage

- L'échantillonnage est l'ensemble des techniques permettant de prélever un échantillon dans une population.
- Chaque technique doit générer des échantillons représentatifs de la population
- Techniques les plus connues :
  - Echantillonnage aléatoire
  - Echantillonnage systématique
  - Echantillonnage stratifié.

Echantillonnage aléatoire simple

- Les individus de la population sont numérotés de 1 à N.
- Chaque individu a la même chance d'être choisi
- En notant alea() une fonction qui génère à chaque appel un nombre au hasard dans [0,1[, on forme l'échantillon de taille n comme suit :

```
Pour i ← 1 à n
    e[i] = E[alea() * N] + 1 // E[...] désigne la partie entière
Fin Pour
```

En toute rigueur, il faut adapter l'algorithme pour qu'un individu ne soit pas choisi plus d'une fois.

Echantillonnage systématique

- Les individus de la population sont numérotés de 1 à N.
- Les individus sont choisis à intervalles réguliers
- En notant alea() une fonction qui génère à chaque appel un nombre au hasard dans [0,1[, on forme l'échantillon comme suit :

```
k ← N div n // div désigne la division entière
dep ← E[alea() * k] + 1
e[1] ← dep
Pour i ← 2 à n
    e[i] = e[i-1] + k
Fin Pour
```

Echantillonnage stratifié (par strate)

- La population est subdivisée en strates.
- Dans chaque strate, on opère un échantillonnage aléatoire simple.
- Il faut connaître au préalable la répartition de la population par strate.
- Le choix des strates est d'autant plus efficace que les caractères étudiés ont une faible variation à l'intérieur de chaque strate.
- Pour estimer les paramètres, les résultats doivent être pondérés par l'importance relative de chaque strate dans la population.

Les caractères étudiés

- On appelle caractère simple une application C :

  C : P $\rightarrow$ R (ou ensemble de codes)

  $\omega \rightarrow X(\omega)$

où P indique la population, ω un individu et C(ω) la modalité pris par l'individu sur le caractère C.

- Le caractère indique une mesure (taille, poids, note) ou un attribut (sexe, CSP) observable sur les individus.
- Quand le caractère est une mesure, on parle d'un caractère quantitatif sinon on parle de caractère qualitatif. C(P) est l'ensemble des modalités du caractère.
  - C(P) $\subset$ R si le caractère est quantitatif
  - C(P) est un ensemble de codes (souvent des entiers)

Les caractères qualitatifs

- Les caractères nominaux : expriment l'appartenance à des catégories et leurs modalités ne sont pas hiérarchisées (Vote, Couleur de cheveu, diplôme, CSP, …).
- Les caractères dichotomiques : caractères nominaux ne pouvant prendre que deux modalités (Présence d'un symptôme, …).
- Les caractères ordinaux : L'ensemble des modalités est hiérarchisé (échelle de notation A-B-C-D-E-F des ECTS, degré de satisfaction pas du tout, peu, moyen, beaucoup, très bien).

