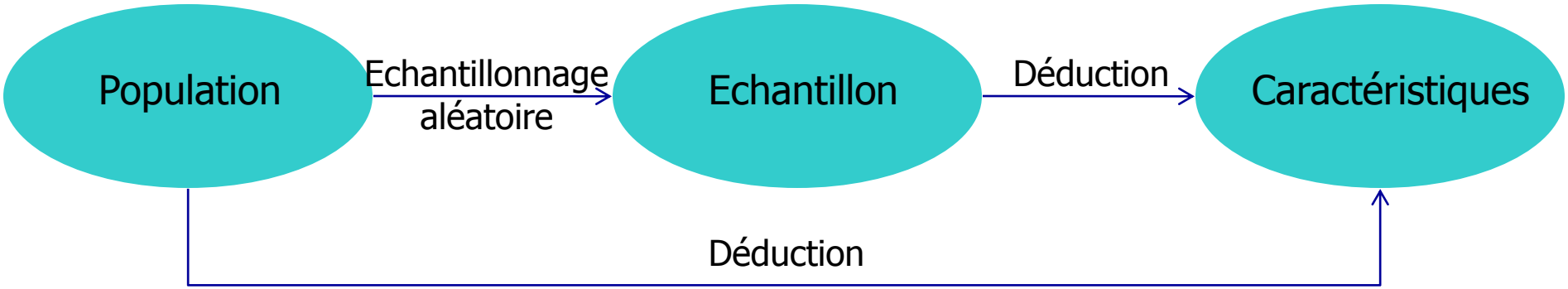
Objectif de la statistique

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

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, des diagrammes en bâton, des 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
  - Ce cours est vu en dans les parcours en Ing2.

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)

  $\omega \rightarrow C(\omega)$

  où P indique la population, ω un individu et $C(\omega)$ 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) sera appelé 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 : ils expriment l'appartenance à des catégories. Les modalités ne sont pas hiérarchisées (Vote, Couleur de cheveu, diplôme, CSP, …).
- Les caractères dichotomiques : caractères nominaux ne pouvant prendre que deux modalités (Sexe, 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).

