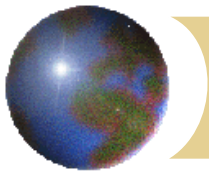


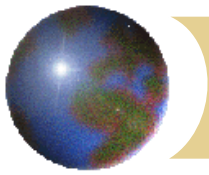
CHAPITRE 2

Statistiques à une dimension



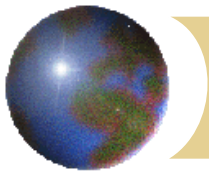
Introduction

- ✚ L'objectif des outils de Statistique descriptive élémentaire est de fournir des résumés synthétique de séries de valeurs, adaptés à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon.
- ✚ **Cas d'une seule variable**, notions classiques de médiane, quantile, moyenne, fréquence, variance, écart-type définies à des représentations graphiques : diagramme en bâton, histogramme, diagramme-boîte, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.
- ✚ **Cas de deux variables**, on s'intéresse à la corrélation, au rapport de corrélation ou encore à la statistique d'un test du 2 associé à une table de contingence. Ces notions sont associées à différents graphiques comme le nuage de points (scatterplot), les diagrammes-boîtes parallèles, les diagrammes de profils ou encore en mosaïque.



Démarche

- ✚ Etude exploratoire à l'aide d'outils en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec des données et de dépister les sources de problèmes : valeurs manquantes, erronées ou atypiques, modalités trop rares, distributions anormales, incohérences, liaisons non linéaires.
- ✚ Prétraitements des données afin de corriger les sources de problèmes et les rendre exploitables par des techniques plus sophistiquées: transformation des variables, codage en classe, données manquantes,...
- ✚ Ensuite, les techniques exploratoires multidimensionnelles permettent des représentations graphiques synthétiques, réductions de dimension pour la compression ou le résumé des données, recherches et représentations de typologies des observations.



Variable quantitative discrète

- ✚ On appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales).

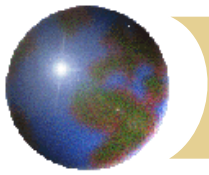
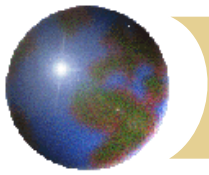


Tableau statistique

- ✚ Tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X ; ces observations sont rangées par ordre croissant et non répétées ; nous les noterons $\{x_l; l = 1; \dots; r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_l , le nombre de réplifications qui lui sont associées ; ces réplifications sont appelées effectifs et notées n_l . Les effectifs n_l sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$, appelées *fréquences* (rappelons que n désigne le nombre total d'observations, c'est-à-dire le cardinal de Ω : $n = \sum_{l=1}^r n_l$)

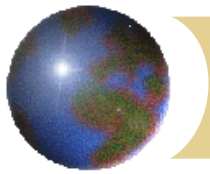


Effectifs cumulés et les fréquences cumulées

- ✚ Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_l = \sum_{j=1}^l n_j \text{ et } F_l = \sum_{j=1}^l f_j$$

On notera que $N_r = n$ et $F_r = 1$



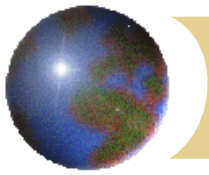
Exemple

- Soit la série statistique brute qui présente l'âge (arrondi à l'année près) des 48 salariés d'une entreprise:

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
 62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
 57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

- Dans le tableau statistique (ci-contre), on a calculé les effectifs, effectifs cumulés, fréquences et fréquences cumulées.