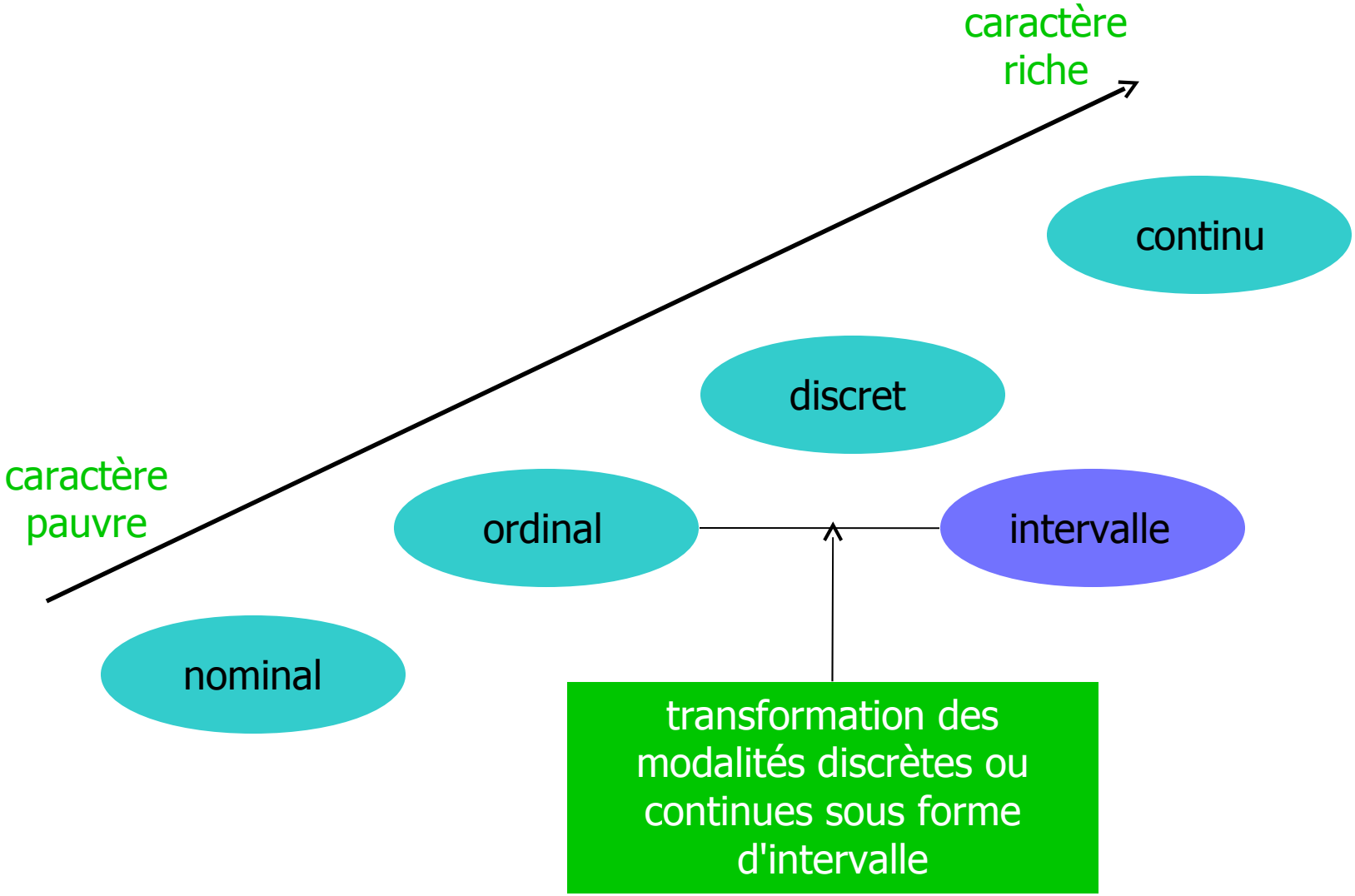
Les caractères quantitatifs

- Les caractères discrets : le caractère ne peut prendre que certaines valeurs dans un intervalle. En général, la mesure est un comptage ou dénombrement (nombre d'enfants, nombre d'accidents dans une période, ...).

- Les caractères continus : le caractère peut prendre toutes les valeurs d'un intervalle donné (le poids, la taille, le salaire, la température).
  Le nombre de valeurs possibles dépend de la précision de la mesure.

- En fait, un caractère continu est un caractère discret qui prend un grand nombre de valeurs.