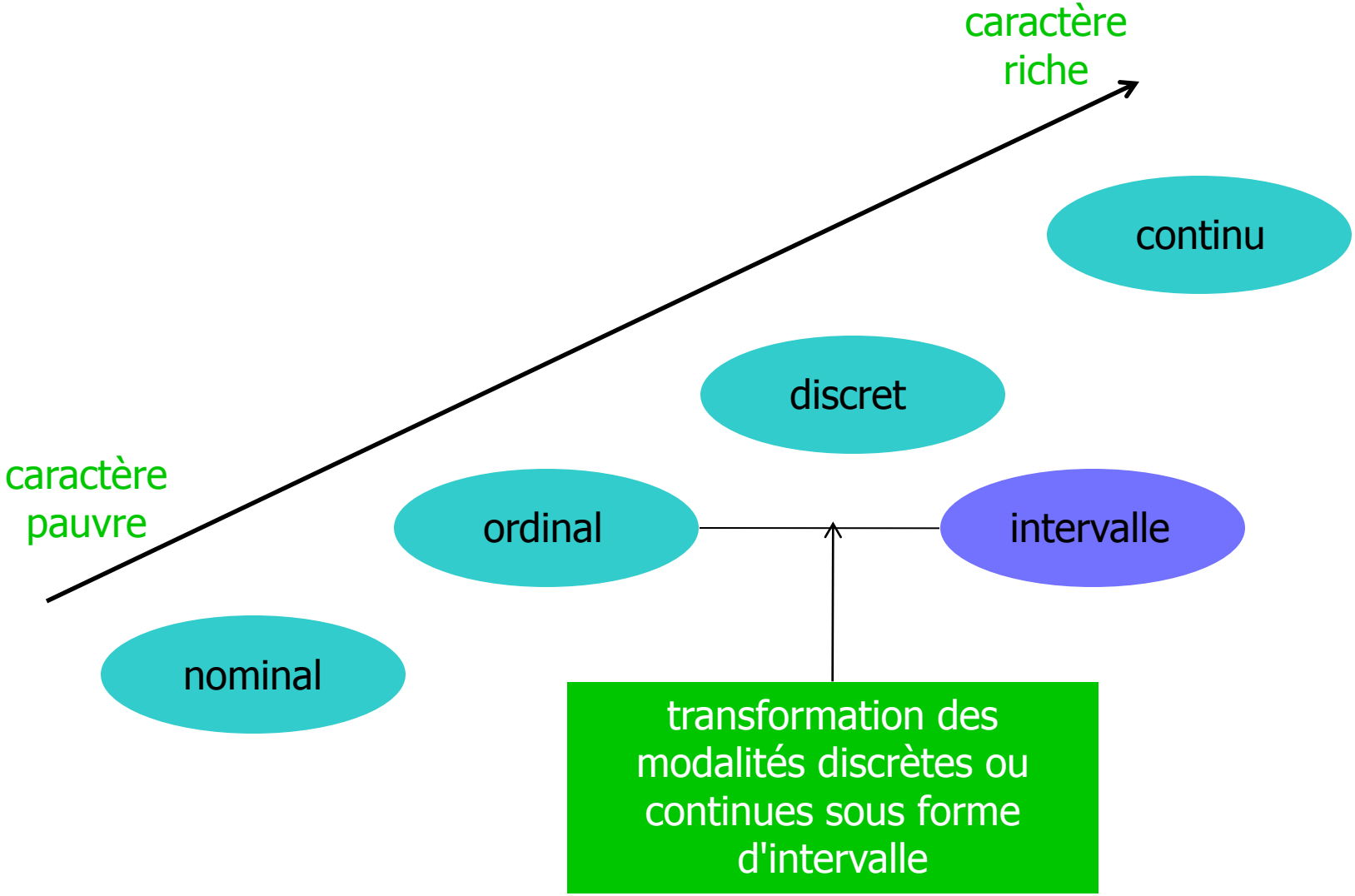
Les caractères quantitatifs

- Les caractères discrets : le caractère ne peut prendre que certaines valeurs d'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 (ensoleillement, nbre de théatres, taux de suicides, infrastructure routière, …)

A chaque type de caractère (qualitatif nominal, quantitatif continu,…) correspond un traitement spécifique.

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
   b) Quelle est la population étudiée ?
      - définition précise de l'unité statistique
      - définition du périmètre spatio-temporel ?
   c) Comment récupérer et stocker l'information
      - Enquête ou données existantes ou un mixte
      - Choix de la technique d'échantillonnage
      - 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

Statistique univariée

- Représentations graphiques
  - Camembert
  - Diagramme en bâton
  - Histogramme
- Résumés numériques
  - Caractéristiques de tendances centrales
  - Caractéristiques de dispersion

Effectifs et fréquences

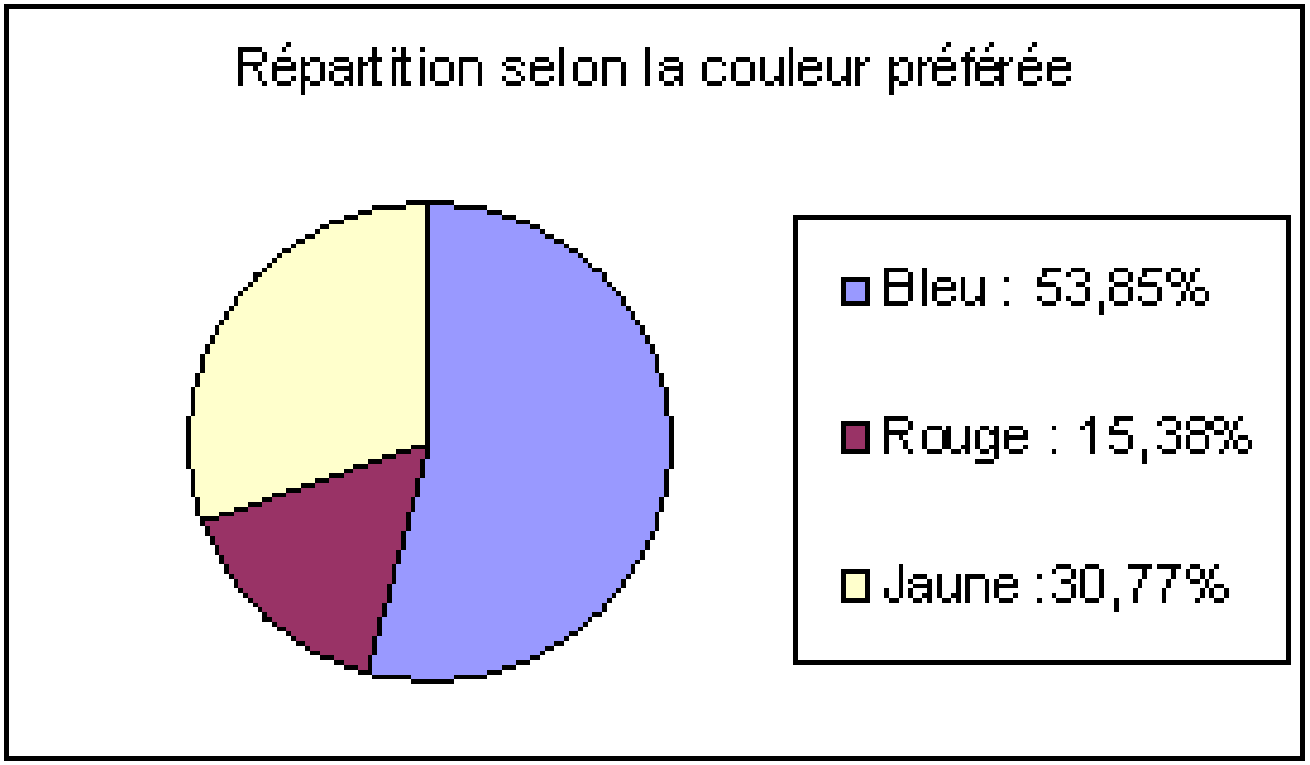
- Un calcul peut être effectué sur tous les caractères : Compter le nombre d'unités statistiques qui ont une modalité donnée.
- On utilisera le terme effectif de la modalité i. On notera cet effectif n_i.
- On aura besoin de ramener les effectifs en pourcentage. On parlera alors de fréquences. On notera cette fréquence f_i.
- $f_i = n_i / n$ où n est l'effectif total.
- Remarque : n_i est aussi appelé fréquence absolue et f_i fréquence relative.

