Chapitre 3

Réduction des données : les paramètres de position et de dispersion

Il est impossible de se contenter de donner l'allure générale d'une série statistique. Les représentations sont riches d'informations mais elles sont parfois lourdes et permettent rarement une discussion ou une comparaison pertinente des différentes séries considérées. Par ailleurs, elles ne disent rien de la population parente.
Notre objectif désormais est de déterminer un nombre limité de valeurs qui permette de résumer de façon satisfaisante certaines propriétés de nos séries.

Commençons par fixer notre vocabulaire :
Supposons que N désigne le nombre d'individus dans la population (souvent inconnu) et que $S = \{a_1, a_2, ..., a_n\}$ soit l'ensemble des résultats observés, non classés.

Définition : On appelle ***série statistique*** tout p-uplets de la forme $((x_i, n_i))_{1 \leq i \leq p}$ où x_i désigne les différentes valeurs observées prises parmis S dans le cas de distributions non groupées (respectivement les points centraux dans le cas de distributions groupées) et n_i désigne la fréquence absolue correspondante, à savoir le nombre d'individus de la série pour lequel le caractère x_i est observé.

- Par convention $x_1 < x_2 < ... < x_p$
- L'entier $n = \sum_{i=1}^{p} n_i$ est appelé ***effectif total***

3.1 Les paramètres de position

3.1.1 Le mode

La manière la plus naïve de résumer une série statistique est de considérer la valeur (respectivement la classe de valeurs) l**a plus fréquente** prise par le caractère observé. On parle de **mode**, (respectivement de **classe modale**) de la série statistique. L'avantage est qu'il n'y a pas besoin de calcul, la réponse se lit immédiatement sur la série ou même sur l'histogramme si vous l'avez déjà tracé !

Exemple 2.1 : dans le cas du diamètres des arbres de la zone décrite en 2.1.2., le mode vaut 16 car la fréquence relative de ce diamètre est maximale et vaut $3/18 = 0,16$ et dans le cas de la distribution groupée la classe modale vaut $[10, 20[$.

Définition : Si cette valeur (respectivement cette classe) est unique, on parle de **série unimodale** sinon, dans le cas d'un histogramme en "dos de chameau" ou en "montagnes russes", avec plusieurs maximums locaux, on parle de **série plurimodale**.

3.1.2 La moyenne arithmétique

C'est souvent la première valeur qui vient à l'esprit pour résumer une série. Qui n'a pas en tête un prof de maths, acariâtre comme il se doit, assènant « Ce devoir est nul, vous n'avez rien compris », avis péremptoire justifié d'un seul chiffre : « la moyenne est de 6 ».

Evacuons la question de la pertinence de ce chiffre pour dire quelque chose de la qualité du travail des étudiants et revenons au moyen de le calculer et à sa signification.
Est-ce utile, de le rappeler ? Il suffit de faire la somme des valeurs, divisée par l'effectif total.

Notation : Nous la désignerons par le symbole habituel $\bar{x}$ lorsque nous travaillons sur un échantillon statistique et par μ lorsque nous prenons en compte l'ensemble des valeurs de la population.

Formule : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$ et $\mu = \frac{1}{N}\sum_{i=1}^{N} a_i$

Remarque 1 : A bien des égards, elle s'apparente au barycentre des points x_i affectés des coefficients $f_i = \frac{n_i}{n}$ et résume la série statistique au même titre que le centre de gravité, en mécanique, réduit votre objet à un point...

Règle d'arrondi : Eviter les arrondis pour les calculs intermédiaires et utillisez pour le résultat final une décimale de plus que dans les données originales.

Remarque 2 : Cette valeur est **toujours unique** et possède une propriété intéressante : elle minimise la fonction ϵ où $\epsilon(x)$ mesure la dipersion des résultats, conçue comme somme des carrés des écarts au réel x. (Je laisse le soin au lecteur méticuleux et avide de justifications de vérifier que $\epsilon'(\bar{x}) = 0$)

$$\epsilon(x) = (a_1 - x)^2 + (a_2 - x)^2 + ... + (a_N - x)^2$$

Cette fonction est positive et ne s'annule que dans le cas où tous vos résultats sont égaux... sans dispersion donc, ce qui est plutôt rassurant sur le plan sémantique.

Définition : La moyenne d'une série statistique $\bar{x}$ est aussi proche que possible, au sens des moindres carrés, des nombres a_i où $(a_i)_{1 \leq i \leq N}$ désigne la série des résultats non classés (càd avec d'éventuelles répétitions...). La moyenne est aussi appelée **mesure de tendance centrale**

3.1.3 la médiane

Définition : La **médiane** est un paramètre de position tel que la moitié des observations lui sont inférieures ou égales et l'autre moitié lui sont supérieures

La moyenne a un défaut car elle dépend de chacune des valeurs de votre échantillon et est particulièrement sensible aux valeurs extrêmes. La médiane, qui sépare en deux parties d'effectifs identiques une série d'observations, est bien moins sensible à ces valeurs.
Elle peut sembler plus "naturelle" que la moyenne d'autant plus qu'**elle minimise aussi les écarts à tous les résultats obtenus**. Une façon de formaliser cet énoncé est de dire que la médiane est un minimum de la fonction ϵ_1 égale à la somme des écarts absolus à un réel x de chacun de vos résultats.