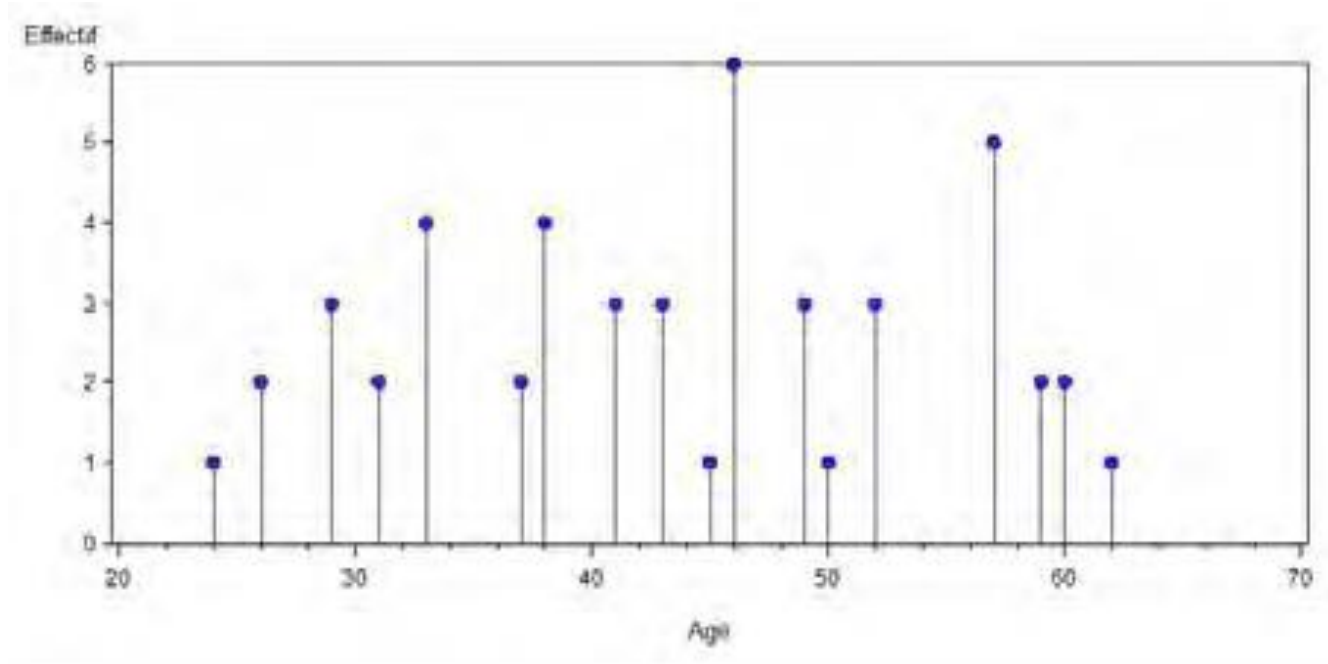
x_l	n_l	N_l	$f_l(\%)$	$F_l(\%)$
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,00

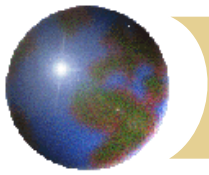


Représentations graphiques

Diagramme en bâtons

- ✚ Il permet de donner une vision d'ensemble des observations réalisées.



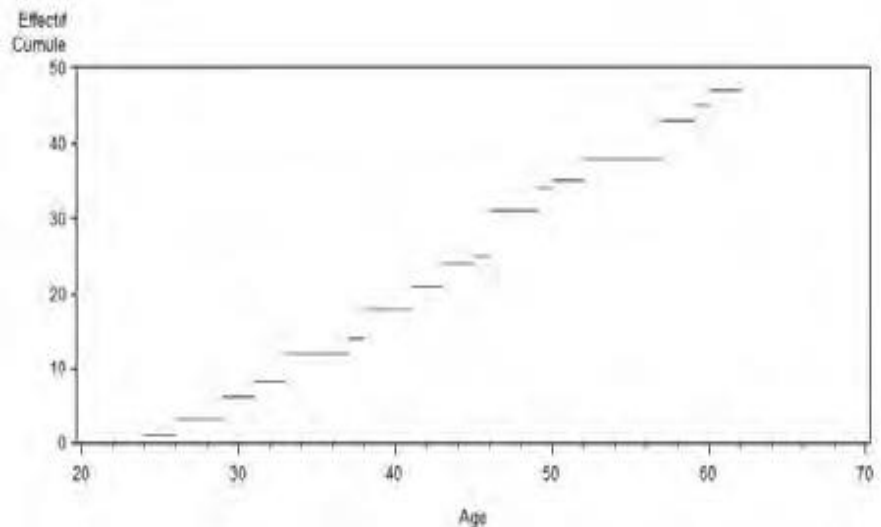


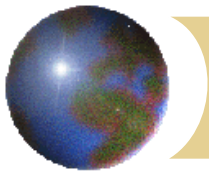
Représentations graphiques

Diagramme cumulatif

- ✚ Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion) d'observations inférieures ou égales à une valeur donnée de la série.
- ✚ Lorsqu'il est relatif aux fréquences, c'est en fait le **graphe de la fonction de répartition empirique** F_X définie de la façon suivante :