Représentations graphiques

Camembert

- Il concerne de préférence les caractères qualitatif nominaux car pas d'ordre

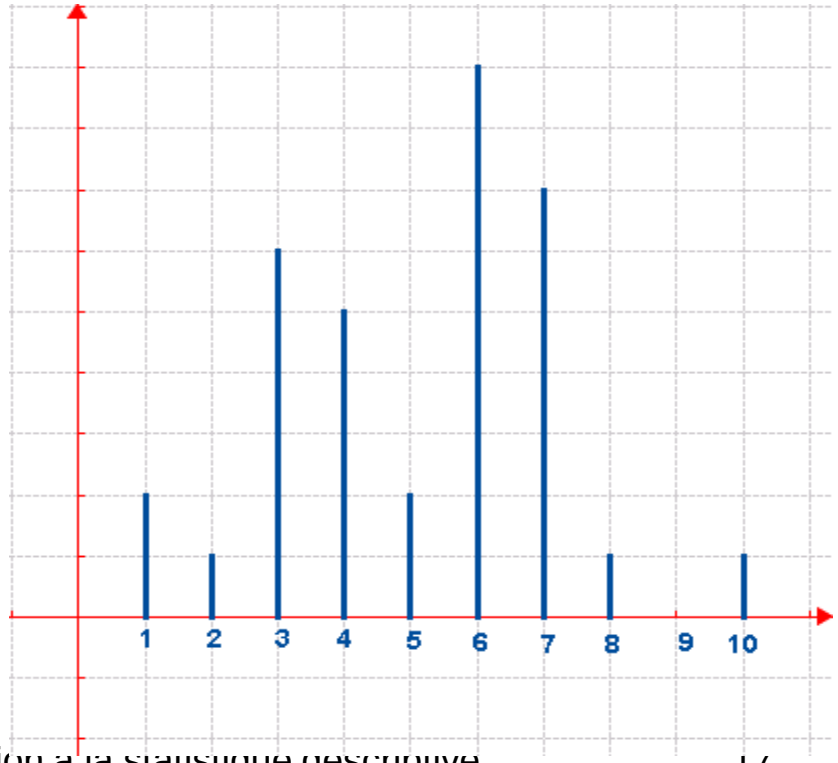
  – Chaque portion du disque est proportionnelle à l'effectif ou la fréquence



Représentations graphiques
Diagramme en bâton

- Il concerne tous les types de caractères sauf le caractère quantitatif continu.
- C'est une représentation planaire.
  - Sur l'axe des x on a les différentes modalités du caractère
  - sur l'axe des y on a la fréquences relatives ou absolues des modalités