$$\epsilon_1(x) = |a_1 - x| + |a_2 - x| + ... + |a_N - x|$$

Dans la pratique, deux cas sont à distinguer pour la calculer :

Dans le cas discret

On commence par **trier les résultats de façon croissante** et on note $S' = \{a'_1, a'_2, ..., a'_n\}$ l'ensemble des résultats observés ainsi organisés. Si la série comporte un nombre impair de valeurs, $n = 2n_1 + 1$, alors la médiane est la $(n_1 + 1)^{\text{ième}}$ valeur.
Si la série comporte un nombre pair de valeurs, on **convient** de prendre pour médiane la demi-somme de la $n_1^{\text{ième}}$ et de la $(n_1 + 1)^{\text{ième}}$ valeur.

Exemple 2.2 :
- Si $S' = \{1; 2; 2; 5; 6; 7; 9\}$ alors la médiane vaut : 5 ;
- Si $S' = \{1; 2; 5; 6; 7; 9\}$ alors la médiane vaut : $(5 + 6)/2 = 5, 5$

✎ ***Remarque 1 :*** Vous noterez que dans le cas d'un effectif total pair ($n = 2n_1$), toute valeur entière comprise entre a'_{n_1} et a'_{n_1+1} vérifie les propriétés de la médiane et pas seulement le milieu entre ces deux extrémités...
En effet, $\forall x \in [a'_{n_1}; a'_{n_1+1}[$:

$$\begin{aligned}
\epsilon_1(x) &= |a'_1 - x| + |a'_2 - x| + ... + |a'_n - x| \\
&= \sum_{i=1}^{n_1}(x - a'_i) + \sum_{i=n_1+1}^{n}(a'_i - x) \\
&= \big(2n_1 - n\big)x + \Big(\sum_{i=n_1+1}^{n} a'_i - \sum_{i=1}^{n_1} a'_i\Big) \\
&= \big(2n_1 - 2n_1)\big)x + \Big(\sum_{i=r+1}^{n} a'_i - \sum_{i=1}^{r} a'_i\Big) \\
&= \Big(\sum_{i=r+1}^{n} a'_i - \sum_{i=1}^{r} a'_i\Big)
\end{aligned}$$

La fonction ϵ' est bien constante sur l'intervalle $]a'_{n_1}; a'_{n_1+1}[$... la valeur médiane n'est pas unique et rend difficile son utilisation.

♠ ***Remarque 2*** : Dans les cas, nombreux, pour lesquels la valeur de la médiane se répète, cette dernière perd sa pertinence statistique. Prenons deux exemples simples :
- Dans le cas impair : Soit $S' = \{12, 45, 45, 45, 52\}$. La valeur médiane devrait être 45 mais 80% des valeurs lui sont inférieures et la même proportion lui sont supérieures...
- Dans le cas pair : Soit $S' = \{12, 45, 45, 45, 52, 60\}$. Le calcul précédent nous donne une médiane égale à 45 (valeur moyenne de 45 et de 45...) mais ce sont 66% des valeurs qui sont lui sont inférieures et 66% des valeurs qui sont lui sont supérieures...

Si la valeur médiane se répète, celle-ci ne sera pas utilisée dans la pratique.

Dans le cas continu

Les calculs menés pour tracer la courbe des fréquences cumulées permettent dans tous les cas d'obtenir la médiane qui cette fois, est unique : il s'agit de l'abscisse $x_{\text{méd}}$ du point du polygône des fréquences cumulées dont l'ordonnée vaut 0.5
S'il existe une valeur dont la fréquence cumulée vaut 0.5 on retiendra cette valeur pour valeur médiane, sinon, on aura recours à la méthode d'interpolation, à savoir :
(i) Soit $x_{\text{méd}}$ la valeur médiane cherchée.
(ii) On suppose que $\exists i \in \{1, ..., N\}/F_c(a'_i) < 0.5$ et $F_c(a'_{i+1}) > 0.5$.
(iii) Alors l'interpolation linéaire consiste à écrire :

$$\frac{0.5 - F_c(a'_i)}{x_{\text{méd}} - a'_i} = \frac{F_c(a'_{i+1}) - F_c(a'_i)}{a'_{i+1} - a'_i}$$

(iv) Ce qui permet d'obtenir la valeur médiane x_{md} cherchée.

Propriétés d'asymétrie

Lorsque la distribution de votre échantillon est symétrique, la moyenne, la médiane et le mode sont sensiblement les mêmes. Dans le cas contraire (asymétrie), celle ci est mesurable grâce au coefficient d'asymétrie suivant, où μ_3 désigne le moment centré d'ordre 3 :

$$\boxed{\alpha = \frac{\mu_3}{\sigma^3}}$$

Il est interressant dans ce cas de noter les positions relatives des trois paramètres de position que sont le mode, la moyenne et la médiane :

- On parle d'**asymétrie négative** ou d'**asymétrie à gauche** lorsque la queue de la distribution est plus longue à gauche qu'à droite. Dans ce cas, la médiane et la moyenne sont à gauche du mode et la moyenne est le plus souvent inférieure à la médiane. Le coefficient α est négatif.
- On parle d'**asymétrie positive** ou d'**asymétrie à droite** lorsque la queue de la distribution est plus longue à droite qu'à gauche. La médiane et la moyenne sont à droite positif.