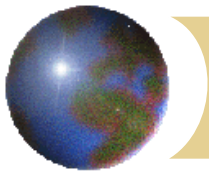
$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1; \\ F_l & \text{si } x \leq x_{l+1}; l = 1; \dots; r - 1 \\ 1 & \text{si } x \geq x_r \end{cases}$$





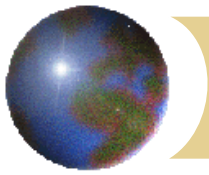
Notion de quantile

- ✚ **Définition:** La fréquence cumulée $F_i (0 \leq F_i \leq 1)$ donne la proportion d'observations inférieures ou égales à x_i .
- ✚ Une approche complémentaire consiste à se donner a priori une valeur α , comprise entre 0 et 1, et à rechercher x vérifiant $F_X(x_\alpha) \cong \alpha$. La valeur x_α (qui n'est pas nécessairement unique) est appelée **quantile** (ou fractile) d'ordre α de la série.
- ✚ Les **quantiles** sont les valeurs qui partagent la population en n parts égales.



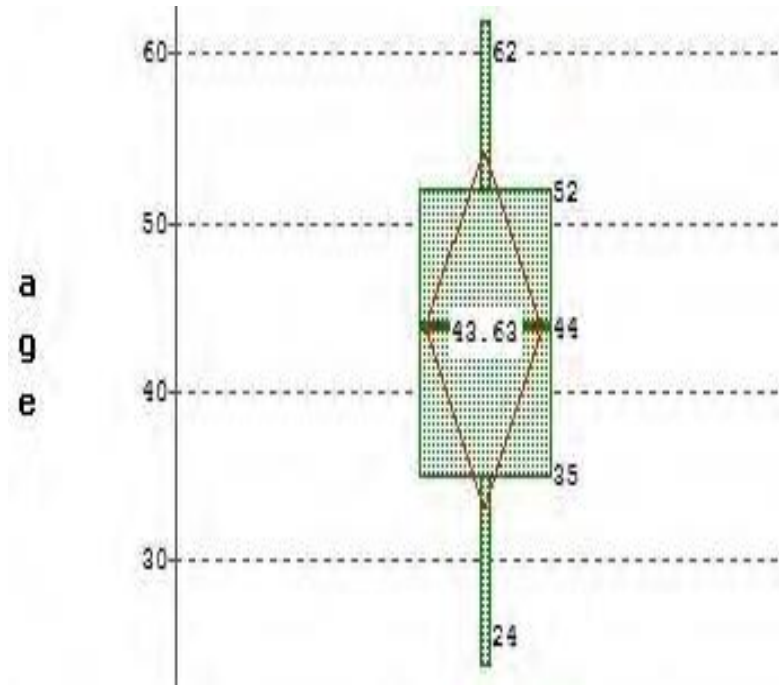
Les quantiles les plus utilisés

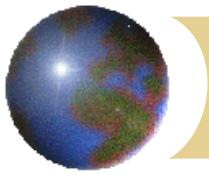
- ✚ Les quantiles les plus utilisés sont associés à certaines valeurs particulières de α .
 - ✚ Si la population est séparée en 2 ($\alpha = \frac{1}{2}$), c'est une **médiane**.
 - La médiane est le quantile d'ordre $\frac{1}{2}$; elle partage donc la série des observations en deux ensembles d'effectifs égaux
 - ✚ Si la population est séparée en 4 ($\alpha = \frac{1}{4}$), ce sont des **quartiles**.
 - Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane).
 - ✚ Si la population est séparée en 5 ($\alpha = \frac{1}{5}$), ce sont des **quintiles**.
 - ✚ Si la population est séparée en 10 ($\alpha = \frac{1}{10}$), ce sont des **déciles**.
 - ✚ Si la population est séparée en 100 ($\alpha = \frac{1}{100}$), ce sont des **centiles**.



Le diagramme-boîte (box-and-whisker plot)

- ✚ Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane.
- ✚ La figure ci-contre donne le diagramme-boîte de l'exemple précédent.
- ✚ Dans cet exemple, on a obtenu $x_{1/4} = 35$, $x_{1/2} = 44$ et $x_{3/4} = 52$; on notera que l'obtention, d'une part de $x_{1/4}$ et $x_{1/2}$, d'autre part de $x_{3/4}$, ne s'est pas faite de la même façon (en fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple).

