Chapitre 3
Les indicateurs

On se place uniquement dans le cas d'une variable **quantitative.**

L'objectif est de résumer l'ensemble des observations par des indicateurs. Il est toujours insuffisant de résumer une série par un seul indicateur.

D'après Guy Brousseau, un modèle doit :

- représenter correctement les observations *(pertinence)*,
- être un résumé plus simple que les observations *(communicabilité)*,
- permettre de reconstituer au mieux l'ensemble des observations *(fidélité)*,
- permettre de comprendre les données, c'est-à-dire de les placer par rapport à des modèles familiers, universels et donc de permettre la comparaison avec d'autres modèles *(intelligibilité)*,
- être accessible au contrôle mathématique *(consistance)*.

I. Les caractéristiques de position ou de tendance centrale

I.1. Le mode

Pour une variable statistique discrète, **le mode** est la valeur la plus fréquente.

Lorsque la variable est continue, on parle de **classe modale** : c'est la classe correspondant « au pic » de l'histogramme (G. Saporta), autrement dit c'est la classe pour laquelle d_i est maximale.

Plus généralement, si X est une variable statistique (resp. aléatoire) absolument continue de densité f, on appelle mode toute valeur de la variable pour laquelle f est maximum.

Bien entendu, il peut y avoir plusieurs valeurs (resp. classes) modales.

Exercice

Le but de cet exercice est de montrer que la classe modale n'est pas nécessairement celle dont l'effectif est le plus grand.

La répartition des salaires annuels, exprimés en milliers d'euros (k €) de 90 employés d'une entreprise est donnée dans le tableau suivant :

Salaires en k €	[13 ; 15[	[15 ;16[	[16 ;17[	[17 ;18[	[18 ;20[	[20 ;22[	[22 ;24[
Effectifs (n_i)	12	12	14	15	17	12	8

Tracer l'histogramme et déterminer la classe modale.

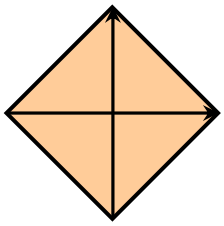
d'après Itinéraires en Statistiques et Probabilités (Ellipses)

1.2. Distances usuelles dans $\mathbb{R}^n$.

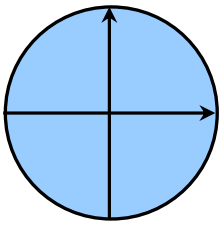
Pour les autres indicateurs de tendance centrale, il s'agit de résumer l'ensemble des observations par une valeur numérique relativement *proche*. La « proximité » se mesurant à l'aide de distances, il est utile de rappeler les distances usuelles dans $\mathbb{R}^n$.

Soit $X=(x_1,x_2,...,x_n)\in \mathrm{R}^n$.	Les distances associées sont :
On définit trois normes usuelles :	Pour tout $(X,Y)\in \mathrm{R}^n\times \mathrm{R}^n$:
$\|\| X \|\|_1=\sum_{i=1}^{n} \| x_i \|$	$d_1(X,Y)=\|\| X-Y \|\|_1$
$\|\| X \|\|_2=\left(\sum_{i=1}^{n} x_i^{\,2}\right)^{\frac{1}{2}}$	$d_2(X,Y)=\|\| X-Y \|\|_2$ (distance euclidienne)
$\|\| X \|\|_\infty=\underset{1\le i\le n}{\mathrm{Max}} \| x_i \|$	$d_\infty(X,Y)=\|\| X-Y \|\|_\infty$

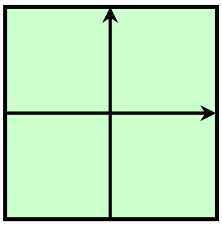
Pour chacune de ces distances, la boule unité dans le plan est donnée ci-après :



Distance d_1



Distance d_2



Distance d_∞

1.3. La moyenne

1.3.1. Cas d'une variable discrète

La moyenne $\bar{x}$ d'une série statistique est la somme de ses éléments divisée par leur nombre : $\bar{x}=\dfrac{x_1+x_2+...+x_n}{n}$.

Si $c_1,c_2,\dots,c_n$ sont les valeurs distinctes prises par les x_i et si n_i désigne l'effectif de la valeur c_i, on a :

$\bar{x}=\dfrac{\sum_{i=1}^{k} n_i c_i}{n}$ ou encore $\bar{x}=\sum_{i=1}^{k} f_i c_i$ avec $f_i=\dfrac{n_i}{n}$

La moyenne minimise la distance d_2. Cela signifie que, parmi les vecteurs "constants" $(a,a,\dots,a)$, le vecteur $\bar{X}=(\bar{x},\bar{x},\dots,\bar{x})$ est le plus proche du vecteur $X=(x_1,x_2,\dots,x_n)$ au sens de d_2.