Exemple 2.3 :

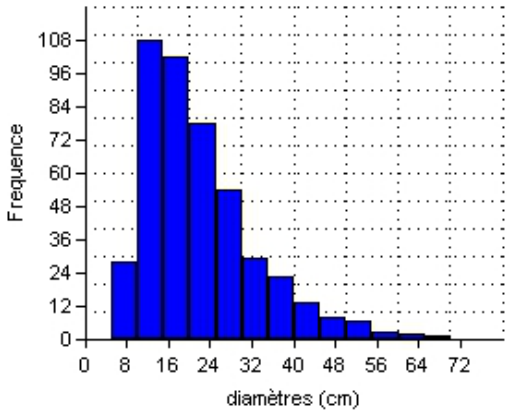

FIG. 3.1 – diamètres (cm) ; pas de 5 cm

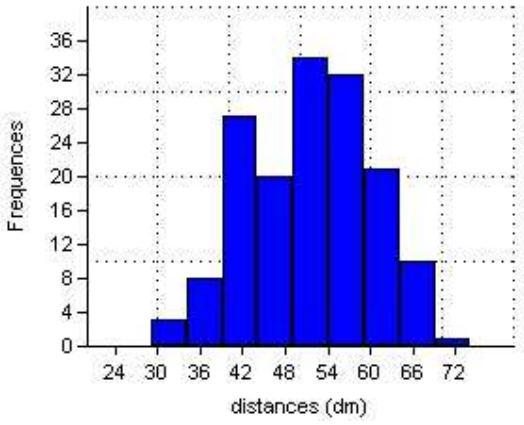

FIG. 3.2 – distances (dm) ; pas de 5 dm

Dans la figure 3.1 la classe modale est $[10; 15[$, la moyenne vaut $22,04$ et la médiane $19,69$. Le coefficient d'asymétrie vaut ici $1,23$. On dit que l'asymétrie est droite.

Dans la figure 3.2 la classe modale est $[49; 54[$, la moyenne vaut $51,50$ et la médiane $51,71$. On pourra dire que la distribution est approximativement symétrique. Pour une étude plus fine, on calculera le coefficient d'asymétrie qui vaut ici -0.06. L'asymétrie est donc gauche.

3.1.4 les quartiles

On les utilise essentiellement dans le cas des variables continues. Le **premier quartile** désigne l'abscisse q_1 du point du polygône des fréquences cumulées dont l'ordonnée vaut $1/4$: 25% des observations lui sont inférieures.
Le **troisième quartile** désigne l'abscisse q_3 du point du polygône des fréquences cumulées dont l'ordonnée vaut $3/4$: 75% des observations lui sont inférieures et 25% lui sont supérieures.

3.2 les paramètres de dispersion

Les paramètres de position sont souvent insuffisant pour caractériser les données recueillies. Nous savons tous qu'une même moyenne peut relever de deux échantillons distincts si on songe à la plus ou moins grande dispersion autour de cette moyenne.

3.2.1 L'amplitude

L'amplitude ou l'étendue d'une série est la plus intuitive des mesures de dispersion.

Formule (amplitude) : $a = max(S') - min(S')$

Cette valeur a le défaut majeur de dépendre exclusivement des données extrêmes, alors même que leur rôle est peut-être mineur dans notre série d'observations. Ainsi, dans l'exemple de la forêt du Gâvre, connaître les diamètres extrêmes d'une parcelle donnée offre bien peu d'informations sur la distribution des diamètres entre ces valeurs...
Il est néanmoins possible d'affiner cette approche en calculant l'**écart-interquartile**

Formule (écart-interquartile :) $\delta = q_3 - q_1$

50% des données sont comprises entre q_1 et q_3. Ainsi, plus δ est faible, moins la série est dispersée.

3.2.2 l'écart-absolu moyen

Il est associé au calcul de la médiane. On la vu, cette dernière minimise la fonction ϵ' définie par $\epsilon_1(x) = |a'_1 - x| + |a'_2 - x| + ... + |a'_n - x|$.

Formule (Ecart-absolu moyen)

$$\epsilon = \epsilon_1(x_{\mathbf{méd}}) = |a'_1 - x_{\mathbf{méd}}| + |a'_2 - x_{\mathbf{méd}}| + ... + |a'_n - x_{\mathbf{méd}}|$$

3.2.3 l'écart-type

C'est le plus utilisé des indices de dispersion.

♠ ***Attention*** : On prendra soin, comme pour la moyenne, de distinguer l'écart-type de l'échantillon (noté s) de l'écart-type de la population (noté σ).