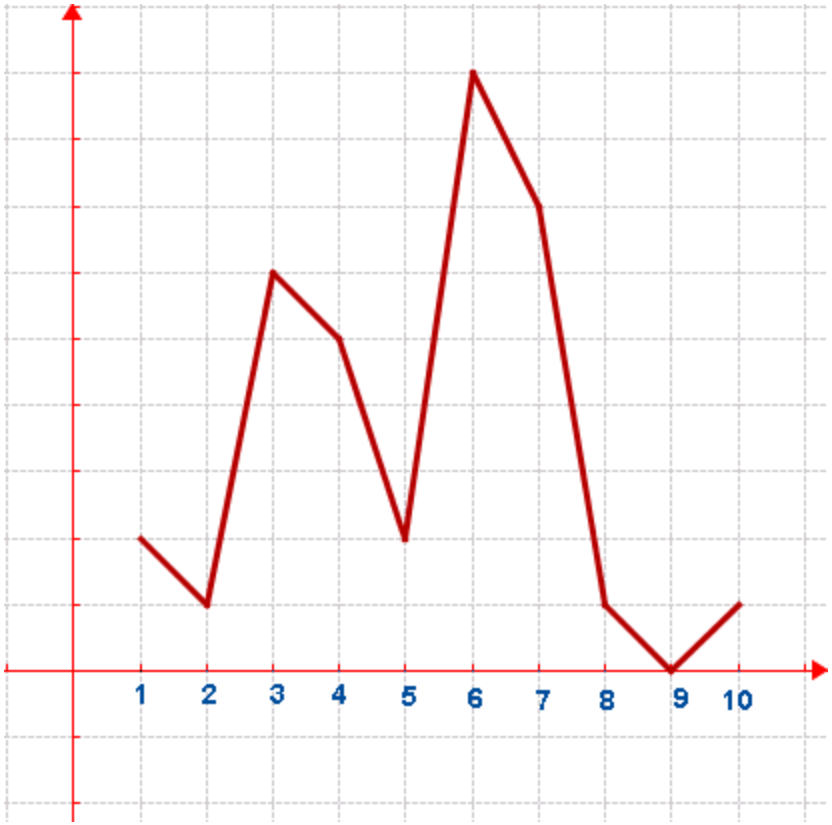
Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1



Représentations graphiques
Polygone de fréquences

- Il concerne les caractères de type qualitatif ordinal ou quantitatif discret.
- On représente dans le plan le polygone (mod_i, n_i)

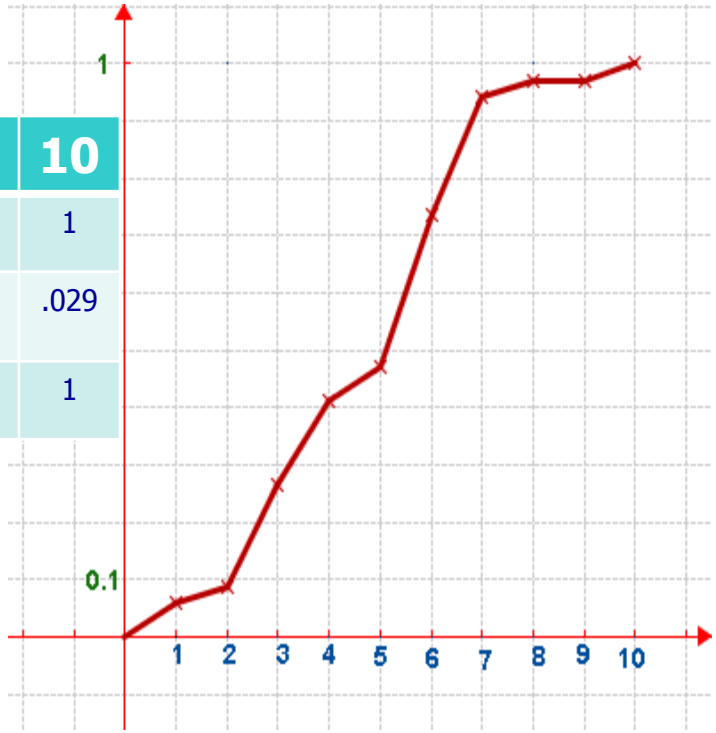
Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1



Représentations graphiques
Courbe de fréquences cumulées

- Il concerne les caractères de type qualitatif ordinal ou quantitatif discret.
- On représente dans le plan le polygone (mod_i,fc_i) où fc_i est la fréquence cumulée des modalités inférieures ou égales à mod_i.

Note	1	2	3	4	5	6	7	8	9	10
n_i	2	1	6	5	2	9	7	1	0	1
f_i	.059	.029	.176	.147	.059	. 264	.206	.029	.0	.029
fc_i	.059	.088	.264	0.411	.470	.734	.940	.969	.969	1



Représentations graphiques

Histogramme

- Il concerne les caractères de type quantitatif continu et accessoirement de type quantitatif discret.
- On partitionne l'ensemble des valeurs du caractères en une famille finie d'intervalles contigus. Les valeurs ponctuelles n'ont pas de sens dans ce graphique.