Preuve :

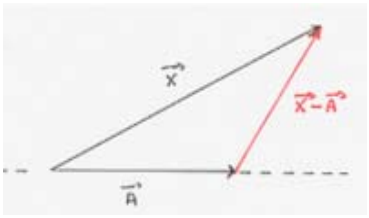
Soit $A=(a,a,...,a)$

$d_2^{\,2}(X,A)=\sum_{i=1}^{n}(x_i-a)^2=f(a)$

La fonction f est du second degré.

$$f(a) = na^2 - 2a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} {x_i}^2$$

f sera minimale pour $A = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$

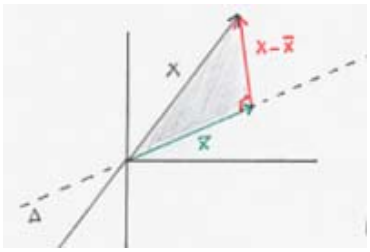


Alors $f(\overline{x}) = {d_2}^2(X, \overline{X}) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = nV(X)$

Le calcul de $f(\overline{x})$ donne alors :

$$nV(X) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} {x_i}^2 - (\overline{x})^2$$

Interprétation géométrique



${d_2}^2(X, A) = \| X - A \|_2$ sera minimum lorsque $\vec{A}$ est la projection orthogonale de $\vec{X}$ sur la droite vectorielle Δ engendrée par (1,1,...,1)

$\| X - \overline{X} \|_2^2 = nV(X) = ns^2$ où s est l'écrt-type.

Propriété : Linéarité de la moyenne

Soit deux séries statistiques $(x_1, x_2, \dots, x_n)$ et $(y_1, y_2, \dots, y_n)$ telles que, pour tout i de [1 ; n], on ait $y_i = ax_i + b$.
Alors $\overline{y} = a\overline{x} + b$.
En particulier, pour $a = 1$ et $b = -\overline{x}$, on a $\overline{y} = 0$. La nouvelle série a une moyenne nulle. On dit qu'on a centré les données x_i.

Propriété : Regroupement ou partition

Soit deux séries statistiques $(x_1, x_2, \dots, x_n)$ et $(y_1, y_2, \dots, y_m)$ de moyennes respectives $\overline{x}$ et $\overline{y}$.
Soit $(z_1, z_2, \dots, z_{n+m})$ la série obtenue par regroupement des deux séries précédentes et $\overline{z}$ sa moyenne.
Alors $\overline{z} = \frac{n\overline{x} + m\overline{y}}{n+m}$.

La preuve est immédiate.

Moyenne élaguée

La moyenne est sensible aux valeurs extrêmes. Pour pallier cet inconvénient, on peut décider de ne pas tenir compte des valeurs extrêmes dans le calcul de la moyenne.

Soit (x_1, x_2, ..., x_n) une série statistique et α un réel de [0 ; 1].
La moyenne élaguée de niveau $1-\alpha$ est la moyenne de la série privée d'un nombre de valeurs extrêmes égal à E($n\alpha$), soit à gauche, soit à droite, soit bilatéralement.
En principe α = 0,05 ou α =0,01.

1.3.2. Cas d'une variable continue

Le regroupement des valeurs en classes entraîne une perte d'information. Dans ce cas on ne peut calculer qu'**une valeur approchée de la moyenne.**

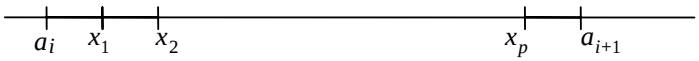
Pour trouver une telle valeur approchée, on considère que toutes les valeurs d'une classe sont rapportées au centre de cette classe. On remplace donc la série initiale par une série discrète.

Bien entendu, cette valeur approchée dépend de la nature du regroupement en classes effectué.

On pourrait prendre l'approximation de la distribution uniforme à l'intérieur d'une même classe : ce modèle conduit à la même valeur approchée que précédemment.