Définition : 1. On appelle **écart-type de la population** le réel positif noté σ_x défini par :

$$(\sigma)^2 = \frac{1}{N}\big((a_1 - \bar{x})^2 + (a_2 - \bar{x})^2 + ... + (a_N - \bar{x})^2\big) = \sum_{i=1}^{N} \frac{(a_i - \bar{x})^2}{N}$$

Le réel $V = (\sigma)^2$ est appelée la **variance** de la population. Il s'agit de la moyenne des carrés des écarts à la moyenne.

Définition : 2. On appelle **écart-type de l'échantillon** S le réel positif noté s défini par :

$$s^2 = \sum_{i=1}^{n} \frac{(a_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{p}[n_i(x_i - \bar{x})^2]$$

On appelle **variance de l'échantillon** le carré de l'écart-type de l'échantillon S. C'est un estimateur **non biaisé** de la variance σ^2 de la population.

✎ ***Remarque 1*** : σ^2, on s'en souvient, est le minimum de la fonction ϵ définie en 3.1.2, atteint pour $x = \bar{x}$. La propriété caractéristique de la moyenne arithmétique $\bar{x}$ est donc que la moyenne des carrés des écarts à tout autre nombre sera suprieure ou égale à $(\sigma)^2$ Attention car l'écart-type est particulièrement sensible aux valeurs extrêmes ou aberrantes et il est bon de les déceler avant de discuter ses résultats.

Règle d'arrondi : Eviter les arrondis pour les calculs intermédiaires et utillisez pour le résultat final une décimale de plus que dans les données originales.

✎ ***Remarque 2*** : Si les données ont une unité, la moyenne et l'écart-type ont même unité mais la variance à comme unité le carré de celle-ci. Ce qui explique qu'on considère le plus souvent l'écart-type, racine carrée de la variance, pour toute interprétation de différences de dispersion sur des échantillons issus d'une même population et donnés dans la même unité.

Si on souhaite comparer les dispersion de deux échantillons pris dans deux populations différentes qui peuvent ne pas être dans la même unité, pour peu que la moyenne soit positive, on utilisera le coefficient de variation qui est lui sans unité :

Définition : On appelle **coefficient de variation** ou **coefficient de variabilité** pour un ensemble de données et on note cv l'expression de l'écart-type en valeur relative ou en pourcentage rapporté à la moyenne, si cette dernière est positive. Soit :

$$cv = s/\bar{x} \text{ ou } cv = \frac{s}{\bar{x}} \times 100$$

Exemple 2.4 : Si on considère à nouveau les diamètres des arbres et la distance au plus proche voisin dont les histogrammes sont fournis dans l'exemple 2.3 p. 21, on trouve respectivement pour $cv1 = 49\%$ et $cv2 = 16\%$. On peut donc conclure que la dispersion sur les diamètres des arbres de la parcelle considérée, prise à 1 mètre du sol, est supérieure à celle des distances aux plus proches voisins.

♠ ***Remarque 3*** : On se méfiera des résultats fournis par les calculatrices, Excel ou Matlab qui, lorsqu'on demande la variance, proposent la valeur de s^2, à savoir $\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{x})^2$. Sous Matlab, l'écart-type par défaut est celui de l'échantillon ($sdt(S)$ ou $std(S,0)$). Si on souhaite connaître l'écart-type de la population, on appellera la fonction « $std(S,1)$ ». Utiliser pour un échantillon S $std(S,1)$ au lieu de $std(S)$ n'est pas dramatique mais il faut savoir que le premier est un meilleur estimateur de σ que le second qui a tendance à proposer des variances d'échantillons qui sous estiment la variance de la population...

3.3 Nouvelles représentations graphiques

Imaginons plusieurs séries de données que nous souhaitons comparer. Il est évidemment possible de calculer leurs mesures de position et de dispersion qu'on rapportera dans un tableau mais il est préférable de donner une image qui facilite le commentaire de vos résultats.

Voilà, à titre d'exemples, deux modes de représentations qui permettent de remplir cet objectif. Je considère cette fois encore les séries statistiques issues de parcelles de la forêt du Gâvre et présentées en 2.1 :

3.3.1 diagrammes en boîte : médiane et quartiles

Ce type de représentation est utile si on s'intéresse à la dispersion des données, à la mise en évidence de valeurs aberrantes ou encore si on cherche à comparer deux séries statistiques associées à une même caractéristique.

Ce diagramme, appelé aussi **boîte à moustache** et chez les anglo-saxons **box plot** résume visuellement les données suivantes :

- les premier et troisième **quartiles** Q_1 et Q_3 forment les extrémités de la boîte dont la longueur est l'*intervalle interquartile* $IQ = Q_3 - Q_1$. 50% des données se trouvent donc à l'intérieure de la boîte.

- La **médiane** est indiquée par un trait horizontal à l'intérieur de la boîte.
50% des données lui sont donc à la fois supérieures et inférieures.