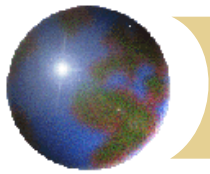




Caractéristiques numériques

Tendance centrale

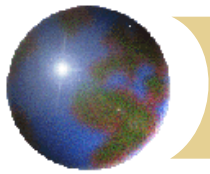
- ✚ Les deux caractéristiques les plus usuelles pour situer le centre, le milieu d'une série étudiée sont :
 - ▣ la médiane,
 - ▣ la moyenne (ou moyenne arithmétique).
- ✚ Formule de la moyenne pour une variable quantitative discrète : $\bar{x} = \frac{1}{n} \sum_{l=1}^r n_l x_l = \sum_{l=1}^r f_l x_l$



Caractéristiques numériques

Dispersion

- ✚ Elle sert à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.
- ✚ Quelques caractéristiques de dispersion:
 - L'étendue ($x_r - x_1$),
 - l'intervalle inter-quartiles ($x_{\frac{3}{4}} - x_{\frac{1}{4}}$),
 - l'écart-moyen à la médiane ($\frac{1}{n} \sum_{l=1}^r n_l \left| x_l - x_{\frac{1}{2}} \right|$),
 - l'écart-moyen à la moyenne ($\frac{1}{n} \sum_{l=1}^r n_l \left| x_l - \bar{x} \right|$),



Ecart-type

- ✪ Mais, la caractéristique de loin la plus utilisée est **l'écart-type**, racine carrée positive de la variance.

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n} \sum_{l=1}^r n_l (x_l - \bar{x})^2$$

L'écart-type de X sera donc noté σ_X .

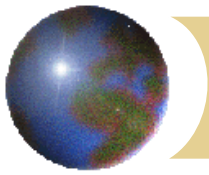
$$= \frac{1}{n} \sum_{l=1}^r n_l (x_l)^2 - (\bar{x})^2.$$

- ✪ Illustration: En utilisant toujours l'exemple précédent, on a calculé :

$$\bar{x} = \frac{1}{n} \sum_{l=1}^r n_l x_l = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans};$$

$$\sigma_X^2 = \frac{1}{n} \sum_{l=1}^r n_l (x_l)^2 - (\bar{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760;$$

$$\sigma_X = \sqrt{\sigma_X^2} \simeq 10,5 \text{ ans}.$$

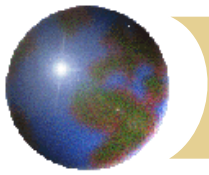


Variable quantitative continue

Généralités

- ✿ Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Le sous-ensemble de $\mathbb{R}$ des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés classes.
- ✿ Deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative:
 - ✦ le grand nombre d'observations distinctes (un traitement en discret serait dans ce cas peu commode)
 - ✦ le caractère "sensible" d'une variable (il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis).

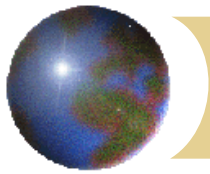
Nous noterons $(b_0 ; b_1), \dots, (b_{r-1} ; b_r)$ les classes considérées. Les nombres b_{l-1} et b_l sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1} + b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée a_l).



Présentation des données

- ✚ On utilise encore un tableau statistique analogue à celui vu pour les variables discrètes, en disposant dans la première colonne les classes rangées par ordre croissant.
- ✚ Les notions d'effectifs, de fréquences, d'effectifs cumulés et de fréquences cumulées sont définies de la même façon que dans le cas discret.
- ✚ **Exemple:** répartition des exploitations agricoles selon la SAU (Surface Agricole Utilisée) exprimée en hectares.

SAU (en ha)	fréquences (%)
moins de 5	24,0
de 5 à 10	10,9
de 10 à 20	17,8
de 20 à 35	20,3
de 35 à 50	10,2
plus de 50	16,8



Représentations graphiques

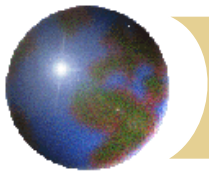
Courbe cumulative

- ✚ C'est le graphe de la fonction de répartition empirique, exprimée au moyen d'interpolations linéaires.