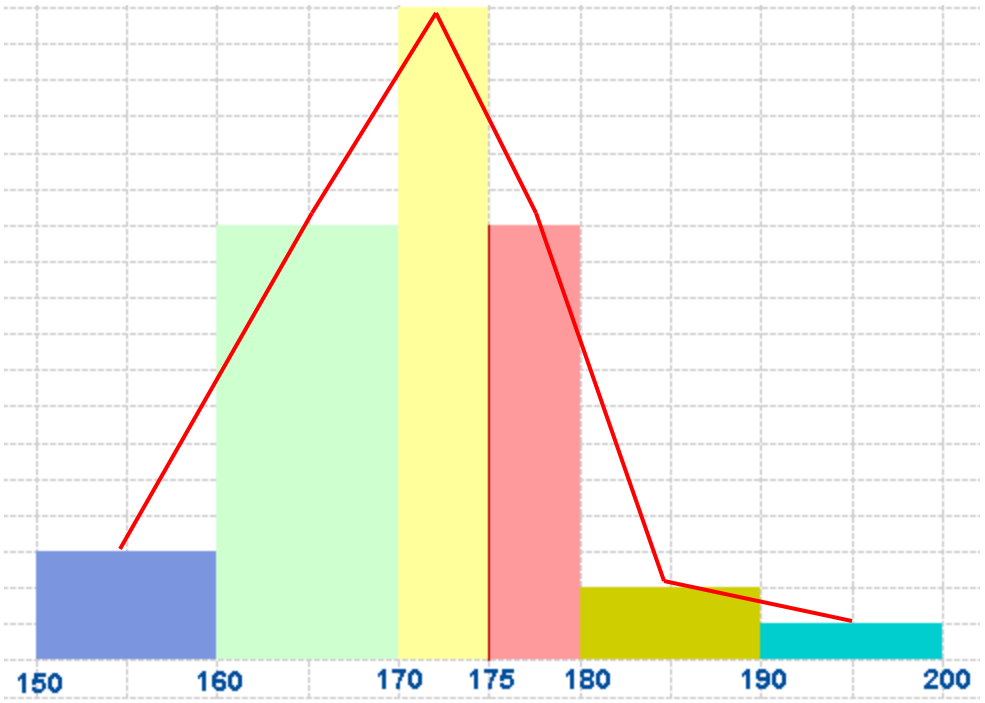
Dans un plan on place sur l'axe des x les différentes bornes des intervalles$[a_i, a_{i+1}[$. Pour chaque intervalle, on trace un rectangle dont la base est $[a_i, a_{i+1}[$ et **la surface est proportionnelle à n_i** où n_i est le nombre d'individus dont la modalité sur le caractère appartient à $[a_i, a_{i+1}[$.

L'histogramme dépend du nombre d'intervalles. On choisit un nombre voisin de

$$E[1 + 10 * \log(n) / 3]$$

Comme pour le diagramme en bâton, on peut tracer un polygone de fréquences. La base en x est le milieu des intervalles. Voir la ligne polygonale en rouge sur le graphique ci-contre.

Taille	[150, 160[	[160, 170[	[170, 175[	[175, 180]	[180, 190]	[190, 200]
Eff.	3	12	9	6	2	1