- Les **valeurs extrêmes** de la série qui s'écartent de plus de $1.5 \times (Q_3 - Q_1)$ de Q_1 ou de Q_3, sont représentées par un cercle si elles sont inférieures à $Q_3 + 3 \times (Q_3 - Q_1)$ ou supérieures à $Q_1 - 3 \times (Q_3 - Q_1)$, par un astérix au delà. Toutes les autres valeurs sont comprises entre les deux traits horizontaux qui constituent la « moustache » de la boîte. Dans la plus grande partie des cas, l'extrémité des moustaches désignent donc le minimum et le maximum de la série.

Résumons en disant que la boîte permet de visualiser au moins cinq valeurs caractéristiques de la série statistique : la médiane, le premier et le troisième quartile, les données de la série situées dans l'intervalle $[Q_1 - 1.5 \times (Q_3 - Q_1); Q_1[$ et dans l'intervalle $[Q_3; Q_3 + 1.5 \times (Q_3 - Q_1)]$ ainsi que les **valeurs extrêmes** dont il s'agira de discuter.

✐ ***Remarque :*** Le coefficient 1.5 proposée par l'inventeur de ce mode de représentation (John W. Tukey, 1977) a une raison probabiliste. Si la série de données suit une loi normale, alors la zone délimitée par la boîte à moustache devrait contenir 99.3% des observations. On ne devrait donc trouver que 0.7% de valeurs hors de ces bornes, qu'on considère comme « atypiques » ou « aberrantes ». Si le coefficient valait 1, la probabilité serait de 0.957 et elle vaudrait 0.999 si le coefficient était égal à 2. Le coefficient 1.5 apparaît donc comme un compromis pour ne rejeter qu'un nombre raisonnable de valeurs.[1]

justification : Toute loi normale pouvant être centrée réduite, nous supposons que $X \hookrightarrow \mathcal{N}(0,1)$. Pour déterminer Q_3 on se rapporte à la table fournie en annexe A.2 dans laquelle on peut lire : $\phi(0,67) = 0,7486$ et $\phi(0,68) = 0,7517$. Par interpolation linéaire, on obtient que :

$$\frac{Q_3 - 0,67}{0,68 - 0,67} = \frac{0,75 - O,7486}{0,7517 - 0,7486} \Leftrightarrow Q_3 = 0,675$$

Dès lors, sachant que $\phi(Q_1) = 0,25 = 1 - \phi(Q_3)$ on en déduit que $Q_1 = -Q_3$ et donc :

$$\begin{aligned} p &= \mathbb{P}[Q_1 - 1,5(Q_3 - Q_1) \leq X \leq Q_3 + 1,5(Q_3 - Q_1)] \\ &= \mathbb{P}[-Q_3 - 1,5(2Q_3) \leq X \leq Q_3 + 1,5(2Q_3)] \\ &= \mathbb{P}(-4Q_3 \leq X \leq 4Q_3) = 2 \cdot \phi(4Q_3) - 1 = 2 \cdot \phi(2,7) - 1 \\ &= 2 \cdot 0,99653 - 1 = 0,993 \end{aligned}$$

[1] « La boîte à moustache de Tukey : un outil pour initier à la Statistique », Monique Le Guen, CNRS-Matisse, *in* http ://matisse.univ-paris.fr/leguen/leguen2001b.pdf

Exemple 1 : Notes du dernier devoir de maths en BCPST2.

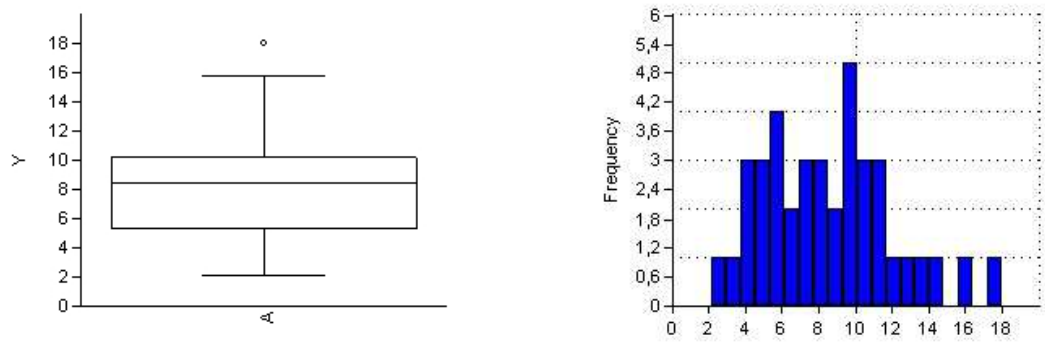

FIG. 3.3 – Boîte à moustache et histogramme des notes

commentaire : Les informations qui viennent compléter l'histogramme des notes sont les suivantes : 50% des étudiants ont une note supérieure à $8,4$ et la moitié d'entre eux à sa note comprise entre $5,3$ et $10,2$. La note minimum vaut $2,1$ et la note maximum vaut 18, considérée comme une valeur extrême. On vérifie que cette note a peu d'impact sur la moyenne car si on supprime le 18, elle passe de $8,46$ à $8,21$.
La positition de la médiane dans la boîte montre une forte asymétrie. On parle d'asymétrie *négative* car la médiane est plus proche de Q_3 que de Q_1 (la « queue » est plus longue à gauche qu'à droite). Retenons que les notes sont moins denses entre $5,3$ et $8,4$ qu'entre $8,4$ et $10,2$.