On appelle fonction de répartition empirique de la variable continue X la fonction F_X définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \dots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

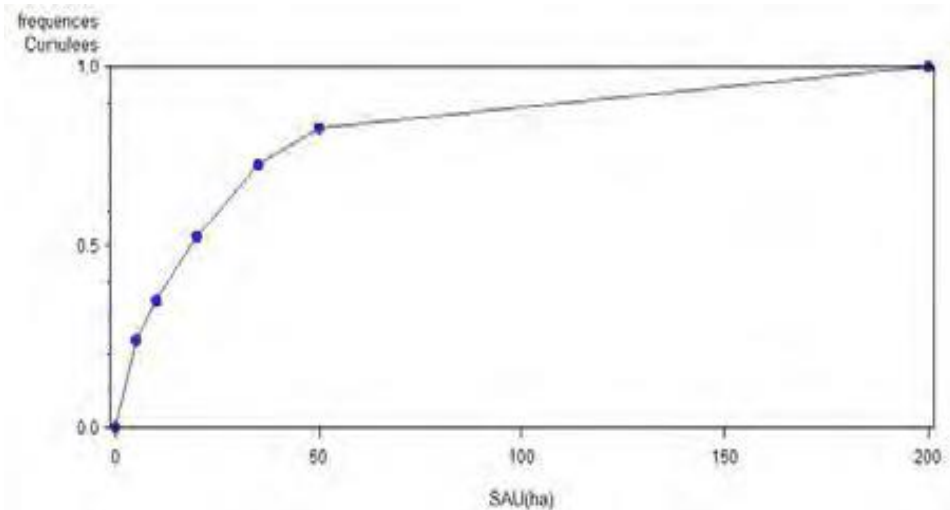
(on a supposé $F_0 = 0$).

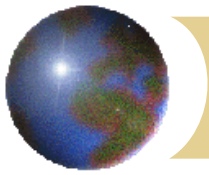


Représentations graphiques

Courbe cumulative de l'exemple

- ✚ La courbe cumulative relative à l'exemple précédent est donnée par la figure ci-contre.
- ✚ On notera que dans cet exemple, comme c'est souvent le cas avec une variable quantitative continue, on a fixé arbitrairement la borne inférieure de la première classe (on a choisi $b_0 = 0$) ainsi que la borne supérieure de la dernière classe (on a choisi $b_6 = 200$, mais d'autres choix étaient possibles).





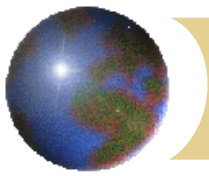
Représentations graphiques

Histogramme

- ⊕ La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses $b_0, b_1, \dots, b_r$. Sa fonction dérivée, éventuellement non définie en ces points, est appelée densité empirique de X et notée f_X . On obtient :