1.3.3. Comparaison des moyennes dans le cas d'une répartition uniforme

Les p valeurs $x_1, x_2, \ldots, x_p$ sont uniformément réparties sur l'intervalle $[\, a_i\, ,\, a_{i+1}\, [$ signifie :



$$\forall k \in [1\,;\,p],\ x_k = a_i + k\frac{(a_{i+1}-a_i)}{p+1}$$

Notons $\overline{x_i}$ la moyenne de $(x_1, x_2, \ldots, x_p)$.

$$p\,\overline{x_i} = \sum_{k=1}^{k=p} x_k = pa_i + \frac{(a_{i+1}-a_i)}{p+1}\sum_{k=1}^{p} k.$$

Or $\sum_{k=1}^{p} k = \frac{p(p+1)}{2}$. Donc $p\,\overline{x_i} = pa_i + p\,\frac{(a_{i+1}-a_i)}{2}$

soit $\overline{x_i} = \frac{a_i + a_{i+1}}{2}$. $\boxed{\overline{x_i} \text{ est donc le milieu du segment } [a_i\,;\,a_{i+1}[}$

Ainsi, quel que soit le modèle d'approximation choisi (regroupement au centre de classe ou répartition uniforme à l'intérieur de l'intervalle), la moyenne obtenue est la même.

Exercice

Lors du regroupement en classes de données abondantes, il y a évidemment perte d'information. Certes on peut espérer que les erreurs introduites par la concentration des données au centre de chaque classe se neutralisent dans le calcul de la moyenne, mais il n'en est pas toujours ainsi, comme le montre l'exemple suivant :

1. Dans une classe, la liste des notes obtenues à un devoir de mathématiques par les élèves classés par ordre alphabétique est la suivante :

8	16	9	18	9	11	9	13	7	3	14	7
10	10	10	17	13	14	10	13	5	15	13	19
10	6	12	5	12	1	9	9	8	8	4	

Déterminer une valeur approchée de la moyenne $\overline{x}$ de cette série statistique.

2. Le professeur décide de classer ses élèves en cinq groupes :

[0 ; 4 [	[4 ; 8 [	[8 ; 12 [	[12 ; 16 [	[16 ; 20 [
faible	médiocre	moyen	satisfaisant	très bon

Déterminer les effectifs de chaque classe.

En utilisant le centre des classes, calculer la moyenne $\bar{y}$ de cette série statistique.

3. Le professeur envisage une autre répartition et refait ses calculs avec le regroupement suivant :

[15 ; 20 [	[10 ; 15 [	[5 ; 10 [	[0 ; 5 [
très satisfaisant	convenable	insuffisant	très faible

Quelle est la moyenne $\bar{z}$ de cette dernière série statistique ?

Réponses : 1° : 10,2 ; 2° : 10,8 ; 3° : 10,5.

d'après Itinéraires en Statistiques et Probabilités (Ellipses)

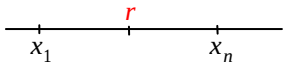
I.4. La moyenne des valeurs extrêmes

La moyenne des valeurs extrêmes d'une série $(x_1, x_2, \dots, x_n)$ est donnée par $r = \dfrac{\min(x_i) + \max(x_i)}{2}$.

Peu usité car très sensible aux valeurs extrêmes, elle minimise d_∞.

Cela signifie que, parmi les vecteurs "constants" $(a, a, \dots, a)$, le vecteur $(r, r, \dots, r)$ est le plus proche du vecteur $(x_1, x_2, \dots, x_n)$ au sens de d_∞.

r minimise la fonction f définie par $f(t) = \underset{1 \le i \le n}{\text{Max}} |t - x_i|$



I.5. La médiane

La médiane d'une série statistique ordonnée $(x_1, x_2, \dots, x_n)$ est x_p si $n = 2p+1$ et $\dfrac{x_p + x_{p+1}}{2}$ si $n = 2p$.

Dans le cas d'une variable continue, la pratique habituelle consiste à tracer la fonction de répartition en faisant l'hypothèse d'une répartition uniforme dans chaque intervalle puis d'exploiter cette représentation graphique pour déterminer l'antécédent de 0,5. D'après le document du GEPS sur les quantiles, cette pratique n'est pas usitée chez les statisticiens. Le GEPS préconise de parler de **classe médiane.**

« *La procédure qui consiste à tracer une courbe dite de fréquences cumulées croissante, continue, obtenue par interpolation linéaire à partir des valeurs F(ai) définies ci-dessus et à définir la médiane comme l'intersection de cette courbe avec la droite d'équation y=0,5, où avec une courbe analogue dite des fréquences cumulées décroissantes n'est pas une pratique usuelle en statistique et ne sera pas proposée au lycée.*

Si des données sont regroupées en classe, on parle de classe médiane. »

La médiane m_e minimise la distance d_1 . Cela signifie que, parmi les vecteurs « constants » $(a, a, \dots, a)$, le vecteur $M_e = (m_e, m_e, \dots, m_e)$ est le plus proche du vecteur $X = (x_1, x_2, \dots, x_n)$ au sens de d_1.