Exemple 2 : Distance au plus proche voisin, forêt du Gâvre.

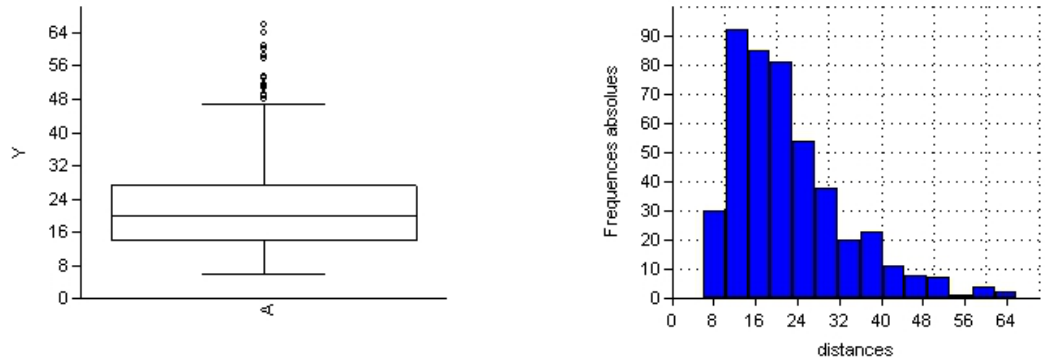

FIG. 3.4 – Boîte à moustache et histogramme des distances

♠ ***commentaire*** : Les valeurs extrêmes ont un impact important sur les calculs de la moyenne et de l'écart-type. Si l'une des valeurs extrême vous semble aberrante, il faut le justifier et l'évaquer des calculs statistiques. Sinon, si ces valeurs extrêmes sont correctes, il sera bon d'en étudier l'impact en traçant les graphiques avec et sans cette valeur, en recalculant les statistiques avec ou sans ces valeurs.

Dans l'exemple ci-dessus, la moyenne passe de $22,03$ à $20,91$; l'écart-type passe de $10,89$ à $9,13$. La différence est significative et il s'agit d'en rendre compte.
En retournant sur le terrain, on note que la parcelle choisie n'était pas homogène. Une partie avait été exploitée récemment et les distances entre les arbres y étaient sensiblement plus grandes. En partitionnant nos données sur les deux sous-parcelles, on obtient les boîtes à moustaches suivantes :

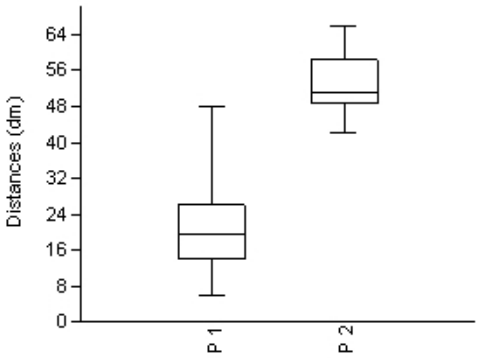

FIG. 3.5 – Boîte à moustache sur deux sous-parcelles 1 et 2

♠ ***Remarque*** : Si on souhaite comparer deux séries de données, ou plus, on place nos boîtes parallèlement...

Exemple 3 : Débits annuels de la Loire.

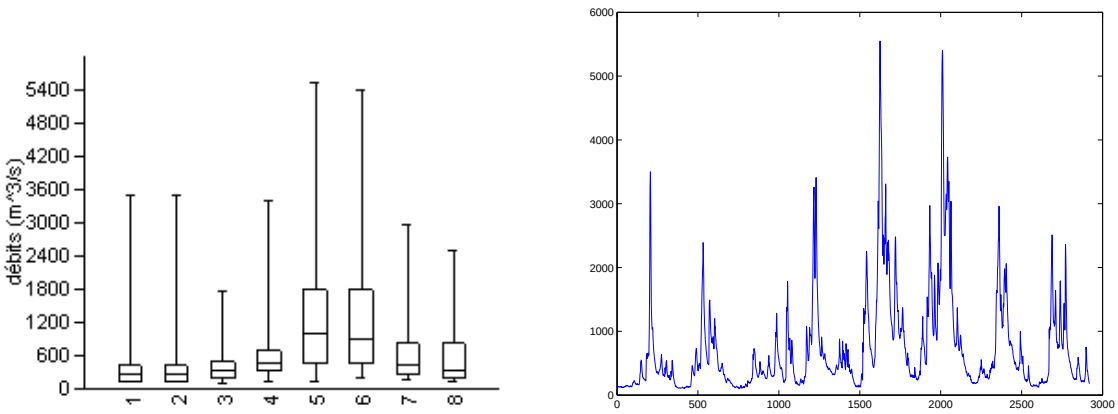

FIG. 3.6 – Boîte à moustache et histogramme des débits regroupés sur 8 années

3.3.2 diagramme en bâtons : moyennes et écart-types

Il est souvent intéressant de présenter un diagramme qui réduise une série de données à sa moyenne et à son écart-type mais deux cas sont à distinguer pour le commenter, selon que votre échantillon a une répartition « en cloche » ou pas. Précisons :