Résumé des représentation graphiques

Variables qualitatives

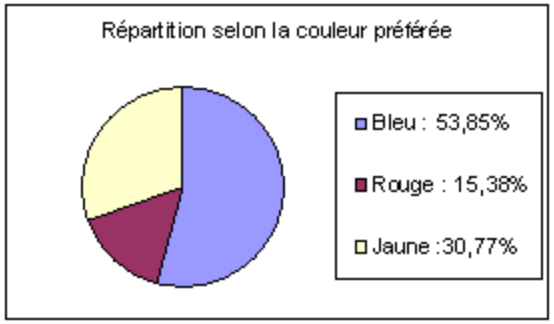
Variables quantitatives

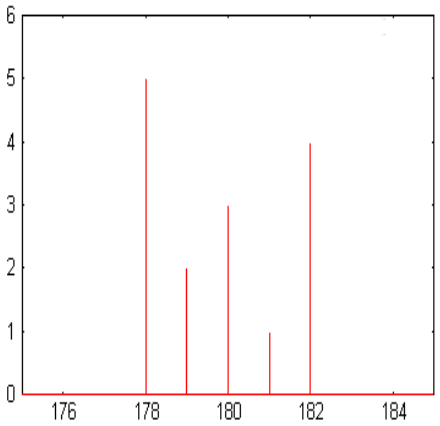
nominales

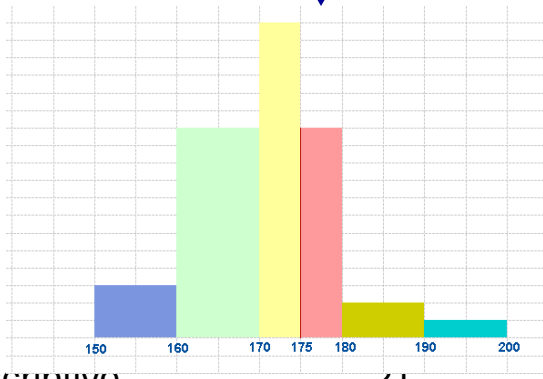
ordinales

discrètes

continues





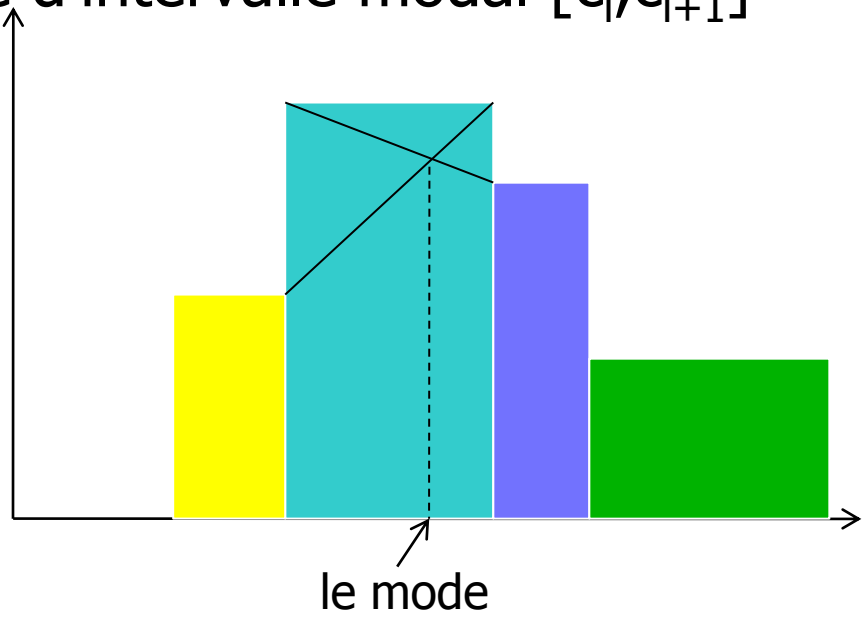


Résumés numériques

- La statistique descriptive synthétise les données pour les rendre intelligibles.
- Les tableaux et graphiques vus précédemment sont les premières synthèses.
- La suite consiste à calculer des synthèses chiffrées.
- Les plus importantes sont :
  - Caractéristiques de tendances centrales pour expliquer ce qui explique principalement le caractère.
  - Caractéristiques de dispersion pour expliquer comment varie le caractère autour des tendances centrales.

Tendance centrale
le mode

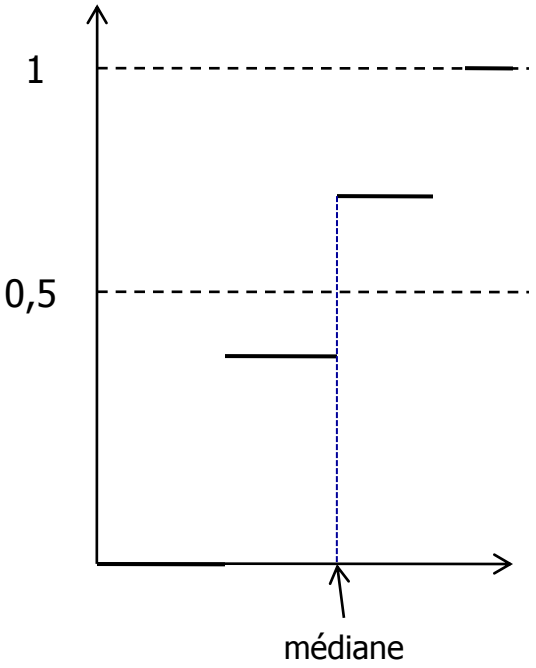
- C'est la modalité observé d'effectif maximum. Elle concerne tous les types de caractères.
- Il sert essentiellement à détecter si la population est homogène ou éventuellement constituée de deux ou plusieurs populations hétérogènes
- Dans le cas du type quantitatif discret, si on a deux effectifs maximum côte à côte on parle d'intervalle modal $[c_i,c_{i+1}]$
- Dans le cas du type quantitatif continu il faut tenir compte des classes adjacentes. Voir graphique ci-contre