Preuve : $d_1(X,A)=\sum_{i=1}^{n}|a-x_i|=f(a)$

La fonction f est continue, dérivable sur chaque intervalle ne contenant pas x_i.

Pour tout $t \neq x_i$, $f'(t)$ sera négatif s'il y a plus de valeur x_i supérieures à t que de valeurs x_i inférieures... et $f'(t)$ sera nul s'il y a autant de valeurs x_i supérieures à t que de valeurs x_i inférieures. D'où le minimum est atteint pour $a = m_e$.

Le programme de 1ère S prévoit la notion de quartile. Le GEPS propose la définition suivante pour une notion plus générale de **quantile** :

> En statistique, pour toute série numérique de données à valeurs dans un intervalle I, on définit la fonction quantile Q, de [0,1] dans I, par : $Q(u) = \inf\{x, F(x) \geq u\}$, où $F(x)$ désigne la fréquence des éléments de la série inférieurs ou égaux à x.

Soit n la taille de la série ; si on ordonne la série par ordre croissant, $Q(u)$ est la valeur du terme de cette série dont l'indice est le plus petit entier supérieur ou égal à u.

Dans le cadre de cette définition, les trois quartiles sont $Q_1 = Q(0,25)$, $Q_2 = Q(0,50)$ et $Q_3 = Q(0,75)$. Les 9 déciles sont les valeurs de $Q(i/10)$, $i = 1...9$, les 99 centiles sont les valeurs de $Q(i/100)$, $i = 1...99$. On définit assez souvent la médiane m_e par $m_e = Q(0,5)$: la médiane est alors le second quartile, le cinquième décile, le cinquantième centile, etc....

Voir le document du GEPS... (PDF, 58 Ko)

2. Les caractéristiques de dispersion

2.1. L'étendue

> L'étendue de la série $(x_1, x_2, ..., x_n)$ est égale à : $\max(x_i) - \min(x_i)$.

Comme la moyenne des valeurs extrêmes, elle est très sensible à ces valeurs extrêmes.

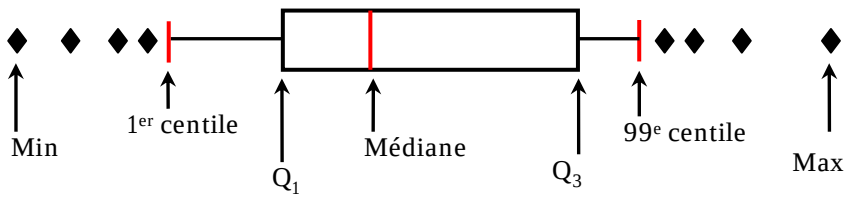
2.2. L'écart interquartile

> **L'écart interquartile** est la quantité $Q_3 - Q_1$.

2.3. Une autre représentation : « la boîte à moustaches ».

Elle est due à JW. Tukey et est appelée « box plot » en anglais.

Le dessin suffit à l'explication :



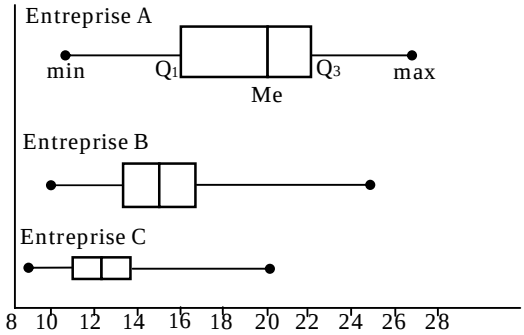
Pour comparer des populations qui n'ont pas le même effectif, on trace la largeur du rectangle **proportionnelle à la racine carrée de la population.**

Exercice

Comparer les salaires dans les trois entreprises suivantes d'un même secteur industriel.