Distributions en cloche

Rappelons que nous connaissons des courbes « en cloches », étudiées dans le cadre des variables aléatoires à densités... Leur équation est :

$$\phi_{(m,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},\ \forall x \in \mathbb{R}$$

Nous avons vu en cours qu'elles sont symétriques par rapport à l'axe verticale d'équation $(x = m)$, elles sont strictement décroissante sur $[m; +\infty[$, possèdent un point d'inflexion au point d'abscisse $x = \sigma$, elles sont concaves sur $[m; m + \sigma]$, convexes sur $[m + \sigma; +\infty[$ et ont leur maximum en $x = m$...
Les présentations étant faites, nous appelerons désormais leurs représentations « courbes de Gauss », plus intimement des « gaussiennes »...

Imaginons maintenant que l'histogramme des fréquences relatives de notre distribution suive approximativement une telle courbe, ce qui est possible puisque toutes les fréquences sont comprises entre 0 et 1 et que l'aire des rectangles formant l'histogramme vaut 1... On peut alors supposer que nos données sont les réalisations indépendantes d'une variable aléatoire X qui suivrait la loi normale de mêmes paramètres que l'échantillon, à savoir $\bar{x}$ et s et dans ces conditions, l'histogramme peut être interprété comme un estimateur de la densité.

Dès lors, et d'après la table de la loi normale centrée réduite fournie en annexe A.2, on a les probabilités suivantes :

- $\mathbb{P}(\bar{x} - s \leq X \leq \bar{x} + s) = \mathbb{P}(-1 \leq X^* \leq 1) = 2 \cdot \varphi(1) - 1 = 0,6826$
- $\mathbb{P}(\bar{x} - 2s \leq X \leq \bar{x} + 2s) = \mathbb{P}(-2 \leq X^* \leq 2) = 2 \cdot \varphi(2) - 1 = 0,9544$

La méthode pour présenter ces résultas consiste à introduire sur les graphes une **barre d'erreur** d'amplitude $2s$ indiquant l'intervalle $]\bar{x} - s; \bar{x} + s[$.

Vous indiquez ainsi au lecteur, de façon implicite, qu'environ 68% **des valeurs sont situés dans cet intervalle.**

Remarque : Comment justifier que votre échantillon suit bien « approximativement » une loi normale ?

- La première méthode consiste à tracer sur votre histogramme la densité de la loi normale de même moyenne et de même écart-type et à mesurer l'asymétrie. Si cette dernière est proche de 0 et que la « gaussienne » et votre histogramme coïncident, on admettra que votre distribution est normale.

- La seconde méthode consiste à utiliser un papier gausso-arithmétique dont l'échelle est telle que toute fonction de répartition d'une loi normale à l'allure d'une droite, appelée **droite de Henry**. Calculez les fréquences cumulées de votre échantillon et tracez les points obtenus sur votre papier millimétré...

- La troisième méthode utilise un test de normalité (*Normal Probability Plot* sous Past) Si vos points sont alignés et suivent la droite de régression, on répondra positivement à la question de savoir si la distribution est normale.

Exemple : Reprenons l'échantillon des 156 valeurs associées à la distance des arbres au plus proche voisin et commençons par vérifier la normalité de la distribution :

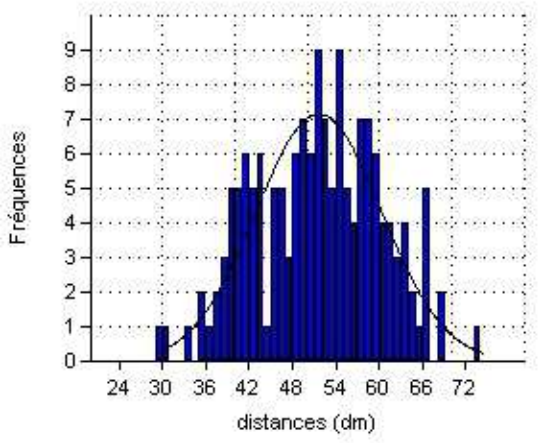

FIG. 3.7 – histogramme et courbe de Gauss

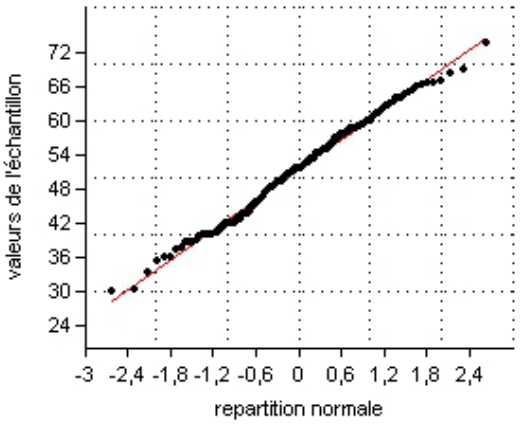

FIG. 3.8 – test de normalité (*Past*)

- La distribution peut être considérée comme « normale ».