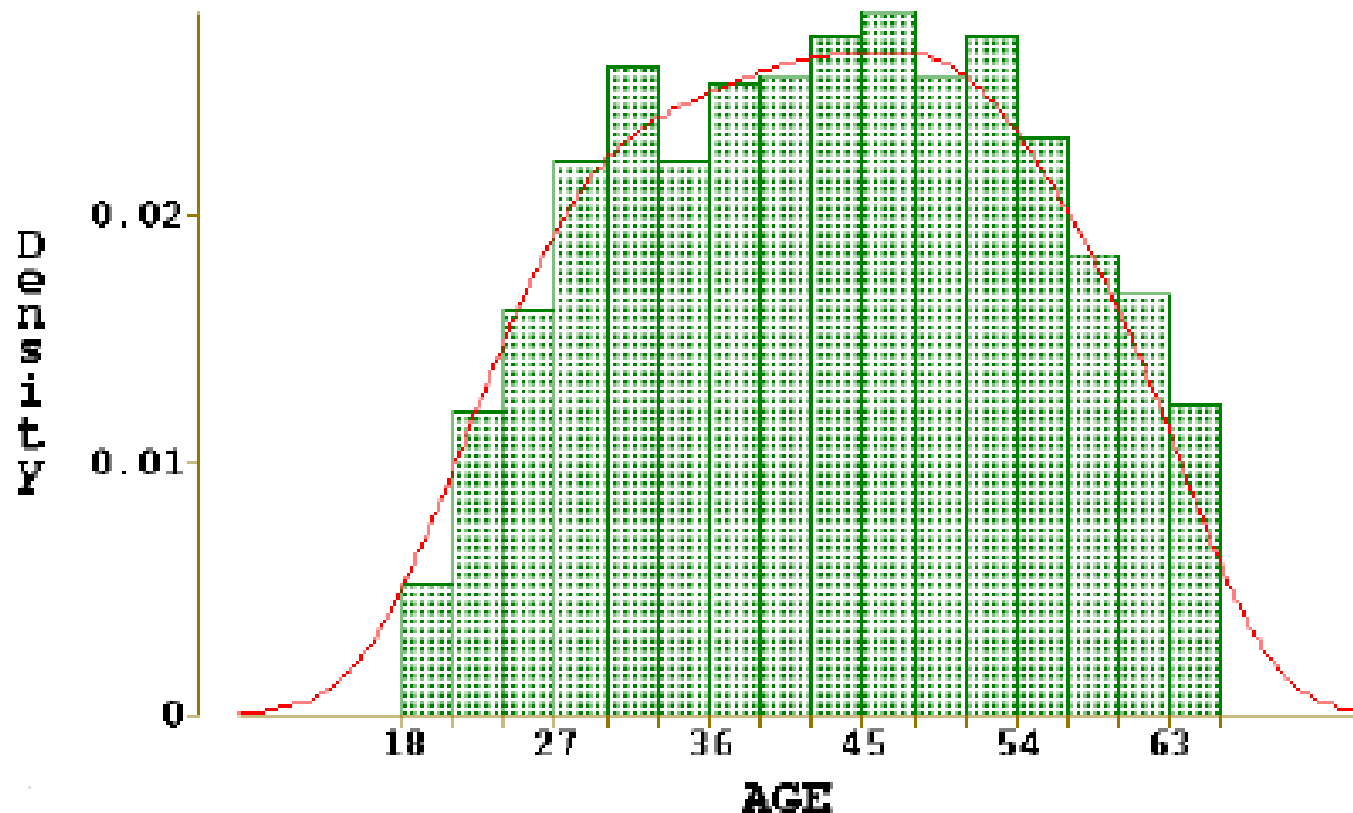
$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \dots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

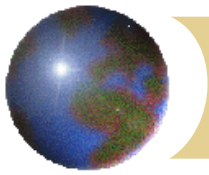
Le graphe de f_X est alors appelé histogramme de la variable X . Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$) et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées *densités de fréquence*. L'aire du $l^{\text{ième}}$ rectangle vaut donc f_l , fréquence de la classe correspondante.



Représentations graphiques

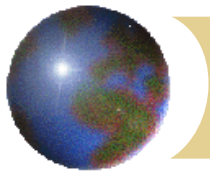
Histogramme de l'exemple





Quantiles, moyenne et écart-type des variables quantitatives continues

- ✚ Les quantiles x_j d'une variable continue peuvent être déterminés de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire (voir exemple plus loin).
- ✚ La moyenne, la variance et l'écart-type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour x_j les centres de classes au lieu des observations (qui ne sont pas connues).
- ✚ Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.



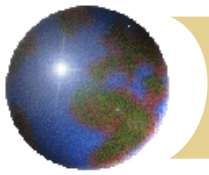
Médiane et Moyenne de l'exemple

- La médiane de la variable présentée dans l'exemple 2.2 se situe dans la classe (10 ; 20), puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante:

$$\begin{aligned}x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \quad \text{La moyenne vaut :} \\ &= 10 + 10 \frac{15,1}{17,8} \\ &\simeq 18,5 \text{ ha.}\end{aligned}$$

$$\bar{x} = \sum_{l=1}^n f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

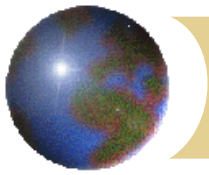
- Dans cet exemple, nous avons choisi $x_6 = 100$ pour calculer la moyenne.



Variable qualitative

Variables nominales et ordinales

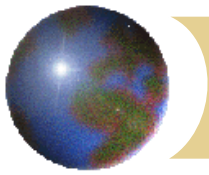
- ✚ Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées modalités.
- ✚ Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite **ordinale**.
- ✚ Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite **nominale**.