Entreprise	Taille	min	Q1	Me	Q3	max
A	125	10 500	16 000	20 000	22 000	27 000
B	75	10 000	13 500	15 000	17 000	25 000
C	25	8 500	11 000	12 500	14 000	20 500

À partir des données, on obtient la représentation suivante :



d'après Itinéraires en Statistiques et Probabilités (Ellipses)

Dans les premiers diagrammes de Tukey, la longueur des « moustaches » est 1,5 fois l'écart interquartile. Les diagrammes de Tukey étaient utilisés dans des secteurs où les données peuvent le plus souvent être modélisées en utilisant une loi de Gauss ; dans ce cas, au niveau théorique, les extrémités des « moustaches » sont voisines du premier et 99e centile : ces diagrammes étaient surtout utilisés pour détecter la présence de données exceptionnelles. On utilise aujourd'hui les diagrammes en boites pour représenter des distributions empiriques de données quelconques, non nécessairement symétriques autour de la moyenne, et le choix de moustaches de longueurs 1,5 fois l'écart interquartile ne se justifie plus. (*Document d'accompagnement des programmes de 1re S*)

2.4. Variance et écart type

On a déjà rencontré la variance dans l'interprétation géométrique de la moyenne.

Pour une série $(x_1, x_2, \dots, x_n)$ de moyenne m, on définit la variance $V(X)$ et l'écart-type s par : $V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ et $s = \sqrt{V(X)}$

Propriétés

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - m^2$$

Pour tous a et b réels :

$$V(aX+b) = a^2\, V(X)$$

$$s(aX+b) = |a|\, s(X)$$

Exercice

On considère deux séries statistiques portant sur le même caractère :

- $(x_1, n_1), \ldots, (x_p, n_p)$, effectif total n, moyenne $\overline{x}$, écart-type σ_x ;
- $(y_1, m_1), \ldots, (y_q, m_q)$, effectif total m, moyenne $\overline{y}$, écart-type σ_y.

On note (z_k, r_k) la série statistique obtenue en regroupant les deux séries, $\overline{z}$ sa moyenne et σ_z son écart-type.

1. Montrer que $\overline{z} = \dfrac{n\overline{x} + m\overline{y}}{n+m}$.
2. Démontrer que : $(m+n)\,\sigma_z{}^2 = n\left(\sigma_x{}^2 + (\overline{x} - \overline{z})^2\right) + m\left(\sigma_y{}^2 + (\overline{y} - \overline{z})^2\right)$

   En déduire : $\sigma_z{}^2 = \dfrac{n\,\sigma_x{}^2 + m\,\sigma_y{}^2}{n+m} + \dfrac{n\,m}{(n+m)^2}(\overline{x} - \overline{y})^2$
3. Un professeur a corrigé n copies d'examen. La moyenne des notes est $\overline{x}$ et l'écart-type de la série de notes est σ.
   Une copie supplémentaire (à corriger) lui est attribuée. On désigne par y la note obtenue pour cette copie.
   Exprimer en fonction des données la moyenne et l'écart-type de la série de $(n+1)$ notes ainsi obtenue. Existe-t-il une valeur de y qui ne modifie pas la moyenne ? l'écart-type ? les deux ?

Dans le cas d'une variable continue et d'un regroupement par classes, on obtient une **valeur approchée** de la variance à l'aide de la modélisation utilisée pour obtenir une valeur approchée de la moyenne, c'est à dire en ramenant toutes les valeurs d'une classe au centre de cette classe. Rappelons que pour avoir une valeur approchée de la médiane, on avait utilisé la modélisation de la répartition uniforme par classe.

2.5. À propos du regroupement en classes

2.5.1. Comparaison des variances

Soit $(x_1, x_2, \ldots, x_n)$ une série répartie en m classes $[a_1 ; a_2[\,;[a_2 ; a_3[\ldots[a_m ; a_{m+1}[$ d'effectifs respectifs n_i pour $1 \le i \le m$ (et $\sum_{i=1}^{m} n_i = n$).

On note $\overline{x}$ la moyenne obtenue par l'un ou l'autre des modèles choisis.

On note σ l'écart-type réel de la série, σ_1 l'écart-type obtenu en ramenant les valeurs au centre de chaque classe, et σ_2 l'écart-type obtenu en supposant une répartition uniforme à l'intérieur de chaque classe.

La variance σ^2 est systématiquement sous estimée par $\sigma_1{}^2$, car on néglige la variation à l'intérieur de chaque classe ($\sigma_1 < \sigma$).

En revanche, $\sigma_2{}^2$ est en général assez proche de σ^2.