- On calcule par le moyen de son choix la moyenne et l'écart-type de l'échantillon. On obtient ici : $\bar{x} = 51,50$ et $s = 8,73$.

- ***Conclusion*** : Nous pouvons assurer qu'environ 68% des distances à l'arbre le plus proche sont comprises entre $51,50 - 8,73 = 42,77$ dm et $51,50 + 8,73 = 60,23$ dm, qu'environ 95% d'entre elles sont comprises entre $51,50 - 2 \cdot 8,73 = 34,04$ dm et $51,50 + 2 \cdot 8,73 = 68,96$ dm.

Remarque 1 : Après vérification sur l'échantillon de données, on trouve 104 valeurs comprise dans l'intervalle $[42.77; 60.23]$, soit $104/156 = 0.6666$. Ce qui confirme la première approximation d'environ 68% des valeurs dans cet intervalle.
On trouve 152 valeurs comprise dans l'intervalle $[34, 04; 68, 96]$, soit $152/156 = 0, 9743$ ou $97, 4\%$ de la population, ce qui reste cohérent avec l'approximation fournie plus haut.

Remarque 2 : L'intérêt principale de ce type de calculs est là encore de pouvoir **comparer** des jeux de données et d'en simplifier la représentation.

Supposons qu'une autre parcelle ait été recencée et qu'on obtienne l'histogramme des distances au plus proche voisin suivant :

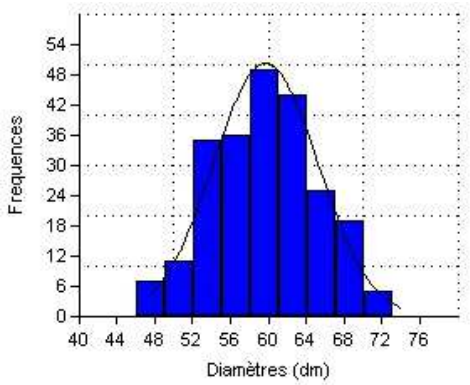

FIG. 3.9 – histogramme et courbe de Gauss

Le distribution étant approximativement normale, on comparera les données par le diagramme en bâton suivant (Excel) :

Les moyennes des distances à l'arbre le plus proche sont significativement plus importantes sur la deuxième parcelle puisque 68% des valeurs sont supérieures à la moyenne de la première parcelle.

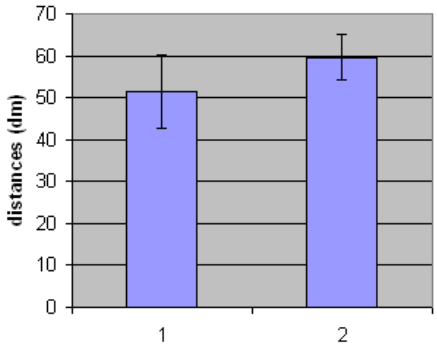

FIG. 3.10 – comparaison des distances entre les arbres les plus proches sur 2 parcelles

Distribution quelconque

Théorème : *Inégalité de Bienaymé-Tchebychev*
Toute variable aléatoire X admettant un moment d'ordre deux vérifie l'égalité :

$$\forall \epsilon > 0,\ \mathbb{P}\left(|X - \mathbb{E}(X)| \geq \epsilon\right) \leq \frac{V(X)}{\epsilon^2}$$

Admettons à nouveau que la série observée S de moyenne $\bar{x}$ et d'écart-type s corresponde à n réalisation successive d'une variable aléatoire théorique X qui a même loi, même espérance et même variance que nous supposerons égales respectivement à $\bar{x}$ et s.

Au regard de la formule de Bienaymé Tchebychev, on a pour tout k entier naturel non nul :

$$\begin{aligned}\mathbb{P}\left(|X - \bar{x}| \geq ks\right) &\leq \frac{V(X)}{(ks)^2} \\ &\leq \frac{V(X)}{k^2 V(X)} = \frac{1}{k^2}\end{aligned}$$

Or : $\mathbb{P}\left(|X - \bar{x}| < ks\right) = 1 - \mathbb{P}\left(|X - \bar{x}| \geq ks\right)$

D'où :

$$\mathbb{P}(\bar{x} - ks < X < \bar{x} + ks) \leq 1 - \frac{1}{k^2},\ \forall k \in \mathbb{N}^*$$

Dans les cas particuliers, $k = 2$ et $k = 3$ on a :

Mesures de dispersions :
- au moins 3/4 de toutes les données (ou 75%) sont à moins de 2 écarts types de la moyenne.
- au moins 8/9 de toutes les données (ou 89%) sont à moins de 3 écarts types de la moyenne.

Remarque : Comme dans toutes les applications de l'inégalité de Bienaymé-Tchebychev, **les majorations sont grossières**. Elles permettent néanmoins, dans les cas de distributions non symétriques, d'interpréter et de comparer des séries de données au regard de leur seule moyenne et écart-type (le graphe, dans ce cas, sera identique à celui donné à la figure 3.10).