Variable qualitative

Traitements statistiques

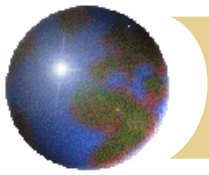
- ✚ On ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale).
- ✚ Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques.
- ✚ Il faut noter que les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).



Variable qualitative

Représentations graphiques

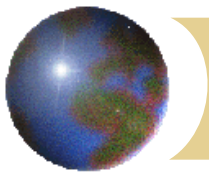
- ✚ Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses.
- ✚ Les trois plus courantes, qui sont aussi les plus appropriées, sont :
 - ✚ le diagramme en colonnes,
 - ✚ le diagramme en barre,
 - ✚ le diagramme en secteurs.
- ✚ Voir illustration ci-après.



Exemple de variable qualitative

- Le tableau ci-dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (Catégorie SocioProfessionnelle):

CSP	effectifs en milliers	fréquences (%)
1. agriculteurs exploitants	1312	6,1
2. artisans, commerçants, chefs d'entreprises	1739	8,1
3. cadres, professions intellectuelles supérieures	2267	10,6
4. professions intermédiaires	4327	20,1
5. employés	5815	27,0
6. ouvriers	6049	28,1



Représentations graphiques de l'exemple

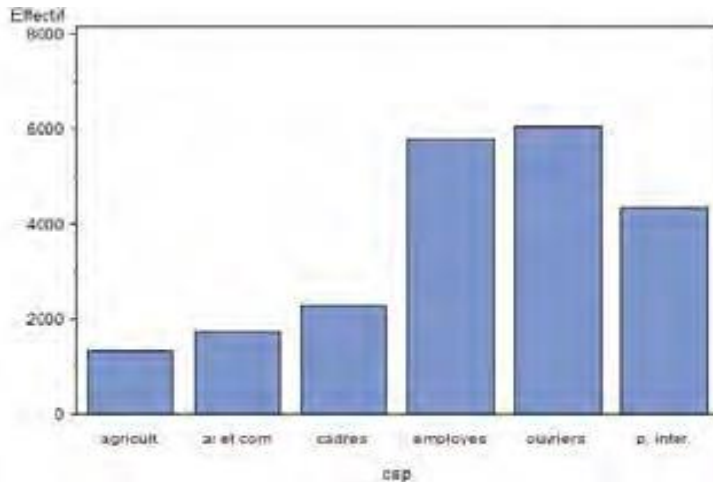


Diagramme en colonnes

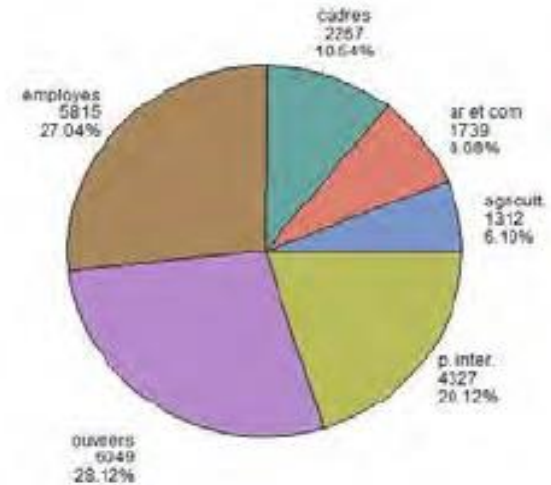


Diagramme en secteurs

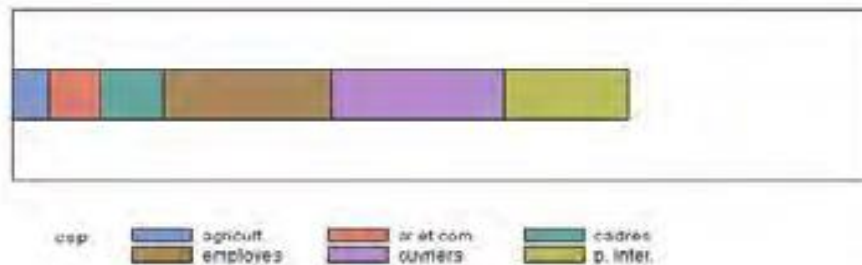


Diagramme en barre