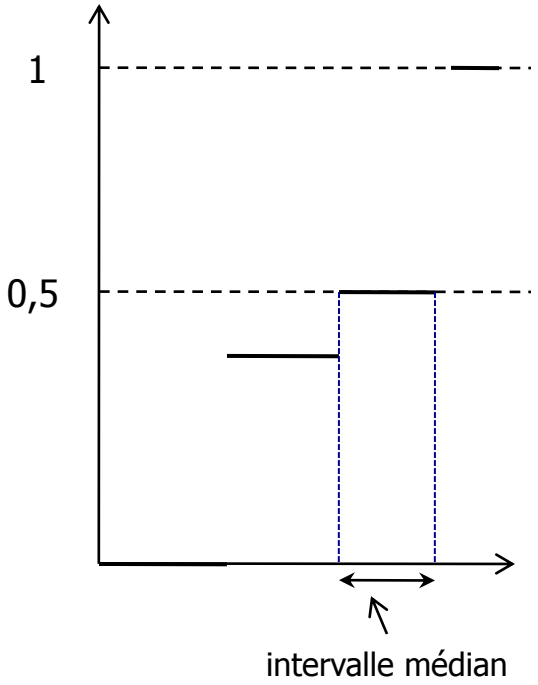


Tendance centrale

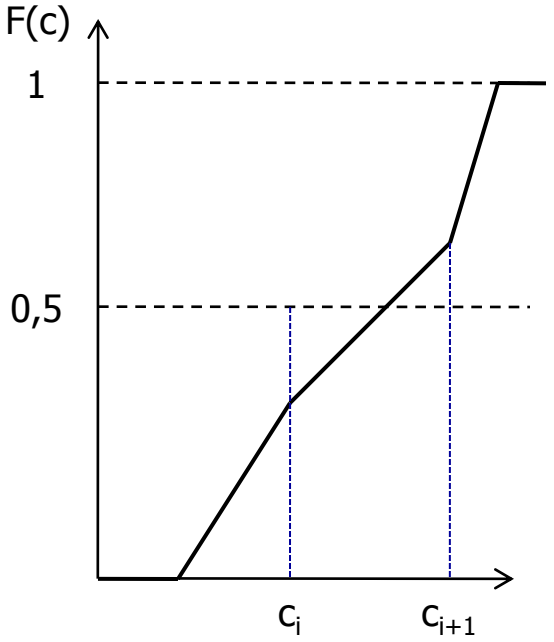
Médiane

- La médiane est la modalité qui sépare la population en deux groupes d'effectifs égaux. Elle n'a de sens qu'avec le type qualitatif ordinal et les types quantitatifs.





Type quantitatif discret



$mé = c_i + (0,5 - F(c_i)/(F(c_{i+1})-F(c_i))$

Les valeurs c_i sont les milieux des intervalles

Type quantitatif continu

Tendance centrale
Moyenne

- La moyenne est une valeur calculée qui est la moyenne des modalités c_i pondérée par les effectifs n_i. Elle n'a de sens qu'avec les types quantitatifs.
- Contrairement à la médiane, ce n'est pas nécessairement une modalité du caractère.
- C'est la tendance la plus utilisée pour des raisons algébriques plus que statistique. Elle est facile à calculer et c'est une fonction dérivable (voir analyse multivariée)
- Contrairement à la médiane, elle est très sensible à l'intensité des valeurs du caractère.
- Notation : Si c est le caractère alors la moyenne est notée $\overline{c}$
- En résumé la moyenne et la médiane sont complémentaires.

Moyenne et Médiane : perte d'information

- Quand on passe de la série statistique (c_i) à la moyenne que perd-on comme information ?

Considérons le problème suivant :

$$\min_{x \in R} err_2(x) = \frac{1}{n} \sum_i n_i (c_i - x)^2$$

$$err'(x) = 0 \Leftrightarrow -\frac{2}{n} \sum n_i (c_i - x) = 0 \Leftrightarrow \sum n_i \cdot c_i = \sum n_i . x = n.x \Leftrightarrow x = \frac{1}{n} \sum n_i \cdot c_i = \bar{c}$$

Choisir la moyenne comme résumé consiste à minimiser l'erreur quadratique.

- Quand on passe de la série statistique (c_i) à la médiane que perd-on comme information ?

De même on peut définir

$$\min_{x \in R} err_1(x) = \frac{1}{n} \sum_i n_i |c_i - x|$$

On peut démontrer que dans ce cas la valeur optimale est la médiane.

Choisir la médiane comme résumé consiste à minimiser l'erreur en valeur absolue. On parle dans ce cas d'écart-médian.