Classification des types de caractères



Les traitements descriptifs

- La statistique univariée : On ne s'intéresse qu'à un seul caractère.
  Ex : Les salaires en Europe, nombre d'enfants par ménage, Temps de visite sur un site, Etude de l'âge des habitants d'un pays par classe d'âge.
- La statistique bivariée : On s'intéresse à l'étude simultanée de deux caractères pour mesurer leur dépendance.
  Ex : le vote est-il différent d'une CSP à l'autre ?
- La statistique multivariée : On s'intéresse à l'étude simultanée de p caractères. Même problématique que pour le bidimensionnel mais avec p assez grand.
  Ex : Etude du bien-être par département à travers des critère géo-socio-économiques (d° ensoleillement, nbre de théâtres, taux de suicides, infrastructure routière, ...)

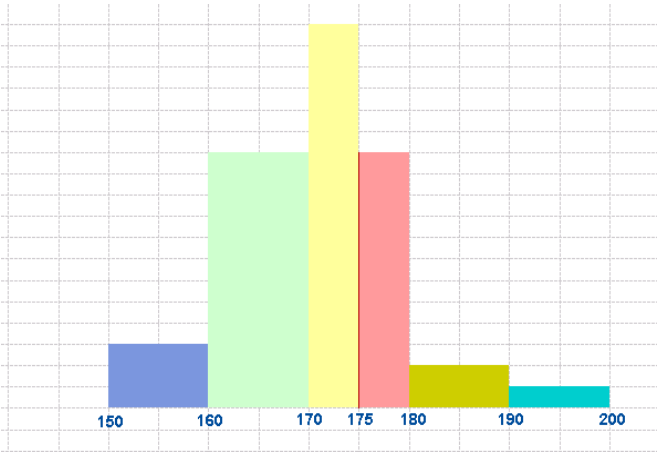
Statistique univariée

On ne s'intéresse qu'à un seul caractère de la population.

Résumés chiffrés des modalités

Tendances centrales : moyenne, médiane, quantile, …
Dispersions : écart-médian, écart-type (variance), moments d'ordre p, coefficient de Gini, ….

Résumés graphiques de répartitions des modalités



Histogramme

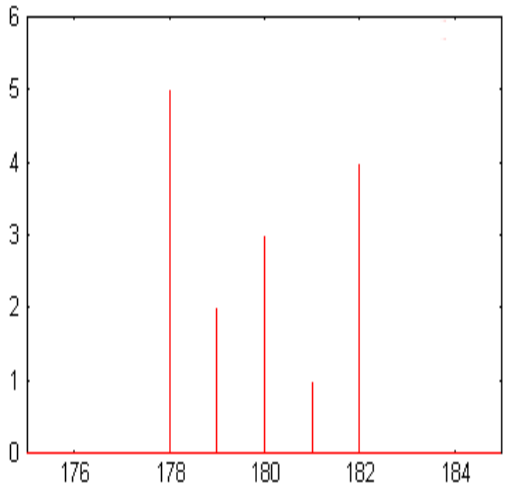
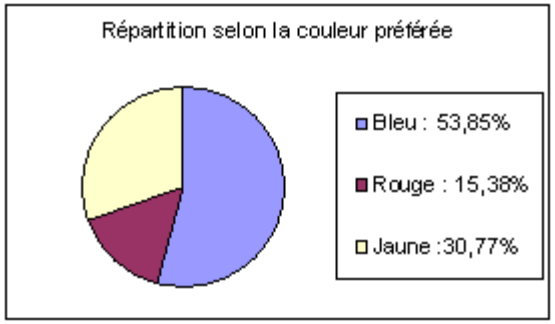