Comparons σ_1 et σ_2 :

$$n\,\sigma_2{}^2 = \sum_{k=1}^{n}\left(x_k - \overline{x}\right)^2$$

$$n\,\sigma_1{}^2 = \sum_{i=1}^{m} n_i\left(c_i - \overline{x}\right)^2 \quad \text{où} \quad c_i = \frac{a_i + a_{i+1}}{2}$$

$n\sigma_2{}^2 = \sum_{i=1}^{m}\left(\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2\right)$ en regroupant les valeurs par classe.

D'après ce qui précède, pour la ième classe, on a :

$x_k = a_i + \frac{k\,d_i}{n_i+1}$ où $d_i = a_{i+1} - a_i$.

$$\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2 = \sum_{k=1}^{n_i}\left(x_k - c_i + c_i - \bar{x}\right)^2 =$$

$$= n_i(c_i - \bar{x})^2 + \sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 + 2(c_i - \bar{x})\sum_{k=1}^{n_i}\left(x_k - c_i\right).$$

Or $n_i c_i = \sum_{k=1}^{n_i} x_k$ donc le dernier terme est nul.

$$\sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 = \sum_{k=1}^{n_i} d_i{}^2\left(\frac{k}{n_i+1} - \frac{1}{2}\right)^2$$

$$= \frac{d_i{}^2}{(n_i+1)^2}\sum_{k=1}^{n_i}k^2 + \frac{1}{4}n_i\,d_i{}^2 - \frac{d_i{}^2}{n_i+1}\sum_{k=1}^{n_i}k$$

$$= \frac{d_i{}^2}{(n_i+1)^2}\times\frac{n_i(n_i+1)(2n_i+1)}{6} + \frac{1}{4}n_i\,d_i{}^2 - \frac{d_i{}^2}{n_i+1}\times\frac{n_i(n_i+1)}{2}$$

$$= \frac{n_i\,d_i{}^2(2n_i+1)}{6(n_i+1)} - \frac{1}{4}n_i\,d_i{}^2$$

$$= \frac{n_i\,d_i{}^2}{12}\left(\frac{n_i-1}{n_i+1}\right)$$

Finalement, $n\,\sigma_2{}^2 = \sum_{i=1}^{m} n_i(c_i - \bar{x})^2 + \sum_{i=1}^{m}\frac{n_i\,d_i{}^2}{12}\left(\frac{n_i-1}{n_i+1}\right)$

Soit : $\sigma_2{}^2 = \sigma_1{}^2 + \sum_{i=1}^{m}\frac{n_i\,d_i{}^2}{12n}\left(\frac{n_i-1}{n_i+1}\right)$

D'après ce qui précède, pour la ième classe, on a :

$x_k = a_i + \frac{k\,d_i}{n_i+1}$ où $d_i = a_{i+1} - a_i$.

$$\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2 = \sum_{k=1}^{n_i}\left(x_k - c_i + c_i - \bar{x}\right)^2 =$$

$$= n_i(c_i - \bar{x})^2 + \sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 + 2(c_i - \bar{x})\sum_{k=1}^{n_i}\left(x_k - c_i\right).$$

2.5.2. Regroupement en classe et précision demandée dans les exercices donnés

Lorsque l'on propose, en temps limité, un exercice de statistique, on est conduit à limiter le nombre de données à traiter (pour diminuer les problèmes de saisie et de calcul). Pour cela on regroupe fréquemment les données en un petit nombre de classes. Il est important de veiller à ce que ce regroupement (qui constitue toujours une perte d'information) soit bien compatible avec la précision demandée par la suite et que les réponses aux questions posées puissent être trouvées sans ambiguïté.

Document de J.P. POUGET IA-IPR, académie de Créteil (PDF, 198 Ko)

2.6. Inégalité de Bienaymé-Tchebychev

Soit une série statistique de moyenne m et d'écart type s. Pour tout α réel strictement positif, on note f_α la fréquence des valeurs comprises entre $m - \alpha s$ et $m + \alpha s$ (c'est-à-dire $|X - m| \leq \alpha s$).

Alors $f_\alpha > 1 - \frac{1}{\alpha^2}$

Cette inégalité, bien que médiocre, est valable quelle que soit la série statistique.

Ainsi plus de 75 % des valeurs sont dans $[\, m - 2s \,;\, m + 2s \,]$,

plus de 88 % des valeurs sont dans $[\, m - 3s \,;\, m + 3s \,]$,

plus de 93,75 % des valeurs sont dans $[\, m - 4s \,;\, m + 4s \,]$,