Dispersion
écart-type, écart-médian

- Quand on utilise la moyenne comme tendance centrale, il faut utiliser la variance et l'écart-type du caractère comme indicateur de dispersion

  Variance : $$s_C^2 = \frac{1}{n}\sum_i n_i (c_i - \bar{c})^2$$

  Ecart-type : $$s_c = \sqrt{s_C^2}$$

- Quand on utilise la médiane comme tendance centrale, il faut utiliser l'écart-médian du caractère comme indicateur de dispersion

$$em(C) = \frac{1}{n}\sum_i n_i |c_i - me|$$

Dispersion
Etendue et quantiles

- Dans tous les cas de figures, on peut utiliser l'étendue du caractère : $\mathbf{et(C) = Max\ c_i - Min\ c_i}$

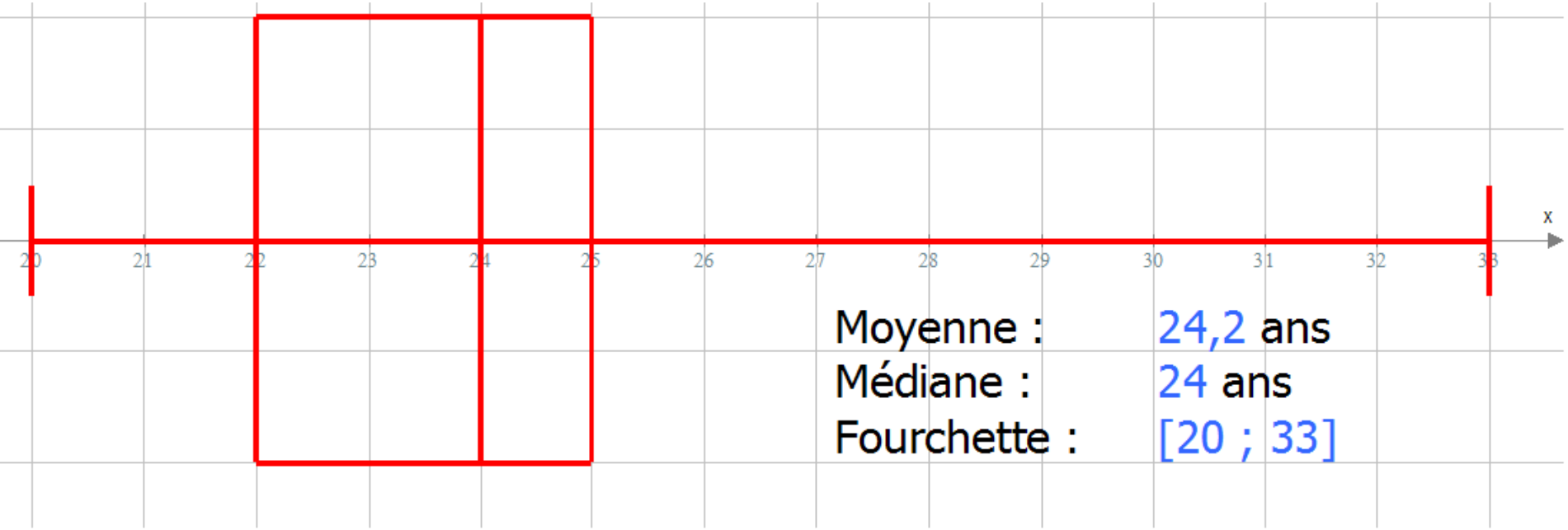
- Les quantiles : On ordonne les valeurs du caractères et on divise cette série ordonnée en paquets égaux. Chaque modalité extrême de chaque paquet est un quantile.

- Quand on divise en quatre paquets, on obtient les quartiles. Le 2ème quartile est la médiane. Le 1er quartile découpe la population en deux sous populations : la première sous population est formée du ¼ des individus, la deuxième est formée des ¾ des individus.

- On utilise souvent le 1er et le 99ème centiles pour éliminer des individus dits aberrants.

Dispersion
Boite à moustache

- C'est un résumé inventé par John Tukey
- Il consiste à afficher sur un graphique cinq indicateurs révélateurs d'un profil : la plus petite valeur, le 1er quartile, la médiane, le 3ème quartile et la plus grande valeur.



Résumé des indicateurs numériques

Série observée

Note	1	2	3	4	5	6	7
Elève L	9	10	8	7	10	9	11
Elève P	14	2	16	5	6	5	16

Série ordonnée

Elève L	7	8	9	9	10	10	11
Elève P	2	5	5	6	14	16	16

Tendance centrale

	Moyenne	Médiane
Elève L	9,1	9
Elève P	9,1	6

Dispersion

	Variance	Ecart-type	Q1	Q3
Elève L	1,81	1,34	8	10
Elève P	35,48	5,96	5	14