Diagramme en bâton



Camembert

Statistique bivariée

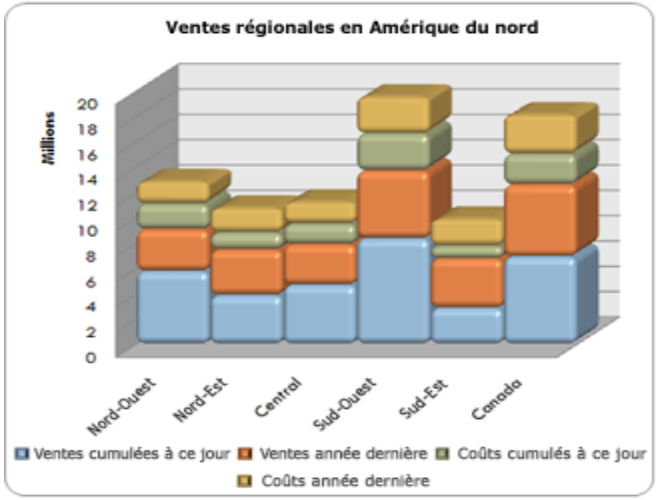
L'analyse bidimensionnelle : On s'intéresse à l'étude simultanée de deux caractères pour mesurer leur dépendance.

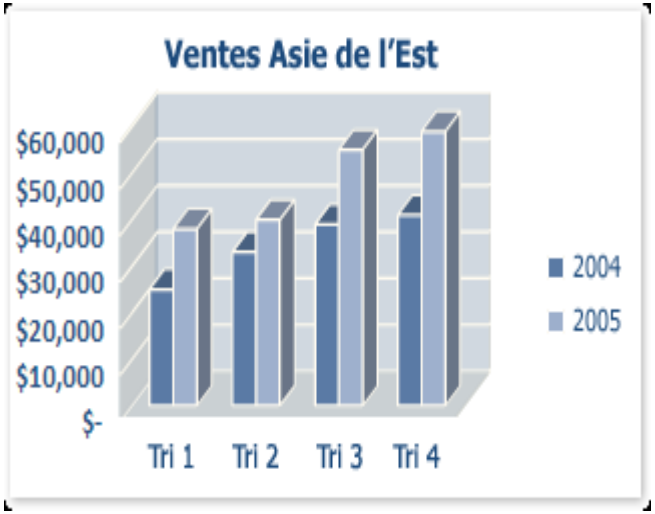
Résumés chiffrés des modalités

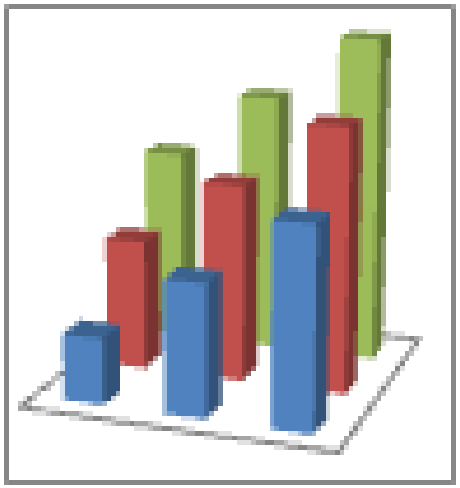
Résumés scalaires : coefficient de corrélation (Pearson), V de Kramer, coefficient de contingence, coefficient phi de Pearson

Résumés matriciels : Tableau de contingence, Répartitions conditionnelles, Répartitions marginales

Résumés graphiques de répartitions des modalités







Histogrammes 2D

Histogrammes 3D

Statistique multivariée

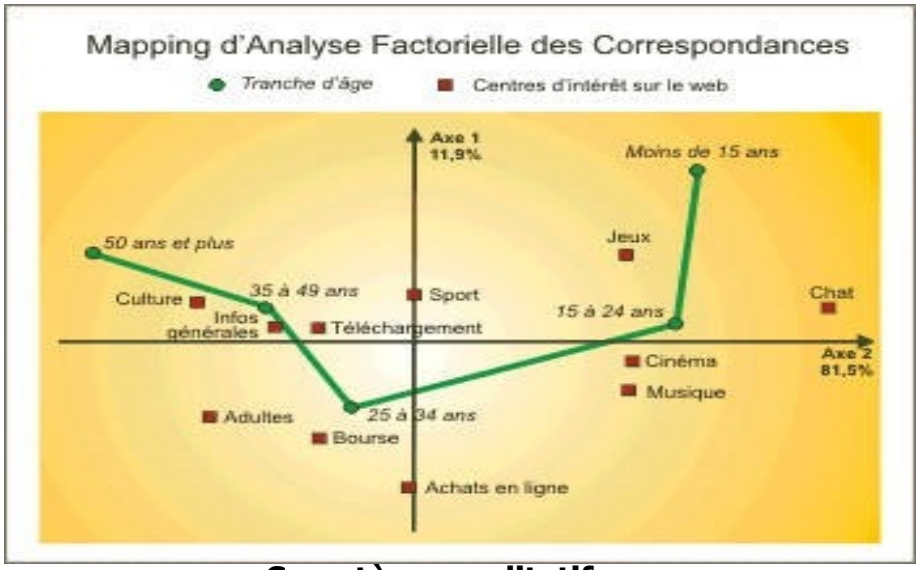
L'analyse multidimensionnelle : On s'intéresse à l'étude simultanée de p caractères. Même problématique que pour le bidimensionnel mais avec p assez grand.

Résumés chiffrés des modalités

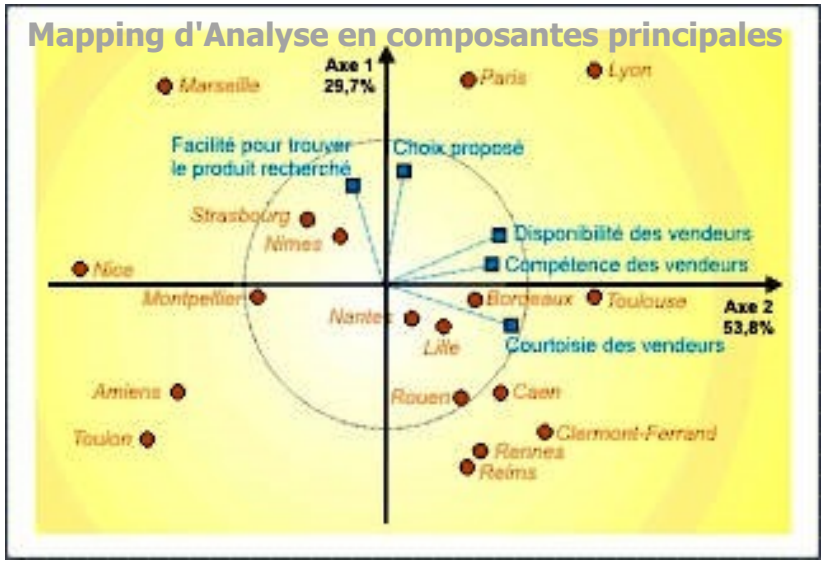
Résumés scalaires : moyenne et écart-type par caractère, fréquences conditionnelles et marginales

Résumés matriciels : Matrice de corrélations linéaires, Tableau de Contingence de Burt, Tableau disjonctif

Résumés graphiques : Mappings



Caractères qualitatifs



Caractères quantitatifs

Méthodologie d'étude statistique

1. Définir précisément le problème étudié :
   a) Quels sont les objectifs de l'étude ?
      - recensement des différentes questions posées
      - déduction des différentes études statistiques à opérer
      - définition des caractères étudiés avec leur type
   a) Quelle est la population étudiée ?
      - définition précise de l'unité statistique
      - définition du périmètre spatio-temporel ?
   a) Comment récupérer et stocker l'information
   b) Une enquête et/ou des données existantes
   c) Choix de la technique d'échantillonnage
   d) Récupération des données et validation des données récupérées
2. Exécution des études statistiques avec les logiciels appropriés
3. Rédaction du document de synthèse
   a) Rappel du contexte : objectifs de l'étude, périmètre de l'étude, définitions des études statistiques
   b) Insertion par étude des résumés chiffrés et graphiques
   c) Interprétation des résultats en cohérence avec le périmètre et les résumés